Violence and Internal Displacement: Insights from Nationwide Mobile Phone Data

Xiao Hui Tai[a], Shikhar Mehra[a], and Joshua E. Blumenstock[a,*]

[a]School of Information, University of California, Berkeley, CA 94720
[*]Corresponding author: jblumenstock@berkeley.edu

This version of text:
May 24, 2021

Abstract

Nearly 50 million people globally have been internally displaced due to conflict, persecution, and human rights violations. However, the study of internally displaced persons (IDPs) — and the design of policies to assist them — is complicated by the fact that IDPs are often missing or under-represented in traditional surveys and official statistics. This paper develops a new approach to measuring the causal impact of violence on human displacement, based on the analysis of anonymized high-frequency mobile phone data. We use these methods to study the effect of violence on internal displacement in Afghanistan, the site of the longest war in U.S. history. Our results quantify the short- and long-term impact of violence on internal displacement, and highlight how displacement depends on the nature and location of the violence. High-casualty events, and violence involving the Islamic State (as opposed to the Taliban), cause the most displacement. We also highlight the important role of provincial capitals in the ongoing conflict: these capitals are resilient to local violence, and act as magnets for people fleeing violence in outlying areas.

Introduction

Every year, tens of millions of individuals and families are forcibly displaced by armed conflict, creating enormous humanitarian, social, and economic costs [1, 2]. Recent estimates indicate that the number of people living in proximity to conflict has doubled since 2007; by 2030, as many as two thirds of the global extreme poor will be living in fragile and conflict-affected situations [3]. Despite the global scale and significance of this crisis, empirical analysis of the link between violence and displacement is complicated by the inherent difficulty of observing displaced populations — a difficulty that is compounded in developing countries and in insecure environments [3, 4].

This paper develops and tests a new approach to studying the impact of violence on internal displacement. Our first contribution is methodological, and illustrates how high-frequency 'digital trace' data can enable new approaches to identifying and estimating the causal effect of violent events on internal displacement. We use a panel event study framework that is only possible because we observe the locations of millions of individuals near-continuously over time: the high-frequency data make it possible to draw causal inferences from discontinuities in spatiotemporal trajectories that coincide with specific violent events. Such an estimation strategy would not be feasible with traditional survey-based data, which track a relatively small number of individuals at very infrequent intervals. This complements recent qualitative [5, 6], survey-based [7, 8, 9], as well as observational studies [10, 11, 12] on conflict and displacement in the developing world. It also builds on recent work using non-traditional sources of digital data to study the movement of human populations [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23].

Our second contribution is to provide rich new evidence on the impact of violence on internal displacement in Afghanistan — a country that has experienced decades of conflict and which contains over two million internally displaced people [1]. We contribute a new quantitative perspective on the nature of this displacement, complementing more traditional approaches based on surveys and administrative reporting [24, 25, 26].

Our analysis is based on the universe of mobile phone activity from Afghanistan's largest mobile phone operator for a 4-year period from April 2013 to March 2017. This dataset contains the anonymized mobile phone metadata of approximately 10 million mobile subscribers (a non-random subset of all individuals in Afghanistan, as we discuss in greater detail below). We separately obtain geo-coded information on approximately six thousand fatal violent events in Afghanistan, which are collected by the Uppsala Conflict Data Program from public media reports [27]. The spatial distribution of mobile phone towers and violent events can be seen in Figure 1.

The empirical approach we develop uses the mobile phone data to observe the movement of subscribers between regions, and a statistical model to estimate the causal effect of violence on this movement (see *Materials and Methods* for a detailed description). To summarize the main steps in this analysis: We first use metadata on the sequence of cell towers used by each mobile subscriber to identify each subscriber's home district (the smallest administrative unit in Afghanistan). We then adapt recent algorithmic advances in the measurement of migration from digital trace data [28] to identify days on which the subscriber's home district changes — we refer to this as a *migration*.[1] These individual migration events are then aggregated to produce estimates of the total population

[1] The International Organization of Migration (IOM) defines *migration* as "the movement of persons away from their place of usual residence, either across an international border or within a State" [29]. Our main analysis focuses on migrations that involve a person leaving their origin district for at least a week, in an effort to maximize recall.

flows between districts on each day. Finally, we use a high-frequency panel event study design to estimate the causal effect of violence on migration, which we measure as the increase in movement of subscribers out of a district impacted by violence, relative to movement from the same district on non-violent days, while controlling for seasonality and other temporal factors. We will frequently refer to this excess migration caused by violence as *displacement*, though we acknowledge that this statistical notion of displacement is more specific than that used by international organizations.[2]

The results presented below illustrate the new perspective on internal displacement that can be achieved through the analysis of population-scale digital trace data. However, this approach has important limitations. Some of these we can address (as discussed in *Materials and Methods*), but others are more fundamental [20, 30, 31, 32, 33]. We discuss a few key limitations — such as issues of population representativity, data access, and privacy — in the *Discussion* section, after describing the empirical findings.

Results

Violence in Afghanistan causes internal displacement (Figure 2). Among subscribers who were present on the day of a violent event, there is an immediate and statistically significant increase in the likelihood of leaving the district (Figure 2A). This increase peaks roughly 10 days after violence, when the odds of being observed in a different district are 4% higher than in the absence of violence. Violence-induced displacement is persistent: even 120 days after the violent event, subscribers who were present for the event are roughly 2% more likely to still be outside of the district. We show that these results are robust to some potential data and modeling issues (*Materials and Methods Section F*).

We also find evidence that displacement often anticipates violence. This is evident in Figure 2B shows the increase in the odds of an individual being in a different district relative to their location 30 days prior (see "k-day displacement" in *Materials and Methods*). Roughly 5 days before violent events are reported in the media, subscribers start to leave the impact regions. In the *Discussion*, we consider several possible explanations for this unexpected result.

The aggregate effects shown in Figure 2 mask substantial heterogeneity in how different types of violence cause different patterns of displacement. Most notably, violence involving the Islamic State (IS), while less frequent, causes significantly more displacement than violence involving the Taliban (Figure 3A). The difference is most pronounced in the immediate aftermath of the event, when the increase in displacement is 10 percentage points higher for violence involving IS. Such evidence is consistent with the fact that IS attacks are particularly brutal and frequently target

[2] For instance, the IOM defines *displacement* as "the movement of persons who have been forced or obliged to flee or to leave their homes or places of habitual residence, in particular as a result of or in order to avoid the effects of armed conflict, situations of generalized violence, violations of human rights or natural or human-made disasters." [29]

civilians; the Taliban have condemned IS attacks as "heinous" [34, 35].

