

This is the **accepted version** of the article:

Mueller, Hannes; Gröger, Andre; Hersh, Jonathan; [et al.]. «Monitoring war destruction from space using machine learning». Proceedings of the National Academy of Sciences of the United States of America, Vol. 118, Issue 23 (June 2021), art. e2025400118. DOI 10.1073/pnas.2025400118

This version is available at <https://ddd.uab.cat/record/254903>

under the terms of the  **CC BY** COPYRIGHT license

MONITORING WAR DESTRUCTION FROM SPACE USING MACHINE LEARNING

Hannes Mueller

Institute of Economic Analysis (IAE-CSIC) and
Barcelona Graduate School of Economics (BGSE)
hannes.mueller@iae.csic.es

Andre Groeger

Department of Economics and Economic History
Universitat Autònoma de Barcelona (UAB) and
Barcelona Graduate School of Economics (BGSE)
andre.groeger@uab.es

Jonathan Hersh

Argyros School of Business
Chapman University
hersh@chapman.edu

Andrea Matranga

Smith Institute for Political Economy and Philosophy
Chapman University
matranga@chapman.edu

Joan Serrat

Computer Science Department and Computer Vision Center
Universitat Autònoma de Barcelona (UAB)
joans@cvc.uab.es

ABSTRACT

Satellite imagery is becoming ubiquitous and is released with ever higher frequency. Research has demonstrated that Artificial Intelligence (AI) applied to satellite imagery holds promise for automated detection of war-related building destruction. While these results are promising, monitoring in real-world applications requires consistently high precision, especially when destruction is sparse and detecting destroyed buildings is equivalent to looking for a needle in a haystack. We demonstrate that exploiting the persistent nature of building destruction can substantially improve the training of automated destruction monitoring. We also propose an additional machine learning stage that leverages images of surrounding areas and multiple successive images of the same area which further improves detection significantly. By combining these steps, we construct an automated classification of building destruction which allows real-world applications and we illustrate this in the context of the Syrian civil war.

Keywords Conflict · Destruction · Deep Learning · Remote Sensing · Syria

1 Introduction

Building destruction during war is a specific form of violence that is particularly harmful to civilians, commonly used to displace populations, and therefore warrants special attention. Yet, data from war-ridden areas are typically scarce, often incomplete, and highly contested, when available. The lack of such data from conflict zones severely limits media reporting, humanitarian relief efforts, human rights monitoring, reconstruction initiatives, as well as the study of violent conflict in academic research. A novel solution to this problem is to use remote sensing to identify destruction in satellite images [1, 2, 3]. This approach is gaining momentum as high-resolution imagery is becoming readily available at ever higher frequency, yielding weekly or even daily images. At the same time recent methodological advances related to deep learning have provided sophisticated tools to extract data from these images [4, 5, 6, 7].

While seminal research has demonstrated the use of automated classifiers for destruction detection, practical applications have so far been hampered by severe problems with labeling, domain transfer and class imbalance in real world imagery from urban war zones. As a consequence, international organizations such as the United Nations, the World Bank, and Amnesty International use remote sensing with *manual* human classification to produce damage assessment case studies

[8, 9, 10]. On the other hand, providers of conflict data for research purposes still rely heavily on news and eyewitness reports which leads to large data publishing lags and potential biases [11, 12, 13, 14, 15, 16, 17]. An automated building damage classifier for use with satellite imagery, which has a low rate of false positives in unbalanced samples and allows tracking on-the-ground destruction in close to real-time, would therefore be extremely valuable for the international community and academic researchers alike.

In this article, we present a new way of combining computer vision techniques and publicly available high-resolution satellite images to produce building destruction estimates that are of practical use to both practitioners and researchers. The standard architectures for this task are convolutional neural networks (CNNs)¹ as they have achieved unprecedented success in large-scale visual image classification with error rates beating humans [18, 19]. We train a CNN to spot destruction features from heavy weaponry attacks (i.e. artillery and bombing) in satellite images such as the rubble from collapsed buildings or the presence of bomb craters.

We make three relevant methodological contributions. First, we introduce a novel label augmentation method for expanding destruction class labels by making reasonable assumptions about the data generating process using contextual information. Second, we introduce a two-stage classification process to control for spatial and temporal noise where the results from the CNN are processed through a random forest model that relies on spatial and temporal leads and lags to improve classification performance. Third, we apply our trained computer vision model to repeated satellite images of the entire populated areas of major Syrian cities, including parks and highways, and produce longitudinal estimates of building destruction over the course of the recent civil war.

We demonstrate that our method yields high performance in out-of-sample tests and validate its ability for destruction monitoring using a separate database of heavy weaponry attacks. Our results highlight the importance of repeated satellite imagery in combination with temporal filtering to improve monitoring performance. As a result, our approach can be applied to any populated area provided that repeated, high-resolution (i.e. sub-meter) satellite imagery is available.

Why Automated War Destruction Monitoring Is Hard

Several studies have demonstrated the use of computer vision on satellite imagery to identify different types of destruction [2, 20, 21, 3, 22, 23, 24, 25, 26]. In many cases this is destruction from natural disasters which tends to be spatially concentrated. While performance results from the literature are encouraging, they typically focus on evaluations at one point in time and training/validating on datasets composed of equal numbers of damaged and undamaged images.

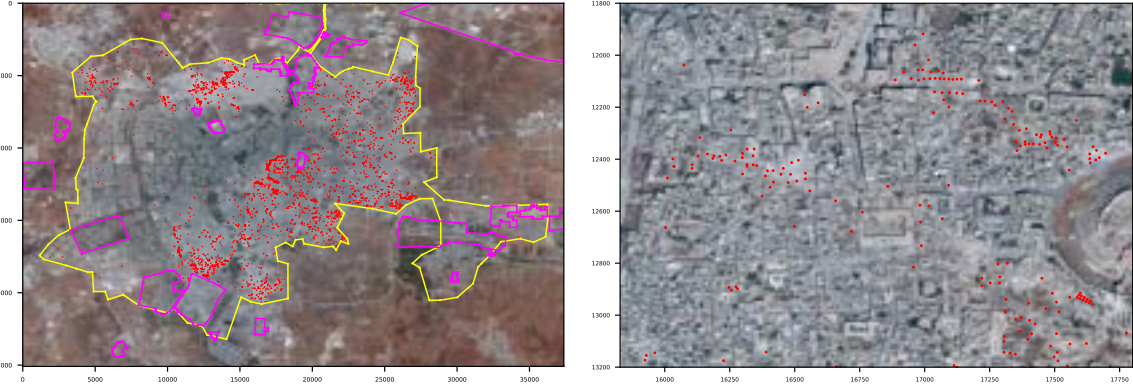
Precision performance in repeated destruction scans of entire cities with heavily unbalanced classes, as in our application, have not been explicitly presented in the literature so far. Part of the reason for this gap is that automated methods need to be able to detect building destruction in an empirical context where the vast majority of images do not feature destruction. Class imbalance is a common problem in machine learning applications but the detection of destruction in war zones faces an extreme level of imbalance. Even in a city which suffered as much destruction as Aleppo, only 2.8% of all images of populated areas contain a building that was classified as destroyed by UNOSAT in September 2016.