High-casualty and high-frequency violence also have larger effects on displacement. Figure 3B shows that the highest casualty events (11 or more casualties, roughly equivalent to the 10% of events with the most casualties) cause more displacement at all periods following the event, relative to lower casualty events. Figure 3C indicates that violent events in regions that have recently experienced violence lead to more displacement than violence in regions that have experienced a period of relative peace. This result is perhaps surprising given prior work suggesting that people in chronically violent areas may acclimate to conflict [36, 37, 38]. However, taken in the context of the prolonged and pervasive Afghan conflict, we take this as evidence that people are more likely to flee their homes after recent and sustained violence, but may be willing to withstand 'idiosyncratic' violence.

We also find that the displacement response depends on whether the violence occurs in a provincial capital (which is typically an urban or peri-urban area) or a more rural non-capital district (Figure 3D). Although provincial capitals appear relatively resilient to violence, with muted and largely insignificant increases in displacement, effects of violence outside of provincial capitals are large and persistent. Afghanistan's 34 provincial capitals are regional seats of government, and are where the Afghan National Security Forces are typically concentrated [39]. As a result, they might be seen to offer relative safety compared to other outlying districts.

The heterogeneous displacement responses shown in Figures 3A-D highlight how each separate characteristic of violent events — the parties involved, the severity and recency of violence, and the location of the attack — relate to subsequent displacement. However, some of these characteristics are correlated: for instance, violence in provincial capitals tends to be preceded by fewer days of peace. For this reason, Figure 3E shows the joint relationship between these characteristics and displacement, i.e., it shows how each factor correlates with displacement, holding the other factors fixed (using a regression model; see *Materials and Methods*). Here, we observe that the general patterns are consistent with earlier results, but the outsized impact of IS is made clear: all else equal, violence involving IS has the largest and most significant impact on short- and long-term displacement.

Where do the displaced go?

Analysis of the anonymized mobile phone metadata can also provide granular insight into the destinations of the displaced. To build intuition, Figure 4 shows the flow of migrants in Afghanistan under normal times—that is, on days when violence does not occur. On such non-violent days, the total volume of mobile subscribers leaving capitals and non-capitals is approximately equal. For those moving from capitals, 73.5% move to a different province and roughly half (47.0%) move to other capitals or major cities; of the subscribers leaving non-capitals, half of them (50.6%) move to another province and 30.0% move to the provincial capital in the same province.

More revealing is how the equilibrium pattern of displacement shifts in response to violence. These results are summarized in Figure 5, which shows how the odds of moving to each type of district change on days with violence. Violence affecting non-capital regions (left panel of figure) makes subscribers more likely to go to the provincial capital of their origin district, and less likely to go to the largest cities (Kabul, Kandahar, Hirat, Mazari Sharif, and Jalalabad) outside their origin province. By contrast, violence affecting capital regions (right panel of figure) tends to drive subscribers away from their home province, and to concentrate either in the 5 major cities or more rural non-capitals. We discuss these and related results below.

Discussion

The preceding results illustrate a new empirical approach to studying conflict-induced displacement. The aggregate finding that violence causes displacement is consistent with prior work on internal displacement in fragile and conflict-affected countries [7, 8, 12]. In Afghanistan, the context of our analysis, IOM survey data from 2019 suggest that conflict had caused roughly two thirds of all current displacement [24].

The main innovation of this new approach is that it permits a granular, dynamic, and quantitative analysis of violence-induced displacement that would be difficult to accomplish using household surveys and qualitative interviews. For instance, our analysis documents the important role of provincial capitals in influencing both the impact of violence, as well as the destinations of displaced people. One specific finding is that when violence occurs in non-capitals, subscribers tend to flee to their home provincial capital. Such evidence corroborates qualitative findings that "people displaced by conflict and violence tended to try to stay as close as possible to their homes, moving from rural areas to the provincial capital or a neighboring province" [40]. The attraction of provincial capitals is likely due to several factors. First, capitals have a higher concentration of government security forces, and in the wake of violence, physical security is likely a crucial consideration. Further, provincial capitals are often the most urbanized area in the region, potentially offering greater economic opportunities [41]. Finally, movement to provincial capitals likely creates a feedback loop, where families are likely to have connections in provincial capitals, thus encouraging further movement to these capitals [42, 25, 43].

We also find that the types of violence associated with more displacement include those related to IS, high-casualty violence, as well as chronic violence. These can be explained by the level of risk perceived by individuals and households in influencing their decision to flee [7, 8]. Apart from indiscriminate attacks, IS has orchestrated vivid displays of violence, such as filmed executions, advertising its brutality and intimidating opponents [44]. The Taliban, by contrast, has support among some segments of society, and are seen by some as a legitimate governing force [45]. Separately, it is plausible that high-casualty and chronic violence similarly create an atmosphere of fear that leads to higher displacement. Casualties have been found to be associated with insurgent recruitment and violence [46]; its link to displacement is perhaps unsurprising.

Our analysis also indicates that people appear to anticipate the occurrence of violence, leaving before it occurs. This is most pronounced with recently experienced violence (*SI Appendix Figure S1*). There are several possible related explanations for this. First, people may not be anticipating a particular event, exactly, but responding to a general period of unrest. Individuals might perceive a threat of violence before a recorded event actually takes place.[3] For example, there might be skirmishes between armed forces that do not lead to fatalities or are not reported in the media. This is consistent with survey-based evidence [7, 8, 9], which find that the the perceived threat of violence or presence of armed forces is sufficient to cause displacement, independent from the actual exposure to violence.

While passively-collected mobile phone data create new possibilities for understanding forced displacement, there are important limitations and considerations (for more systematic reviews, see [20, 30, 31, 32, 33]). In particular, our measures of migration and displacement only cover inter-district movement of phones within the country of Afghanistan, thus likely underestimating the overall effect of violence on displacement, which includes international and within-district movements.

More generally, there are several issues related to the representativeness of the data and the potential biases that may result. Our data reflect the displacement patterns of mobile phone owners who have an active account on one specific commercial network, and not the full Afghan population. While mobility inferences of mobile phone owners have been shown to correlate with the mobility of non-owners in certain contexts [47], it is possible that violence induces different types of displacement among phone owners and non-owners. Specifically, the most vulnerable populations, such as women, children, and lower income individuals, tend to be underrepresented in mobile phone data, as they are less likely to own a phone [48, 49]. Related, the data only indicate the intermittent and approximate location of mobile phones, not of actual subscribers. When phones are shared, powered off or disused, this can introduce measurement error. Vulnerable populations may use their phones less often, resulting in larger errors for these users [50]. We address some of these issues through data processing, estimation methods, and robustness checks (see *Materials and Methods* and *SI Appendix Figure S3*), but cannot eliminate all such concerns.