Figure 1 depicts this quite clearly. In the left panel (a) we see the full extent of Aleppo, with all destroyed building annotations depicted as red dots. The right panel (b) zooms into the central area of Aleppo, just east of the historic Citadel, which was heavily attacked. The red dots coincide clearly with patterns of destruction from heavy weaponry attacks in the satellite images. But destruction only affected a small fraction of buildings, even in this heavily affected part of the city.

With such class imbalance even a small false positive rate (FPR) will result in an unacceptable absolute number of false positive predictions in applications which would yield destruction data that are practically useless due to high measurement error. A simple example illustrates this: Suppose we have 100,000 sample images of which 1000 are destroyed. A "low" FPR of 15% together with a true positive rate (TPR) of 90% implies that the classification model will produce 14,850 false positives and 900 true positives, resulting in a precision below 6%. In other words, conditional on predicting destruction, such a classifier would be wrong more than 94% of the time. Note that the same classifier produces a "high" precision score of 86% on a 1:1 balanced sample.

The task of automated monitoring over time is typically further complicated by a lack of training data, i.e. the low number of destruction labels available in any given city. This can quickly lead to overfitting in machine learning as the training set consists of a narrow selection of building types, neighborhoods, sun and satellite angles, changing vegetation or weather phenomena like snow and cloud coverage. These problems are known as spatial and temporal

¹For a glossary of technical key terminology, see the Supplementary Information (SI)



(a) Full Extent of Aleppo, Syria.

(b) Zoom Showing Destruction is Sparse

Figure 1: Imagery of Aleppo 09/18/2016. Red dots indicate UNOSAT annotations as *destroyed*. Areas enclosed by magenta lines are *no analysis zones*, excluded from the UNOSAT damage assessment due to being non-civilian. The yellow line encloses the populated areas of Aleppo under analysis. *Sources*: Google Earth/Maxar satellite imagery and UNITAR/UNOSAT damage annotations.

domain shift [27]. Temporal domain shift is a particularly serious problem in our application as destruction monitoring requires the generation of a reasonable timeline with repeated scans of the same city. This emphasizes the need for a robust solution to this problem which ensures some comparability across time.

Our approach aims at solving these problems. We exploit the time dimension of the images and labels to alleviate the domain shift problems and extreme class imbalance. We also make a point of reporting precision performance in unbalanced samples to provide realistic insights into the potential performance in applications.

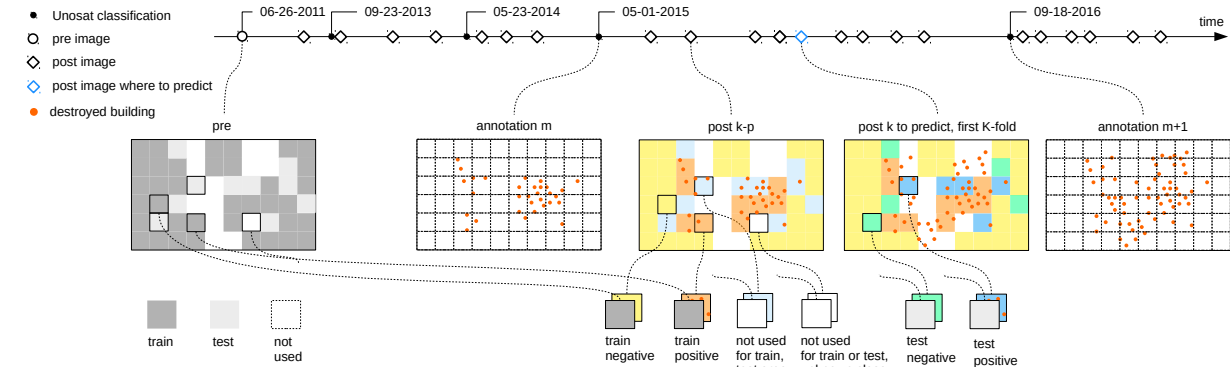


Figure 2: Image Sampling and Prediction Process. Timeline shows 23 Aleppo images. The first image, from 06/26/2011 is used as pre-war image when training the classifier. All other 22 images are used as post images. Images are split into over 95,000 patches which serve as a unit of analysis and are separated into test and training sample before the analysis. Labels for the patches come from UNITAR/UNOSAT annotation dates shown as black dots on the timeline. Annotations are extended forward and backward in time beyond these dates under the assumption that buildings which are labeled destroyed at some point remain destroyed throughout the period of observation. Those which are labeled as not destroyed at a given time were not destroyed before. Patches which are not destroyed at an annotation date but destroyed at a later annotation date have an unknown class. All patches which are not classified as destroyed in the last annotation date are of unknown class (set to missing) after that date.

Table 1: Sample overview

City	(1) Total images	(2) Total patches	(3) Total labeled images	(4) Total labeled patches	(5) Share destroyed patches
Aleppo	22	2,106,412	4	1,626,920	1.83%
Daraa	13	202,462	4	125,231	1.00%
Deir-Ez-Zor	7	98,602	4	84,723	2.86%
Hama	9	285,057	3	224,365	3.73%
Homs	5	200,035	2	83,941	8.26%
Raqqa	8	180,184	3	112,481	1.96%
All	64	3,072,752	20	2,257,661	2.26%

Note: Column (1) reports the number of "post-" satellite images / time periods, excluding the first "pre"-image for each city. Column (2) reports the resulting number of patches in the populated areas of the respective city based on available imagery. Column (3) refers to the number of images / time periods for which UNITAR/UNOSAT labels are available. Column (4) is the number of patches for which UNITAR/UNOSAT damage labels for the "destroyed" class are available after label augmentation. Column (5) is the share of "destroyed" labels over the number of labeled patches. Sources: Author calculations based on Google Earth/Maxar satellite imagery and UNITAR/UNOSAT damage annotations.