Another obstacle to the widespread use of these methods to study internal displacement is that mobile phone data are not always easily accessible and may require partnerships with data owners. However, there are encouraging signs that private companies (such as Facebook, Google and Safegraph [51, 52]) may be interested in making mobility data available to researchers and humanitarian organizations [53, 30, 54]. Finally, the analysis of phone data must respect the privacy of

[3] This does not violate the causal identification assumption of no unobserved time-varying confounders (see *Materials and Methods*), since the perceived threat of violence does not cause violence to occur, and is thus not a confounder.

individual subscribers, particularly when dealing with displaced and otherwise vulnerable populations [55, 56]. Our analysis involves only anonymized data that is aggregated geographically (by district) and temporally (by day). More generally, [57] and [58] others provide broader frameworks for the privacy-conscientious use of mobile phone data.

Despite these important caveats, we remain optimistic that mobile phone and other digital trace data offer great potential for the study of internal displacement. Our empirical analysis provides new insight into the nature of violence-induced displacement in Afghanistan, and helps to quantify some of the human costs of violence that would be difficult to measure using traditional methods. For while there are definite limitations to what can be observed through mobile phone data, conflict-prone regions are often also the places where traditional survey-based data are least reliable and most difficult to obtain. Our hope is that this new approach can complement traditional perspectives on displacement, and eventually contribute to the design of effective policies for prevention and mitigation.

Materials and Methods

A Data on violence in Afghanistan

We obtain violent events data from the Uppsala Conflict Data Program,[4] a leading source of data on conflict events [59]. This open-source collection of metadata on armed conflict and organized violence is collected from media reports, so is likely to be biased toward salient events in more populous regions. The criteria for inclusion of an event is "the incidence of the use of armed force by an organized actor against another organized actor, or against civilians, resulting in at least one direct death in either the best, low or high estimate categories at a specific location and for a specific temporal duration" [60]. In Afghanistan from 2013 to 2017, 5,984 events were recorded where the event location is known to a district level, and event time is known to a specific day. We discard events which are not recorded with this level of precision (47% of all events). Afghanistan is divided into 398 districts in 34 provinces; our analysis is conducted on a district level.

B Mobile phone Call Detail Records (CDR)

Our analysis of displacement is based on a large dataset of pseudonymized mobile phone data from Afghanistan's largest mobile phone operator. We obtain Call Detail Records (CDR) that provide metadata for every mobile phone call and data packet transfer that occurred on this network between April 2013 to March 2017 — a total of roughly 20 billion events. For each such event, we observe a pseudonymized unique identifier for the subscriber (hashed from their phone number), the date and time of the event, as well as the identifier of the physical mobile phone tower through

[4] UCDP Georeferenced Event Dataset (GED) Global version 19.1 (`https://ucdp.uu.se/downloads/`)

which the transaction was routed. We also know the exact location of each tower, which allows us to approximately identify each subscriber's location at the time of the event, to within roughly 500 meters in urban areas and roughly 10km in rural areas.

In total, there are 13,315 active towers during this period, many of which are very close together; we group these towers into 1,439 tower groups by combining towers less than 100 meters apart. These cell tower groups are plotted in Figure 1. Only districts with cell towers are included in this analysis, though we note that there is substantial overlap between districts with violence and districts with cell towers (84% of the events in Figure 1 occur in a district with a cell tower). These generally correspond to the more populated districts in Afghanistan.

C Measuring migration

From the original call detail records, we follow a sequence of steps to determine if and when a migration event occurs. We adopt the International Organization for Migration (IOM)'s definition of *migration*, which is "The movement of persons away from their place of usual residence, either across an international border or within a State" [29], and focus on internal migration ("within a State"), where place of usual residence is measured to a district-level precision. We capture trips that last approximately at least a week (five full days and two travel days). The migration that we measure is therefore an inter-district movement. Complete details on this process are in the *SI Appendix*; a brief summary is provided here.

Our first step is to derive a "daily modal location" for each subscriber for each day (24-hour period from 6 A.M. to 6 A.M. the next day), which is the district in which the subscriber is observed most frequently on that day. Similar methods have been used and validated in other work [61, 62, 63, 64, 13, 31]. Some of this work uses night-time hours; we use all hours, which allows the inclusion of more users. For example, comparing locations of each user each day (user-day) in April 2013, data are available for approximately 31 million user-days using night-time hours (6 P.M. to 6 A.M.), while using all hours (6 A.M. to 6 A.M.), 61 million user-days are available. Among the 32 million, 89% record the same district as using all hours. Now, these daily modal locations tend to be sparse and noisy: for instance, many people do not use their phones on every single day; people may take short trips to nearby (non-residential) locations, and so forth.

Thus, our second step employs an unsupervised scanning algorithm [28] to identify contiguous segments in which a subscriber is, with high probability, resident in a single district. Contiguous segments with sparse data are interpolated within the bounds of input parameters.

The third step is to identify migration events using discontinuous breaks in these contiguous segments. The second and third steps use the open-source `migration_detector` Python package,[5] which is specifically designed to infer migration events in transaction log data. The accompanying

[5] `https://github.com/g-chi/migration_detector`

paper [28] validates the use of these methods to measure migration. Tuning parameters are set to identify changes in locations resulting from stays of at least five full days in origin and destination districts. See *SI Appendix* for full details.

The above procedures allow us to measure migration events from the mobile phone CDR. Many such events are not indicative of *displacement*, which the IOM defines as "The movement of persons who have been forced or obliged to flee or to leave their homes or places of habitual residence, in particular as a result of or in order to avoid the effects of armed conflict, situations of generalized violence, violations of human rights or natural or human-made disasters." [29] Given the limited contextual information available in the CDR, we cannot directly observe whether each inferred migration event should be considered a displacement. Instead, as we discuss in *Materials and Methods Section E*, we focus our analysis on the increase in out-migrations from a district that appear to be caused by violence in that district.

D Data validation

To validate the measures of migration derived from the mobile phone CDR, we compare our derived *migration* metrics to *displacement* measures published by the IOM.[6] To our knowledge, there is no official or other published data measuring inter-district migration as we do; while we try as far as possible to produce analogous measures, the IOM data measure fundamentally different quantities, and we do not expect comparisons to be identical. Generally speaking, we might expect province shares of *migration* and *displacement* to be similar if the fraction of displaced people among those who move for any reason are similar across provinces. This might not always be the case, for example, we might expect the capital Kabul to have a much smaller share of displaced people. Nevertheless, we make the comparison, as the IOM data are the closest published dataset on internal migration or displacement in Afghanistan.

The IOM collects data at the settlement (village) level through key informant interviews, focus group discussions, and direct observation [24]. They use these data to estimate counts of outgoing and incoming internally displaced persons (IDPs) in assessed settlements over fixed periods. IDPs are categorized into "returnee IDPs," "arrival IDPs," and "fled IDPs." We use the data collected in the year 2016; to our knowledge, this means that these individuals were recorded as being IDPs anytime during the year. We group "returnee" and "arrival" IDPs together as incoming IDPs, treat "fled IDPs" as outgoing IDPs, and sum the total number of incoming and outgoing IDPs for each province. We then compute each province's share of the total incoming and outgoing IDPs.

Now, to construct an analogous metric from the CDR, we compare the district locations of each subscriber at the beginning and end of three four-month periods in 2016 (Jan-Apr, Apr-Aug, and Aug-Dec; summed to obtain a measure of movement in 2016, since the longest we track users is

[6] Available at `https://data.humdata.org/dataset/afghanistan-displacement-data-baseline%2Dassessment-iom-dtm`.

for 120 days). Since each district could have different cellphone penetration rates, for each period and each district, we estimate the total number of people who moved in and out of the district, by scaling the number of recorded subscribers who moved by $\frac{\texttt{district population}}{\texttt{\#recorded subscribers}}$, where the district population is as estimated by Afghanistan's Central Statistical Office.[7] We then aggregate these to the province level for 2016, and compute province shares in a similar manner. Figure 6A shows the share of each province estimated to leave; Spearman's correlation between CDR and IOM statistics at the province level is $\rho = .49$. Figure 6B does the same for incoming individuals, with $\rho = .56$.

Indeed, in Figure 6 we see that many provinces see similar shares of migration and displacement, with some obvious differences in Kabul province, where migration far exceeds displacement, and Hilmand, where displacement far exceeds migration (Hilmand province is a Taliban stronghold and frequently sees heavy fighting [65]).

E Panel Regressions: Measuring k-day displacement

We combine the violent events data and migrations observed in the CDR into a district-day panel dataset, which we use to estimate the "average" impact of violence on out-migration from the impacted district. We estimate this effect by adapting widely used panel regression models to our context (e.g., [66]), which allows us to estimate the total displacement caused by violence, while controlling for unobserved district- and time-related factors that might influence both the occurrence of violence and displacement, and subject to the identifying assumptions discussed below. This estimation of displacement as a increase in migration due to violence also partially addresses the concern that place of usual residence might be incorrectly measured using CDR. If violence does not impact the measurement error, for example if the likelihood of a subscriber being misallocated to the district of their workplace instead of their home does not change due to violence, then the misallocation will not bias the estimated displacement.

For each value of k from 1 to 120, we estimate the following regression:

$$g(\mathbb{E}(Y_{dt,k}|X_{dt}, T_{d,t+\tau})) = \gamma_d + \lambda_t + \sum_{\tau=-30}^{180} \beta_\tau T_{d,t+\tau} \tag{1}$$

where d indexes the district, t the time (calendar date), covariates X_{dt} are given by γ_d, district fixed effects, and λ_t, time fixed effects. $T_{d,t+\tau}$ are the treatment variables (whether or not violence occurs) for district d at time t, at a lag of τ days. Lags of $\tau \in [-30, 180]$ are used, representing violence in the district 30 days in the future to 180 days in the past. The outcome variable, $Y_{dt,k}$, is the proportion of those in district d at time $t - k$, that are in a different district at time t. Subscribers present k days ago in district d but with missing location on day t are included in the denominator but not in the numerator in this computation. The parameter k is introduced

[7] Population data available at `https://data.humdata.org/dataset/estimated-population-of-afghanistan-2015-2016`

to capture the fact that displacement has to be measured relative to some time in the past. $g()$ is the logit link function. Since the outcome variable is a proportion, we model it using a beta distribution, a family of continuous distributions in the interval from 0 to 1, taking a variety of possible shapes depending on the values of its parameters. We fit a beta regression using maximum likelihood estimation [67]. Standard errors are clustered at a district level.

These coefficients can be interpreted as with a logistic regression: for each τ, e^{β_τ} is the multiplicative change in odds of being in different district today (time t), for $T_\tau = 1$ (when violence occurs) relative to $T_\tau = 0$ (days without violence), holding the other variables constant. To interpret β_τ as the causal effect of violence on displacement, the target parameter is the causal conditional odds ratio, and the set of necessary identification assumptions are positivity, consistency, conditional exchangeability and correct model specification [68]. In our context, this specification assumes that there are no spatial spillovers, meaning that violence in one district does not have an effect on displacement in other districts. Carryover effects of the violence are limited to 180 days after the violence, and effects from up to 30 days prior are allowed. These daily effects are estimated independently and do not modify one another. The effects are also assumed to be identical for all districts, and do not vary over the measurement period (2013-2017). The confounders are limited to district, time, and treatment in the surrounding window of time, and these enter additively. This implies that there are no unobserved time-varying confounders, and also that past outcomes do not affect current treatment (this is plausible since in most cases the number of displaced people is not large enough to affect military strategy). We relax several of these assumptions in subsequent analysis, for instance by allowing for heterogeneous effects for different types of violence in different types of locations.

In estimating these regressions, we exclude district-days in which the outcome variable is 0, 1, or missing. The rationale is that these zeros and ones are likely due to data sparsity. For instance, if no subscribers were recorded as being in a different district, it could be that their locations were simply missing (e.g., they did not use their phones, there was no cell service, they switched providers, etc.). On the other hand, it is unlikely that all subscribers would have left a district on any day; a recorded 1 could indicate cell tower outages in the origin district, for example.

F Impact of a violent day

To distill the impact of a single violent day, for each $k \in [1, 120]$ we consider the coefficient for T_τ for $\tau = k$. This coefficient captures the effect of violence occurring at a τ day lag, on movement measured at time t, compared to district locations k days ago. When $\tau = k$, the outcome variable is measured with respect to those in the district on the day of the violence. In this way, extracting the relevant coefficients from regressions where the outcome variable are different values of k, gives us the impact of a single violent day, on the subscribers in the district on that day. We demonstrate the robustness of these results to potential data issues, such as the presence of outliers, as well as modeling issues such as the inclusion of additional time-varying controls (*SI Appendix Figure S3-4*).