Method

Satellite Data

Most of our sample comes from Aleppo which we use as our main proof-of-concept due to the size of the city and the high availability of repeated images and labels. To train and evaluate our model, we use 22 high-resolution satellite images from Aleppo and a total of 42 images from five other Syrian cities (see Table 1). All images used in this analysis are obtained from Google Earth [28], are georeferenced, orthorectified, and feature three bands (RGB) as well as a ground sampling distance of ca. 50cm per pixel.

Sample images cover the period 2011 to 2017, after the onset of the civil war in Syria, during which extensive destruction from heavy weaponry attacks occurred across all sample cities. We use an additional, early image for each city (for example, 26 June 2011 in Aleppo) as the “pre” image and call the later 64 images as the “post” images. Our method relies on change detection – i.e. when classifying images, the pre image is compared to the respective post image.

To move as close as possible to the automated monitoring task we transform all images into millions of 64x64 pixel sub-images that we call *patches*. These patches are the unit of observation for training and testing, and the final step which we call *scanning* or *dense prediction* in which the classifier is used to produce fitted values for every patch in the study areas. Ground area coverage of each patch can vary slightly, but is approximately 1,024 (i.e. 32×32) square meters. Importantly, the size of a specific patch remains constant over time.

Column (2) in Table 1 reports the sample size in terms of patches for the six cities in our sample. For Aleppo, for example, we have over 95,000 patches per image times 22 images, which gives approximately 2.1 million patches. Importantly, this is panel data where images of the same patch are repeated 22 times.

Destruction Labels

We combine the imagery data with georeferenced building damage labels from the United Nations Operational Satellite Applications Programme (UNOSAT) of the United Nations Institute for Training and Research (UNITAR) [8]. Over the course of the Syrian civil war, UNOSAT produced building destruction annotations by manual inspection of satellite images for severely affected Syrian cities. For Aleppo, these manual assessments were conducted at four different dates, one each year between 2013 and 2016. Column (3) in Table 1 reports the number of these assessments.

UNOSAT damage annotations were categorized in three degrees of damage: moderate and severe damage as well as completely destroyed. In our analysis we rely on the latter class due to the fact that destruction patterns for the other

labels were not always clearly visible in the satellite images. Our method classifies the satellite images as destroyed if at least one UNOSAT annotation of destruction is inside a patch.

Our analysis of building destruction focuses on the urban areas of Syrian cities. For Aleppo this is depicted by the area enclosed by the yellow line in Figure 1. Areas enclosed by magenta lines correspond to so-called "no analysis" areas, which have been left out by UNOSAT in their damage annotations due to these zones hosting non-civilian buildings. Consequently, these areas are also excluded from the training process. But we scan these areas and make use of these scans for out-of-sample validation. Sample image patches for destroyed areas pre- and post- destruction are presented in Figure S1 and non-destroyed ones, including damaged buildings, are shown in Figure S2 in the SI.

The ideal annotation dataset to analyze this problem would be composed of pixel-wise classification of all damaged and non-damaged buildings across the sample cities for all time periods. Labels like this could then be used to train models to identify the footprint of destroyed buildings using satellite images [3, 22]. However, because of the significant cost of annotating destruction footprints, UNOSAT only provides point coordinates (centroids) of destroyed buildings. We match these point labels to our image patches by attributing a label to the closest patch centroid. One issue with this method of generating labels is that buildings have different sizes and, therefore, some UNOSAT labels are surrounded by more visible destruction than others. We address this issue through a second stage, described below, in which we exploit spatial information.

Contextual Label Augmentation and Test Sample

The computer vision task is to train an algorithm to detect destruction from the visual bands of high-resolution daylight satellite images. Training deep learning architectures typically requires large training datasets including thousands of labels, which are extremely rare in our empirical context.

Consequently, as reported in column (3) of Table 1, we have a maximum of four UNOSAT annotation dates to work with for certain cities, for others three or only two (i.e. Homs). Compared to the number of annotations, we usually have significantly more raw images available, as shown in column (1). In addition, few label dates perfectly coincide with the date of a satellite image. This generates an "uncertain class" in which patches cannot be attributed clearly to either the destroyed or not destroyed class because destruction could have occurred between the labeling date and the date of the image.

To increase the number of labeled data points we exploit the fact that reconstruction was largely absent in the areas of interest during the study period between 2013 and 2017 (see Table S4). Our label augmentation approach assumes that positive samples at time t_i also remained positives at subsequent times $t_j > t_i$, i.e. that destruction persists throughout the period of the civil war. And conversely, that negative samples at time t_j also had to be negatives at times $t_i < t_j$.

We solve two problems using this approach. First, we expand the size of our training data set by boosting the number of labels to close to 2.3 million of which approximately 51,000 show destruction. Second, by including additional time periods in our training sample, we improve the performance of our classifier in its ability to handle domain shift. Our method of label augmentation is conservative given that we assign missing values to all patches that remain in the uncertain class - those for which we cannot know with certainty whether destruction has occurred in the past or those for which we do not know with certainty that they will be labeled not destroyed in the future.

Figure 2 illustrates our method for generating training and test samples. Given the temporal and spatial structure of the data, extra care must be taken when splitting the sample for training and testing to avoid overfitting. Standard cross-sectional cross-validation procedures are not appropriate since they could show the network patches from different times, but the same location in training and testing. We therefore use the patch identifier to perform sample splitting, whereby 70% of patches are reserved for training and 30% for testing across temporal periods. All performance measures reflect accuracy as measured from data reserved in the test set.

CNN Architecture and Two-stage Classification Procedure

Another innovation in our approach is the use of a two-stage classification procedure that feeds the predicted destruction estimates from the initial CNN model into a random forest classifier. With respect to the CNN architecture, we experimented with several different types of CNNs. For each of these we optimized hyper-parameters according to accuracy results in the validation set. The results of these experiments suggested the use of a relatively flat CNN architecture as described in section 1 of the SI.

To the output of the CNN model we apply a second machine learning stage, intended to exploit the temporal and spatial clustering of destruction. Specifically, the labels and predicted values from the CNN are used to train a random forest model that relies on information from two spatial lags around each patch location and two temporal leads/lags around

each date. The random forest uses these spatial and temporal features from the raw CNN scores plus the spatial standard deviation to generate a prediction for the test sample and the dense prediction.