G Heterogeneous effects

To allow for the possibility that the displacement response may differ for different types of violence, or for specific types of locations, the results of heterogeneous effects models are shown in Fig. 3. These results are estimated by creating separate treatment indicators for different types of events (e.g., low-casualty vs. high-casualty), which replace the treatment indicators in (1). For instance, letting $H_{d,t+\tau}$ denote the occurrence of high-casualty (>10 casualties) violence, and $L_{d,t+\tau}$ denote the occurrence of low-casualty violence, we estimate:

$$\begin{aligned} &g(\mathbb{E}(Y_{dt,k}|X_{dt}, H_{d,t+\tau}, L_{d,t+\tau})) \\ &= \gamma_d + \lambda_t + \sum_{\tau=-30}^{180} \beta_{H,\tau} H_{d,t+\tau} \;\; + \sum_{\tau=-30}^{180} \beta_{L,\tau} L_{d,t+\tau} \end{aligned} \tag{2}$$

When analyzing heterogeneity of response by location (e.g., for provincial capitals), we estimate prior regressions on the relevant subsets of the data, i.e., by only including observations pertaining to provincial capitals.

H Controlling for multiple dimensions of heterogeneity

To account for multiple dimensions of heterogeneity varying jointly, we analyze 30-day displacement by first fitting (3) *separately for each of the events*, using ordinary least squares.

$$log\left(\frac{Y_{dt,30}}{1-Y_{dt,30}}\right) = \gamma_d + \lambda_t + \sum_{\tau=-30}^{180} \beta_\tau T_{d,t+\tau} + \epsilon_{dt} \tag{3}$$

Here $T_{d,t+\tau}$ indicates a single event at a time (each treatment indicator indicates whether or not the specific event occurs at district d at time t, at a lag of τ days). Only events in which all β_τ coefficients can be estimated are included, meaning that if the outcome variable is unavailable in any day that is 30 days preceding the event to 180 days after the event, it is not included in the analysis. This results in a total of 2359 events being studied. For each included event, we take the mean of the estimated coefficients for β_τ, for τ=1-15, 16-30, 31-45, 46-60, 61-75 and 76-90. We treat these as outcome variables, and model each of these derived outcomes O_i as

$$\begin{aligned} O_i = \beta_0 + \beta_1 \texttt{provCap}_i + \beta_2 \texttt{log(population)}_i \\ + \beta_3 \texttt{IS}_i + \beta_4 \texttt{casualties11}_i + \beta_5 \texttt{peace60}_i + \epsilon_i \end{aligned} \tag{4}$$

where i is the event, $\texttt{provCap}_i$ is a binary variable denoting whether the event occurs in a provincial capital, $\texttt{log(population)}_i$ is log of the population of the district in which the event occurs (added as a control), $\texttt{IS}_i$ is a binary variable denoting whether the event involved the Islamic State, $\texttt{casualties11}_i$ is a binary variable denoting whether the event was associated with 11 or more casualties, and $\texttt{peace60}_i$ is a binary variable denoting whether the event was preceded by 60 or more days of peace. Figure 3E shows the estimated coefficients for each of the outcomes.

I Destinations of displaced people

To investigate where the individuals displaced by violence go, we first examine migrant flows during non-event days (Figure 4) and event days (*SI Appendix Figure S2*). We consider all recorded moves in any 30-day period, and split these into days in which violent events occurred at the start of the 30-day period ("event days"), and those which no events were recorded ("non-event days"). We repeat the following analysis for each. First, we categorize recorded moves as originating in either capital districts or non-capital districts. Then, destination districts are categorized into mutually exclusive categories by first recording whether they are in the same or different province from the origin district; these destinations are then partitioned into three different types of districts – the major urban cities (Kabul, Kandahar, Hirat, Mazari Sharif, Jalalabad), other capital districts, or non-capital districts.

To estimate the effect of violence on the destination of displacement, we use a similar setup as (1). Instead of the outcome variable being the fraction of population moved on days k, we use the fraction of *movers* (those in a different district at time t compared to k days ago) on days k, observed to be at specific types of destination districts, as described above. We use outcomes for $k = 7, 30, 90$, and fit separate regressions for provincial capitals, and non-capitals, and for each outcome. As before, district-days in which the outcome variable is 0, 1, or missing are excluded from the analysis (e.g., if no subscribers were recorded as being in a different district, it could be that their locations were simply missing, or that the mobile phone operator did not have coverage in the specific destination district).

Data availability

The data that provides the foundation for the paper is a dataset containing detailed information on roughly 20 billion mobile phone calls in the country of Afghanistan. These data contain proprietary and confidential information belonging to a major telecommunications operator. We are hence unable to publicly release the full dataset. Instead, several template datasets that contain a small sample of the simulated data are available from the authors upon reasonable request. Information on how someone else could obtain the proprietary data from the mobile phone operator in question can also be provided. All other data used are publicly available and sources have been listed in the text.

Code availability

All code to reproduce the findings of this study is available at `https://github.com/shikharmehra/afghanistan-internal-displacement`.

References

[1] United Nations. Global Trends: Forced Displacement in 2019 (United Nations High Commissioner for Refugees, 2020).

[2] Blattman, C. & Miguel, E. Civil War. Journal of Economic Literature **48**, 3–57 (2010). URL https://pubs.aeaweb.org/doi/10.1257/jel.48.1.3.

[3] Corral, P., Irwin, A., Krishnan, N., Mahler, D. G. & Vishwanath, T. Fragility and conflict: on the front lines of the fight against poverty (World Bank Group, Washington, 2020).

[4] Azose, J. J. & Raftery, A. E. Estimation of emigration, return migration, and transit migration between all pairs of countries. Proceedings of the National Academy of Sciences **116**, 116–122 (2019).

[5] LUCAS, R. E. Internal migration in developing economies: An overview of recent evidence. Geopolitics, History, and International Relations **8**, 159–191 (2016). URL https://www.jstor.org/stable/26805968.

[6] Steele, A. Seeking safety: Avoiding displacement and choosing destinations in civil wars. Journal of Peace Research **46**, 419–429 (2009).

[7] Adhikari, P. Conflict-Induced Displacement, Understanding the Causes of Flight: *CONFLICT-INDUCED DISPLACEMENT*. American Journal of Political Science **57**, 82–89 (2013). URL http://doi.wiley.com/10.1111/j.1540-5907.2012.00598.x.

[8] Engel, S. & Ibáñez, A. Displacement due to violence in colombia: A household-level analysis. Economic Development and Cultural Change **55**, 335–365 (2007). URL http://www.jstor.org/stable/10.1086/508712.

[9] Schon, J. Motivation and opportunity for conflict-induced migration: An analysis of Syrian migration timing. Journal of Peace Research **56**, 12–27 (2019). URL `http://journals.sagepub.com/doi/10.1177/0022343318806044`.

[10] Ball, P. & Asher, J. Statistics and Slobodan: Using Data Analysis and Statistics in the War Crimes Trial of Former President Milosevic. CHANCE **15**, 17–24 (2002). URL `http://www.tandfonline.com/doi/full/10.1080/09332480.2002.10554820`.

[11] Schon, J. Focus on the Forest, Not the Trees: A Changepoint Model of Forced Displacement. Journal of Refugee Studies **28**, 437–467 (2015). URL `https://academic.oup.com/jrs/article-lookup/doi/10.1093/jrs/feu037`.

[12] Melander, E. & Öberg, M. The threat of violence and forced migration: Geographical scope trumps intensity of fighting. Civil Wars **9**, 156–173 (2007).

[13] Deville, P. et al. Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences **111**, 15888–15893 (2014).

[14] Lu, X., Bengtsson, L. & Holme, P. Predictability of population displacement after the 2010 haiti earthquake. Proceedings of the National Academy of Sciences **109**, 11576–11581 (2012).

[15] Sutton, J. et al. Warning tweets: Serial transmission of messages during the warning phase of a disaster event. Information, Communication & Society **17**, 765–787 (2014).

[16] Wesolowski, A. et al. Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. Proceedings of the National Academy of Sciences **112**, 11114–11119 (2015).

[17] Maas, P. Facebook Disaster Maps: Aggregate Insights for Crisis Response & Recovery. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 3173–3173 (ACM, Anchorage AK USA, 2019). URL `https://dl.acm.org/doi/10.1145/3292500.3340412`.

[18] Li, T., Dejby, J., Albert, M., Bengtsson, L. & Lefebvre, V. Detecting individual internal displacements following a sudden-onset disaster using time series analysis of call detail records. arXiv preprint arXiv:1908.02377 (2019).

[19] Finger, F. et al. Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. Proceedings of the National Academy of Sciences **113**, 6421–6426 (2016).

[20] Salah, A., Pentland, A., Lepri, B. & Letouzé, E. Guide to Mobile Data Analytics in Refugee Scenarios (Springer, 2019).

[21] Mazzoli, M. et al. Migrant mobility flows characterized with digital data. PloS one **15**, e0230264 (2020).

[22] Pastor-Escuredo, D., Imai, A., Luengo-Oroz, M. & Macguire, D. Call detail records to obtain estimates of forcibly displaced populations. In Guide to Mobile Data Analytics in Refugee Scenarios, 29–52 (Springer, 2019).

[23] Singh, L. et al. Blending noisy social media signals with traditional movement variables to predict forced migration. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1975–1983 (2019).

[24] International Organization for Migration. Baseline mobility assessment summary results iom afghanistan round 9 (2019). URL `https://displacement.iom.int/system/tdf/reports/IOM-Afghanistan-Baseline-Mobility-Assessment-Summary-Results-December-2019-English_0.pdf?file=1&type=node&id=9016`.

[25] Marchand, K. et al. Afghanistan migration profile (2014).

[26] Internal Displacement Monitoring Centre. Afghanistan displacement associated with conflict and violence: Figure analysis -grid 2020 (2020). URL `https://www.internal-displacement.org/sites/default/files/2020-04/GRID%202020%20-%20Conflict%20Figure%20Analysis%20-%20AFGHANISTAN.pdf`.

[27] Pettersson, T. & Öberg, M. Organized violence, 1989–2019. Journal of peace research **57**, 597–613 (2020).

[28] Chi, G., Lin, F., Chi, G. & Blumenstock, J. A general approach to detecting migration events in digital trace data. PloS one **15**, e0239408 (2020).

[29] International Organization for Migration. Glossary on Migration (International Organization for Migration, 2019).

[30] on Big Data for Official Statistics, U. G. W. G. Handbook on the Use of Mobile Phone Data for Official Statistics. 9 (UN, 2019).

[31] Vanhoof, M., Reis, F., Ploetz, T. & Smoreda, Z. Assessing the quality of home detection from mobile phone data for official statistics. arXiv preprint arXiv:1809.07567 (2018).

[32] Grassini, L. & Dugheri, G. Mobile phone data and tourism statistics: a broken promise? National Accounting Review **3**, 50–68 (2021).

[33] Tiru, M. Overview of the sources and challenges of mobile positioning data for statistics. In International Conference on Big Data for Official Statistics. United Nations Statistics Division (UNSD) and Nation 28–30 (2014).

[34] Joscelyn, T. The taliban is still the main driver of violence in afghanistan (2020). URL https://vitalinterests.thedispatch.com/p/the-taliban-is-still-the-main-driver.

[35] Aggarwal, N. K. Exploiting the islamic state-taliban rivalry for counterterrorism messaging. Journal of Policing, Intelligence and Counter Terrorism **12**, 1–15 (2017).

[36] Allen, L. Getting by the occupation: How violence became normal during the second palestinian intifada. Cultural Anthropology **23**, 453–487 (2008).

[37] Nossiter, A. 'there is no safe area': In kabul, fear has taken over. The New York Times **1** (2021). URL https://www.nytimes.com/2021/01/17/world/asia/afghanistan-kabul-taliban-violence.html.

[38] Margold, J. A. From 'cultures of fear and terror'to the normalization of violence: An ethnographic case. Critique of Anthropology **19**, 63–88 (1999).

[39] McFate, J. L., Denaburg, R. & Forrest, C. Afghanistan threat assessment: The taliban and isis. Institute for the Study of War (ISW) Backgrounder **10** (2015).

[40] Internal Displacement Monitoring Centre. Afghanistan patterns of displacement (2020). URL https://www.internal-displacement.org/countries/afghanistan.

[41] Hakim, A. A. & Boz, I. Influencing factors of rural families' migration to urban area: The case of kabul, afghanistan. Asian Journal of Agricultural Extension, Economics & Sociology 1–10 (2019).

[42] Weiss, Z. Environmental factors driving migrants into afghanistan's cities (2012). URL https://www.stimson.org/2012/environmental-factors-driving-migrants-afghanistans-cities/.

[43] Clark, W. A. & Lisowski, W. Extending the human capital model of migration: The role of risk, place, and social capital in the migration decision. Population, Space and Place **25**, e2225 (2019).