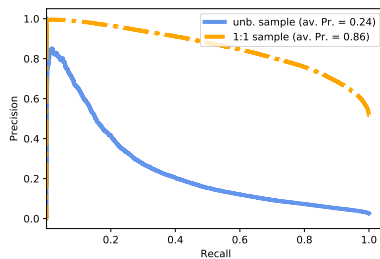
The logic behind this second stage approach is that destruction is not only serially correlated, but also spatially clustered. We separate this step from the deep learning stage for maximum flexibility and modularity. This allows us to vary the information set that we use in the second stage model. In particular, we experimented with using only spatial information and different temporal lag structures and discuss their relative importance below.

Data Generation

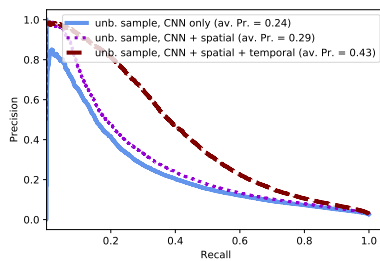
As a final step we train the second stage on all available data and predict values for every patch-period combination in our data. This simulates the data generation problem where the trained architecture is used to interpret all patches at all points in time including those patches that had missing labels. The result is what we call *dense* predictions and this forms the raw material for additional validation exercises. As reported in column (2) of Table 1, the result is a panel dataset of destruction predictions at the patch level for six cities with varying time periods with over 3 million patch-time observations.

Results

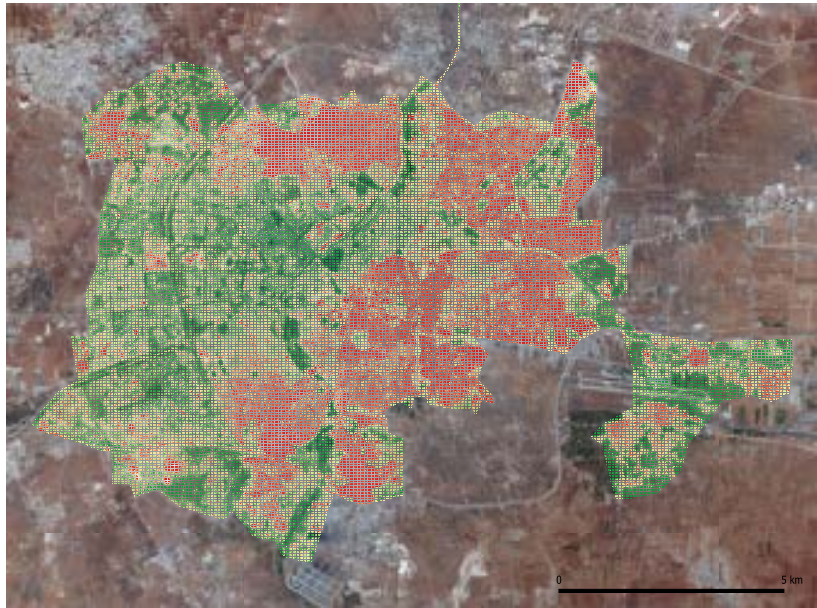
Overall Performance



(a) Precision Using First-stage Only.



(b) Second-stage Precision Improvement.



(c) Patch-Wise Second Stage Destruction Prediction Scores for Aleppo City, Syria.

Figure 3: (*Top Left*) Precision-Recall Curve, Unbalanced Versus Balanced Sample. Reported performance is in the 30% training sample either by up-sampling the positives to reach a 1:1 sample (orange curve) or by evaluating at the original sample proportions (blue curve). (*Bottom Left*) Precision-recall curve, unbalanced sample. First stage model versus two alternative second stage models. As in Figure a) blue curve shows performance after the first stage. Dashed maroon curve shows performance after the second stage which uses training of a random forest on temporal and spatial leads and lags in the training sample. Dotted purple curve shows performance when using only spatial lags and no additional temporal information. (*Right*) Average second stage dense patch-wise destruction prediction scores for Aleppo city, Syria. Green color indicates low prediction scores, red color indicates high prediction scores. Color bins reflect deciles of second stage fitted values with full spatial and temporal smoothing. Sources: Google Earth/Maxar satellite imagery, UNITAR/UNOSAT damage annotations, and author calculations.

Our first stage CNN classifier achieves an Area Under the Curve (AUC) of 0.86 in the test sample of the first stage (i.e. with the raw output from the CNN) and an AUC of 0.92 after the second stage random forest procedure (see Figure S4). The associated ROC curve implies a true positive rate of 0.8 is achieved a false positive rate of 0.17. At

Table 2: Model performance when varying second-stage module in the unbalanced sample

City	(1) First-stage (CNN)	(2)	(3)	(4)
	raw	Second-stage (CNN+RF)		
		with spatial leads/lags	with spatial & temporal leads/lags	with spatial & temporal leads/lags
	<i>precision</i>	<i>precision</i>	<i>precision</i>	<i>AUC</i>
Aleppo	16.1	16.9	35.7	91.5
Daraa	4.2	4.6	11.7	89.0
Deir-Ez-Zor	11.0	12.1	21.7	80.0
Hama	54.5	65.2	68.0	91.0
Homs	25.8	34.9	55.2	85.7
Raqqa	12.8	17.4	32.1	87.6
All	24.5	28.7	42.5	90.7

Note: first-stage predictions from convolution neural network (CNN) and second-stage predictions from random-forest model (CNN+RF) with spatial leads/lags (column 2) and spatial and two temporal leads/lags (columns 3 and 4). Columns (1) through (3) report the average precision and column (4) the "Area Under the Curve" (AUC). Sources: Author calculations based on Google Earth/Maxar satellite imagery and UNITAR/UNOSAT damage annotations.

a more conservative, higher threshold for a positive classification a true positive rate of 0.5 is associated with a false positive rate of only 0.025. However, the class imbalance is extremely relevant here. The ROC curve and its AUC are classification performance measures which are not affected by class imbalance in the sample and therefore do not allow us to discuss the impact of class imbalance in our sample. In what follows, we therefore focus on precision statistics to highlight the problem of unbalanced classes in applications of automated destruction detection.