[44] Ibrahimi, N. & Akbarzadeh, S. Intra-jihadist conflict and cooperation: Islamic state–khorasan province and the taliban in afghanistan. Studies in Conflict & Terrorism **43**, 1086–1107 (2020).

[45] LYALL, J., BLAIR, G. & IMAI, K. Explaining support for combatants during wartime: A survey experiment in afghanistan. American Political Science Review **107**, 679–705 (2013).

[46] Condra, L. N., Felter, J. H., Iyengar, R. K. & Shapiro, J. N. The effect of civilian casualties in afghanistan and iraq. Tech. Rep., National Bureau of Economic Research (2010).

[47] Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W. & Buckee, C. O. The impact of biases in mobile phone ownership on estimates of human mobility. Journal of the Royal Society Interface **10**, 20120986 (2013).

[48] Sekara, V. et al. Mobile phone data for children on the move: Challenges and opportunities. In Guide to Mobile Data Analytics in Refugee Scenarios, 53–66 (Springer, 2019).

[49] Blumenstock, J. E. & Eagle, N. Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda. In Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and 6–12 (ACM, 2010). URL `http://dl.acm.org/citation.cfm?id=2369225`.

[50] Leo, Y., Fleury, E., Alvarez-Hamelin, J. I., Sarraute, C. & Karsai, M. Socioeconomic correlations and stratification in social-communication networks. Journal of The Royal Society Interface **13**, 20160598 (2016).

[51] Aktay, A. et al. Google covid-19 community mobility reports: Anonymization process description (version 1.0). arXiv preprint arXiv:2004.04145 (2020).

[52] Ilin, C. et al. Public mobility data enables covid-19 forecasting and management at local and global scales. Tech. Rep., National Bureau of Economic Research (2020).

[53] Blondel, V. D. et al. Data for development: the d4d challenge on mobile phone data. arXiv preprint arXiv:1210.0137 (2012).

[54] Oliver, N. et al. Mobile phone data for informing public health actions across the covid-19 pandemic life cycle (2020).

[55] Taylor, L. No place to hide? the ethics and analytics of tracking mobility using mobile phone data. Environment and Planning D: Society and Space **34**, 319–336 (2016).

[56] Blumenstock, J. Don't forget people in the use of big data for development (2018).

[57] de Montjoye, Y.-A. et al. On the privacy-conscientious use of mobile phone data. Scientific data **5**, 1–6 (2018).

[58] Mayer, J., Mutchler, P. & Mitchell, J. C. Evaluating the privacy properties of telephone metadata. Proceedings of the National Academy of Sciences **113**, 5536–5541 (2016).

[59] Eck, K. In data we trust? a comparison of ucdp ged and acled conflict events datasets. Cooperation and Conflict **47**, 124–141 (2012).

[60] Sundberg, R. & Melander, E. Introducing the ucdp georeferenced event dataset. Journal of Peace Research **50**, 523–532 (2013).

[61] Phithakkitnukoon, S., Calabrese, F., Smoreda, Z. & Ratti, C. Out of sight out of mind–how our mobile social network changes during migration. In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Internationa 515–520 (IEEE, 2011).

[62] Toomet, O., Silm, S., Saluveer, E., Ahas, R. & Tammaru, T. Where do ethno-linguistic groups meet? how copresence during free-time is related to copresence at home and at work. PLoS One **10**, e0126093 (2015).

[63] Hong, L., Wu, J., Frias-Martinez, E., Villarreal, A. & Frias-Martinez, V. Characterization of internal migrant behavior in the immediate post-migration period using cell phone traces. In Proceedings of the tenth international conference on information and communication technologies and developmen 1–12 (2019).

[64] Hankaew, S. et al. Inferring and Modeling Migration Flows Using Mobile Phone Network Data. IEEE Access **7**, 164746–164758 (2019). URL `https://ieeexplore.ieee.org/document/8895964/`.

[65] Mashal, M. Marines return to helmand province for a job they thought was done. The New York Times **4** (2017). URL `https://www.nytimes.com/2017/04/29/world/asia/marines-return-to-helmand-province-for-a-job-they-thought-was-done.html`.

[66] Angrist, J. D. & Pischke, J.-S. Mostly harmless econometrics: an empiricist's companion (Princeton University Press, Princeton, 2009). OCLC: ocn231586808.

[67] Ferrari, S. & Cribari-Neto, F. Beta regression for modelling rates and proportions. Journal of applied statistics **31**, 799–815 (2004).

[68] Hernán, M. A. & Robins, J. M. Causal inference: what if. Boca Raton: Chapman & Hill/CRC **2020** (2020).

Acknowledgements

We are grateful for helpful conversations with Michael Callen, Oeindrila Dube, Tarek Ghani, Sham Kakade, and Jacob Shapiro. Guanghua Chi, Rohan Gajendragadkar, Shekhar Mittal, and Musa Jalis provided excellent research assistance on this project. This material is based upon work supported by the National Science Foundation under Grants IIS-1942702 and CCF-1637360, and the Office of Naval Research under award N00014-17-1-2313. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Office of Naval Research.

Author contributions statement

X.H.T., S.M., and J.E.B. designed research, performed research, analyzed data, and wrote the paper.

Additional information

The authors declare no competing interest.

Figures

Figure 1: Map of cell towers and violent events in Afghanistan, 2013-2017. District boundaries are represented by gray lines. There are 1439 cell tower groups (black dots) and 5984 violent events (red dots). Event locations are marked by exact geocoordinates when available, or with the district centroid when only the district name is known.