Figure 3 summarizes our main results across cities. The top left panel (a) presents two precision-recall curves from the test sample which depict the out-of-sample performance of our classification approach. The dashed orange curve plots the precision-recall trade-off in the balanced sample. The average precision here is 0.86 and the curve suggests a very mild trade-off with a precision of over 0.9 at a recall rate of 0.5, for example. In contrast, the solid blue line depicts the performance of the same model when taking into account unbalanced classes that the automated destruction detection would face in the actual application in the test sample. Clearly precision is much lower with the average precision being a mere 0.24. For a recall rate of 0.5 the first stage reaches a precision of below 0.2. This illustrates impressively how class imbalance in real application can change the precision-recall trade-off in this exercise.

In the bottom left panel (b) we illustrate the improvement in precision that we achieve by applying the second stage. The figure compares precision-recall curves for the first stage (solid blue line), as in panel (a), with the improvements from the second stage models, all evaluated in the unbalanced test sample. The second stage average precision increases to 0.29 with only spatial smoothing (dotted purple line) and 0.43 with temporal and spatial smoothing (dashed maroon line). This highlights a key insight from our experiments with the modular second stage. The use of temporal smoothing is absolutely crucial for reaching better precision in the second stage. The gains of the spatial smoothing are relevant in some cases but the real boost in performance arises when using temporal information to validate predictions coming out of the first stage.

In panel (c) of Figure 3, we show an example of the final output of our methodology - the continuous dense prediction scores generated from the second stage. The figure shows the average patch-wise dense predictions across the entire city of Aleppo, including no-analysis zones. Red color indicates high predicted scores and green indicates low scores. Generally, the red areas coincide with the destruction annotations in Figure 1. In addition, roads and parks are clearly visible as dark green (lowest destruction probability) or yellow patches. This is not only evidence of the power of our approach in picking up housing destruction, but it also shows how the classifier has *learned* that roads and parks are never destroyed buildings.

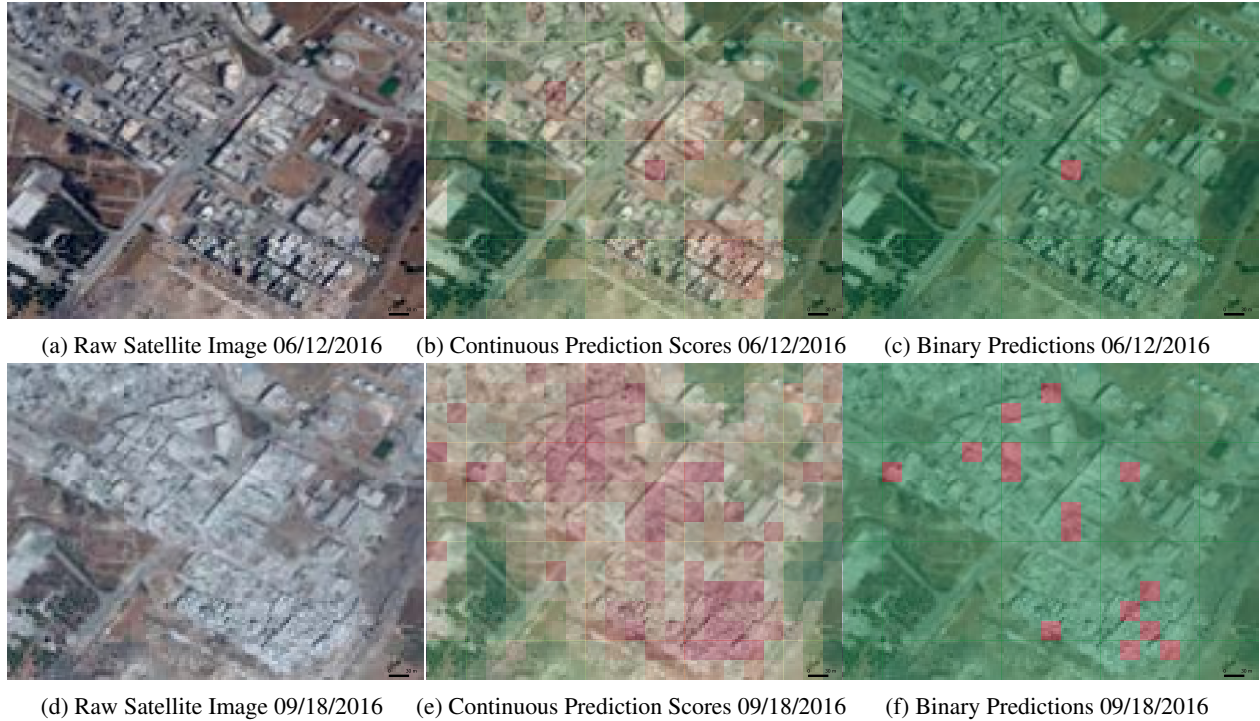


Figure 4: Example of Raw Satellite Images (left panel) and Second-stage Patch-wise Continuous Predictions Scores (middle panel) and Binary Classification (right panel) for Ramouse Neighborhood of Aleppo, Syria. Before (first row) and after (second row) heavy weaponry attacks. Green color indicates low prediction scores, red color indicates high prediction scores. Color bins reflect deciles of fitted values. Binary classification cutoff optimized to reach 50 percent recall in the test sample. Satellite image recording dates: 06/12/2016 (before) and 09/18/2016 (after). Approximate image centroid location: 36.1525 decimal degrees North and 37.1332 East. Sources: Google Earth/Maxar satellite imagery and author calculations.

The Role of the Second Stage Module

The second stage plays a key role for boosting performance to levels that imply practical gains from automatizing destruction monitoring in our sample. It is important to consider that, while the cities in our sample are all in the same country, they are of different size, have different building types and are situated in different landscapes with a variety of vegetation and seasonal changes. In addition, label and image availability differ dramatically. As shown in Table 1 the vast majority of images in our sample comes from Aleppo due to its large size and elevated image availability - less than one third of all images come from other cities (Table S3 summarizes the results from training on Aleppo exclusively). If our approach can adapt to these very different conditions it means we can be optimistic about applications elsewhere.

Table 2 provides details on the performance improvements through the second stage procedure by city. In column (1) we report performance of the first stage by city. This reveals strong differences in performance across cities with average precision ranging from a mere 4.2 percent for Daraa to an impressive 54.5 percent for Hama (for corresponding precision-recall curves, see Figure S5). To a large degree this is driven by sample imbalances where Daraa suffered only 1 percent of destroyed patches on average whereas Hama suffered almost four times as much.

The second stage boosts this performance substantially. This is most notable for the worst performing cities for which precision improves two- to threefold in the full model (column 3). How does the full model achieve this improvement in performance? Table 2 confirms the role of the temporal smoothing shown in Figure 3. However, the city-by-city analysis also reveals interesting differences across cities where Homs and Hama seem to benefit more from the spatial smoothing. In both cities, destruction is indeed clustered heavily in some neighborhoods so that this clustering might be useful in reinforcing patch-wise predictions in the second stage. Our predictions for Daraa, Deir-Ez-Zor and Aleppo rely much more on repetition and temporal smoothing. We confirm the role of temporal smoothing in Table S5 by varying temporal lags and providing performance estimates without spatial smoothing.

The improvements with temporal smoothing suggest the domain shift problem across time plays an important role when angles, lighting, vegetation and seasons change. Our results therefore highlight the potential role of repeated high-frequency imagery and temporal smoothing for providing useful destruction monitoring. The extreme imbalance combined with small samples imposes serious trade-offs for monitoring but we will show in the following section that monitoring can be brought to work even in the case of Aleppo which has one of the more unbalanced samples in our dataset.

External Validation Exercises

We conduct two validation exercises to illustrate the merits of our approach. We first make use of the no-analysis areas in Aleppo (see Figure 1) that have been entirely excluded from the training process. One of these zones corresponds to the Ramouse neighborhood in the southernmost tip of our study area in Aleppo – an area which our classifier identified as heavily destroyed as depicted in panel (c) of Figure 3.

In Figure 4 we show satellite imagery from a subarea of the Ramouse neighborhood at two points in time, before (06/12/2016, top row) and after (09/18/2016, bottom row) a major heavy weaponry attack. We show raw satellite images (left panel), patch-wise visualizations of the second-stage continuous predictions scores (middle panel), and a binary classification (right panel). Due to the classifier not having been trained on this area, this exercise serves as a good out-of-sample validation test. Visual inspection of the raw images shows no destruction before (panel a), but extensive building destruction after the attacks (d). Comparing the continuous prediction scores before (b) and after the attack (e), shows a significant increase in predicted destruction by our approach which coincides clearly with the locations of actual destruction of buildings in the area. Note that the model also classifies correctly areas without building destruction, such as the industrial compounds in the Northeast and Southwest of the image, as not destroyed at both points in time. The same applies to the fields and roads in the East and the forest in the West. The panel on the left shows one way of converting continuous prediction scores into a binary classification. The threshold chosen here is optimized to reach a level of 50 percent recall in the test sample. One can observe that the before period is consistently classified as non-destroyed (with one exception), whereas destruction is indicated in affected areas after the attacks.

Figure 4 demonstrates that the classifier is able to identify destruction in parts of the city which were not part of the training sample. This is important as it shows that we are able to successfully solve the spatial and temporal domain shift problems within Aleppo and thus generate a time series of destruction data in this way. If our automated method was to augment human monitoring this is the kind of data that would be passed to human verification.

Given our strategy of expanding labels forward and backward in time, it becomes particularly important to verify the ability of our approach to approximate the timing of destruction. We therefore validate our dense predictions in an event study framework which relies on an external dataset of georeferenced bombing events in Syria. In particular, we rely on 731 bombing events with precise location information from the Live Universal Awareness Map project (LiveUAmap). We merge these events with our pooled sample of dense predictions at the patch-time level. We then conduct an event study regression on a sample of over 2.8 million observations to test whether our prediction scores increase in the aftermath of an externally reported bombing event (see SI section 2 for details).

We present a coefficient summary plot for two second-stage modules in Figure 5. The graph shows clearly that bombing events are positively and significantly correlated to the destruction scores at the time- and patch-level. Note that the baseline hazard of destruction, i.e. the mean of the dependent variable, is very small in our sample (see Table S2). Compared to the baseline level of the respective destruction score, the point estimates imply increase of 29% and 37%, respectively, after a bombing event is reported in a given cell. This is a substantial increase if one keeps in mind that not all bombing events will result in the destruction of a building, introducing attenuation bias in the regression. The Figure also shows that temporal smoothing implies big gains in overall signal strength, with the coefficients from the full model represented by the red diamonds being consistently above the spatial only model as depicted by the blue squares.

Discussion

Building destruction due to heavy weapon attacks is a particularly salient form of war-related violence. Destruction is often used as a military strategy to displace population and is responsible for tremendous human suffering beyond the loss of life. Likewise, organizations like the Red Cross warn that massive destruction of urban infrastructure (also called *urbicide*) has dramatic knock-on effects on health as it implies the destruction of water and power supplies as well as hospitals. Therefore, reliable and updated data on destruction from war zones plays an important role for humanitarian relief efforts, but also for human rights monitoring, reconstruction initiatives, media reporting, as well as the study of violent conflict in academic research. Studying this form of violence quantitatively, beyond specific case studies, is currently impossible due to the absence of systematic data.

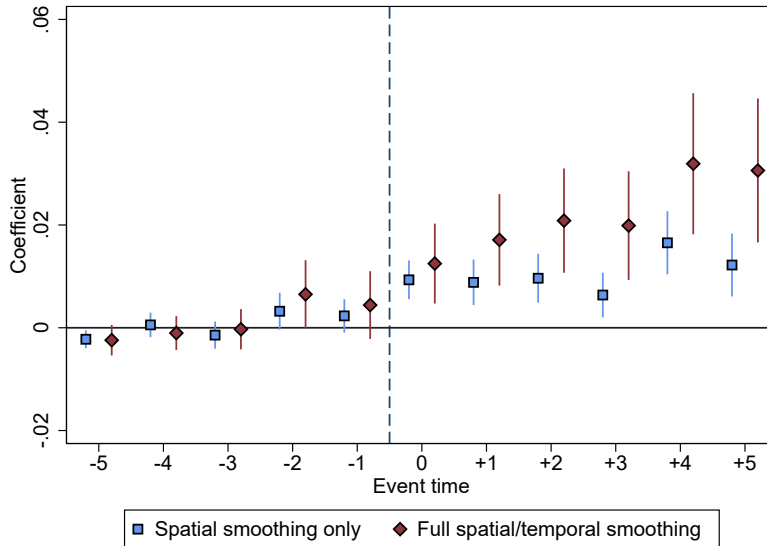


Figure 5: Event Study Validation Exercise Pooled Sample. External bomb event data from *LiveUMap* is positively and significantly correlated with satellite predicted war destruction at the patch level. The figure shows coefficients from a regression of 5 leads and lags of bombing events identified in the event data against our continuous destruction prediction score from the second stage. Point estimates depicted by blue squares correspond to second stage continuous prediction scores with spatial smoothing only and red diamonds correspond to the full model with spatial and temporal smoothing. Error bars represent 95% confidence intervals. Dashed line indicates the occurrence of a bombing event in the event data and coefficients capture the response in predicted damage. The full regression specification and results are reported in SI section 2 and Table S2, respectively. Sources: LiveUMap event data and author calculations.

Our method of identifying building destruction combines the existing state of the art of computer vision methods with an additional post-processing step, and exploits the time dimension of destruction data to expand the training data set. This allows us to exploit the repetition of imagery to bring down error rates in when classifying destruction. Thanks to these advances, we are able to achieve an AUC of above 0.9 and an average precision of over 0.42 in the unbalanced sample from six Syrian cities. We also show that our approach is able to identify the timing and location of building destruction out-of-sample, i.e. in areas of Aleppo that have not been used for training the classifier.

These results are encouraging and allow applicability for automated destruction classification, and even close to real-time tracking for policy purposes. Our method is particularly well-placed to take advantage of the ever increasing temporal granularity of imagery. Our calculations suggest that human manual labeling of our entire dataset would cost approximately 200,000 USD and additional repetitions of imagery would increase these costs almost proportionally. With an automated method like ours, higher image frequency helps precision and comes at only marginal extra costs. However, our results also suggest limitations where average precision falls, e.g. if only a very low share, less than one percent, of a city is destroyed. For applications requiring high precision in heavily imbalanced prediction problems such as the monitoring of several cities, we believe that the real use case for our approach will be in a decision support framework in which the predictions are combined with human verification to create much faster and accurate on-the-ground violence detection. Iterations between machine learning and human verification can also help improving the training process [29] and could be easily integrated in our approach.

The performance of our method could be further improved by increasing the size of the training dataset, which could also help adapt it to classify destruction in other war zones around the globe. Further performance improvement could be achieved through *fine tuning*, a common practice in deep learning in which the network is pre-trained with a large sample of building destruction from a variety of contexts in the first step, and then refined by training on heavy weaponry destruction. This could be implemented by using a recent public dataset of natural disaster destruction imagery that provides a sample of 98,000 annotated buildings across 3 levels of damage [30]. Moreover, domain adaptation techniques developed for deep learning could be used to try to further minimize the remaining domain biases [31].

Our label augmentation technique is driven by strong assumptions and should therefore be regarded only as a first step in understanding the dynamic classification of building destruction over time. A particularly fruitful direction for

future research could be to model the data generating process of what we call the "uncertain class" between changing labels and after the last label date. This should then be combined with label smoothing to generate probabilistic labels [32]. Such a holistic approach would also need to think about label priors regarding the reconstruction process. Future applications of such an approach would then be able to augment the human-classification process of verifying violence – so-called digital humanitarians [33] – and track the post-war recovery within the same classifier model.

The destruction data that can be generated with monitoring approaches such as the one presented in this article opens up possibilities for a set of new research agendas in the social sciences [34]. For example, our approach may advance the academic literature on understanding the micro-level determinants of violence [35, 36, 37, 38, 39, 40, 41]. At what stage in a conflict is building destruction used? What can be done to reduce civilian fatalities during urban warfare? What are the effects of building destruction on displacement compared to other kinds of violence such as small firearms? Can reporting-based violence data be used to reduce error in the remote sensing exercise or can combined measures be developed [42, 43]? Can destruction data be used to reveal biases in reporting-based measures? An additional potential application of our method is conflict forecasting systems like the "Violence Early-Warning System" (ViEWS) which rely on spatial violence dynamics in their forecasts [44].

Finally, there are important ethical concerns in war destruction monitoring which should be considered. Research in the social sciences has shown that monitoring tends to reduce armed violence between states, but there are also examples where the opposite is true [45, 46]. Theoretically, we can identify specific scenarios in which monitoring worsens the situation on the ground. If local actors are using the flow of information about atrocities to displace population and do not fear repercussions linked to the monitoring of these atrocities then monitoring itself can increase violence and should, therefore, not be conducted publicly.

Acknowledgements

We would like to thank Eli Berman, Joshua Blumenstock, Mathieu Couttenier, Joan Maria Esteban, Clément Gorin, Edward Miguel, Sebastian Schütte, and Jacob Shapiro for useful comments and discussions. We are grateful to Bruno Conte Leite, Jordi Llorens, Parsa Hassani, Dennis Hutschenreiter, Shima Nabiee, and Lavinia Piemontese for excellent research assistance. We are particularly grateful to Javier Mas for his research assistance which produced the coding backbone to this project. We thank seminar participants at the Applied Machine Learning, Economics, and Data Science, University of California Berkeley, Empirical Studies of Conflict Project Annual Meeting, AI for Development conference by the Center for Effective Global Action and the World Bank Development Impact Evaluation Group University of Bozen/Bolzano, International Institute of Social Studies of Erasmus University, University of Economics Ho Chi Minh City, Lyon University, Trinity College Dublin, Barcelona Graduate School of Economics, Institute for Economic Analysis of the Spanish Council for Scientific Research, Universitat de Barcelona, Berlin Network of Labor Market Research Winter Workshop, PREVIEW workshop at the German foreign office and Violence Early-Warning System workshop in Uppsala. A.G. and H.M. acknowledge financial support from the "la Caixa" Foundation project grant number CG-2017-04, title: "Analysing Conflict from Space", and from the Spanish Ministry of Science and Innovation, through the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000915-S). H.M. acknowledges financial support from the Spanish Ministry of Science, Innovation and Universities through grant PGC-096133-B-100. A.G. also acknowledges financial support from the Spanish Ministry of Science, Innovation and Universities through grant PGC2018-094364-B-100. J.H. and A.M. acknowledge support from the Chapman University Faculty Opportunity Fund. A.M. acknowledges support from the Smith Institute of Political Economy and Philosophy at Chapman University. Any remaining errors are our own.

References

- [1] Frank Witmer. Remote sensing of violent conflict: eyes from above. *International Journal of Remote Sensing*, 36(9):2326–2352, 2015.
- [2] L. Gueguen and R. Hamid. Large-scale damage detection using satellite imagery. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1321–1328, June 2015.
- [3] F. Kahraman, M. Imamoglu, and H. F. Ates. Battle damage assessment based on self-similarity and contextual modeling of buildings in dense urban areas. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5161–5164, July 2016.
- [4] Yoshua Bengio LeCun, Yann and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

- [5] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.
- [6] R. Engstrom, D. Newhouse, and J. Hersh. Poverty from Space: Using High Resolution Satellite Imagery for Estimating Economic Well-being and Geographic Targeting. *World Bank Policy Research Working Paper*, 2017.
- [7] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications*, 11(2583):1–11, 2020.
- [8] UNITAR Operational Satellite Applications Programme. Damage Density in the City of Aleppo, Syria, 2016.
- [9] World Bank. The Toll of War: The economic and social consequences of the conflict in Syria, 2017.
- [10] Amnesty International. Strike Tracker. Decode how US-led bombing destroyed Raqqa, Syria, 2020.
- [11] Nils Petter Gleditsch, Peter Wallensteen, Mikal Eriksson, Margareta Sollenberg, and Havard Strand. Armed conflict 1946-2001: A new dataset. *Journal of Peace Research*, 39(5):615–637, 2002.
- [12] Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. Introducing acled: An armed conflict location and event dataset: Special data feature. *Journal of Peace Research*, 47(5):651–660, 2010.
- [13] Meredith Reid Sarkees and Frank Wayman. *Resort to War: 1816 - 2007*. Washington DC: CQ Press, 2010.
- [14] Ralph Sundberg and Erik Melander. Introducing the ucdp georeferenced event dataset. *Journal of Peace Research*, 50(4):523–532, 2013.
- [15] Megan Price, Anita Gohdes, and Patrick Ball. Documents of war: Understanding the syrian conflict. *Significance*, 12(2):14–19, 2015.
- [16] Nils B. Weidmann. A Closer Look at Reporting Bias in Conflict Event Data. *American Journal of Political Science*, 60(1):206–218, 2016.
- [17] Therése Pettersson and Magnus Öberg. Organized violence, 1989–2019. *Journal of Peace Research*, 57(4):597–613, 2020.
- [18] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.
- [19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
- [20] Austin J. Cooner, Yang Shao, and James B. Campbell. Detection of urban damage using remote sensing and machine learning algorithms: Revisiting the 2010 haiti earthquake. *Remote Sensing*, 8(10), 2016.
- [21] L. Gueguen and R. Hamid. Toward a generalizable image representation for large-scale change detection: Application to generic damage analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 54(6):3378–3387, June 2016.
- [22] F. Kahraman, M. Imamoglu, and H. F. Ates. Disaster damage assessment of buildings using adaptive self-similarity descriptor. *IEEE Geoscience and Remote Sensing Letters*, 13(8):1188–1192, Aug 2016.
- [23] Jiangye; Yuan. Automatic Building Extraction in Aerial Scenes Using Convolutional Networks. *arXiv*, 2016.
- [24] N. Attari, F. Ofli, M. Awad, J. Lucas, and S. Chawla. Nazr-cnn: Fine-grained classification of uav imagery for damage assessment. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 50–59, Oct 2017.
- [25] A. Fujita, K. Sakurada, T. Imaizumi, R. Ito, S. Hikosaka, and R. Nakamura. Damage detection from aerial images via convolutional neural networks. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 5–8, May 2017.
- [26] F. Nex, D. Duarte, F.G. Tonolo, and N. Kerle. Building damage detection with deep learning: Assessment of a state-of-the-art cnn in operational conditions. *Remote Sensing*, 11(23):2765, 2019.
- [27] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016.
- [28] Google Earth. <https://www.google.com/earth>, 2020. Accessed: 2020-09-07.
- [29] M. Colaresi and Z. Mahmood. Do the robot: Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research*, 54(2):193, 2017.

- [30] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeew, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
- [31] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *Advances in Computer Vision and Pattern Recognition*, pages 1–35, 09 2017.
- [32] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *ArXiv*, (1906.02629), 2020.
- [33] Patrick Meier. *Digital humanitarians: how big data is changing the face of humanitarian response*. Routledge, 2015.
- [34] Kristian Skrede Gleditsch, Nils W Metternich, and Andrea Ruggeri. Data and progress in peace and conflict research. *Journal of Peace Research*, 51(2):301–314, 2014.
- [35] Timothy Besley and Hannes Mueller. Estimating the peace dividend: The impact of violence on house prices in Northern Ireland. *American Economic Review*, 102(2):810–833, 2012.
- [36] Oeindrila Dube and Juan F. Vargas. Commodity price shocks and civil conflict: Evidence from Colombia. *Review of Economic Studies*, 80(4):1384–1421, 2013.
- [37] Marshall Burke, Solomon M. Hsiang, and Edward Miguel. Climate and conflict. *Annual Review of Economics*, 7(1):577–617, 2015.
- [38] Stelios Michalopoulos and Elias Papaioannou. The long-run effects of the scramble for Africa. *American Economic Review*, 106(7):1802–1848, 2016.
- [39] Natalija Novta. Ethnic Diversity and the Spread of Civil War. *Journal of the European Economic Association*, 14(5):1074–1100, 2016.
- [40] Nicolas Berman, Mathieu Couttenier, Dominic Rohner, and Mathias Thoenig. This mine is mine! How minerals fuel conflicts in Africa. *American Economic Review*, 107(6):1564–1610, 2017.
- [41] Marco Manacorda and Andrea Tesei. Liberation technology: Mobile phones and political mobilization in africa. *Econometrica*, 88(2):533–567, 2020.
- [42] J. Vernon Henderson, Adam Storeygard, and David N. Weil. Measuring economic growth from outer space. 102(2):994–1028, 2012.
- [43] Kristian Lum, Megan Emily Price, and David Banks. Applications of multiple systems estimation in human rights research. *The American Statistician*, 67(4):191–200, 2013.
- [44] Håvard Hegre, Marie Allansson, Matthias Basedau, Michael Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Höglbladh, Remco Jansen, Naima Mouhleb, Sayyed Auwn Muhammad, Desirée Nilsson, Håvard Mogleiv Nygård, Gudlaug Olafsdottir, Kristina Petrova, David Randahl, Espen Geelmuyden Rød, Gerald Schneider, Nina von Uexkull, and Jonas Vestby. ViEWS: A political violence early-warning system. *Journal of Peace Research*, 56(2):155–174, 2019.
- [45] Grant Gordon. *Violence and Intervention*. PhD thesis, Columbia University, 2016.
- [46] Bryan R. Early and Erik Gartzke. Spying from space: Reconnaissance satellites and interstate disputes. *Journal of Conflict Resolution*, 2021.