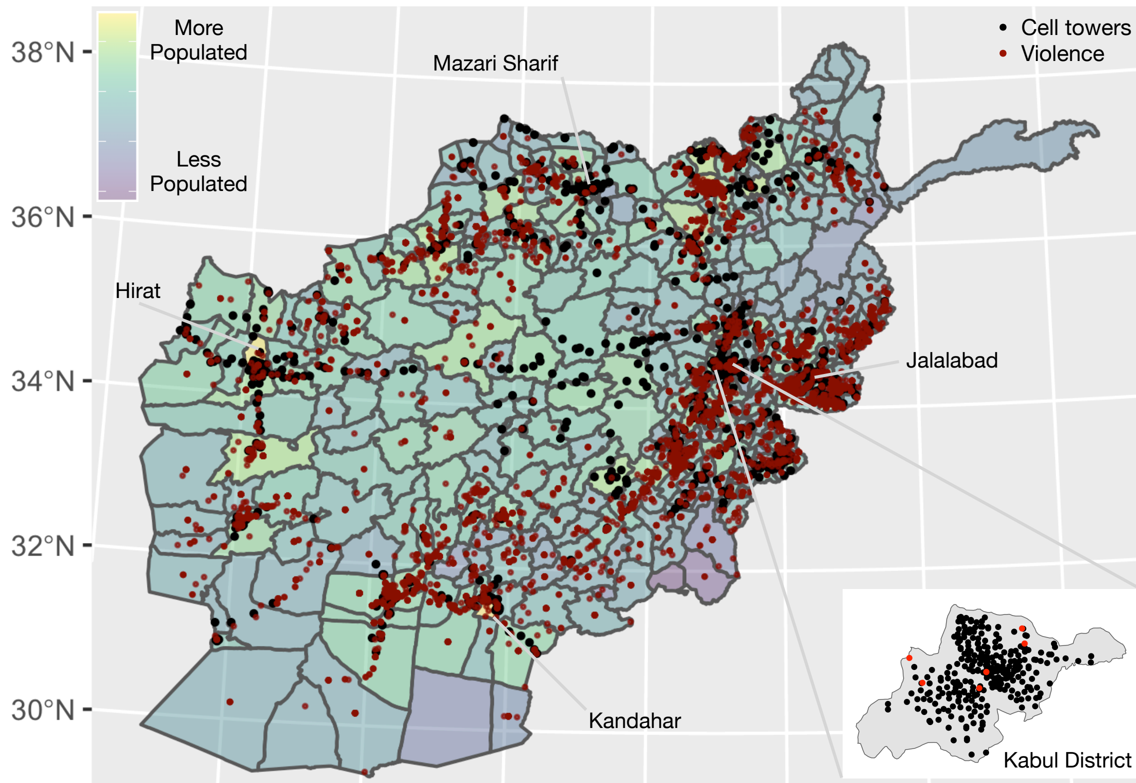
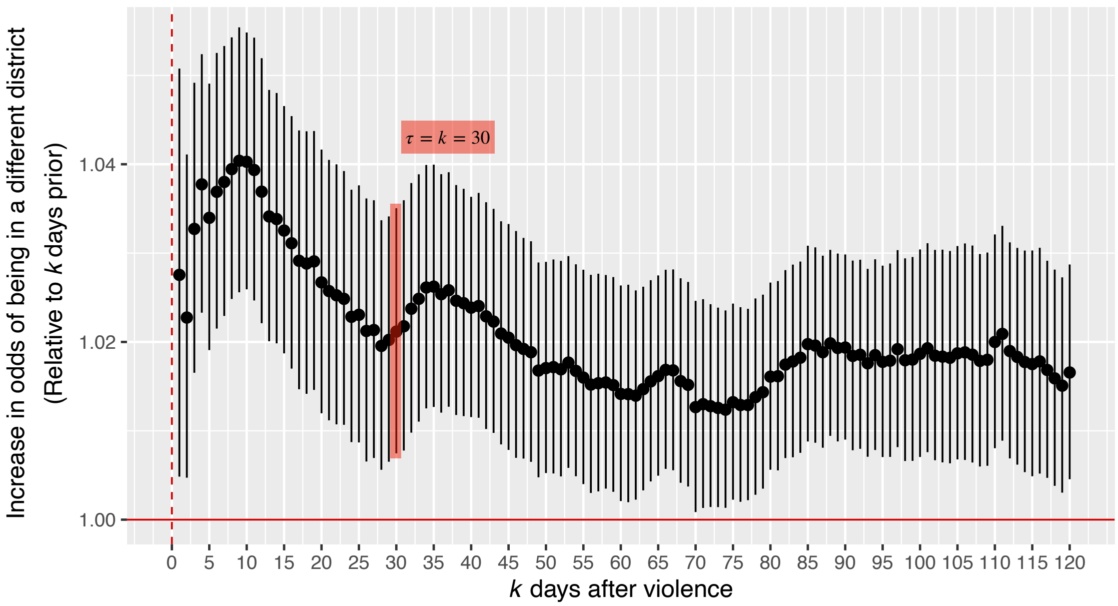
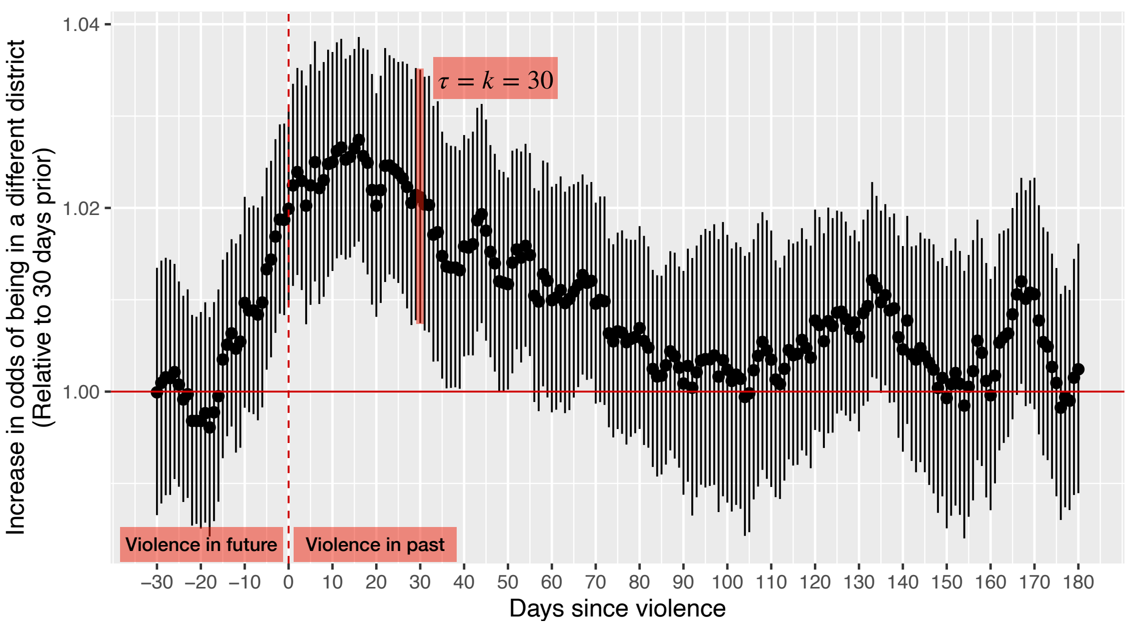


Figure 2: The effect of violence on internal displacement.

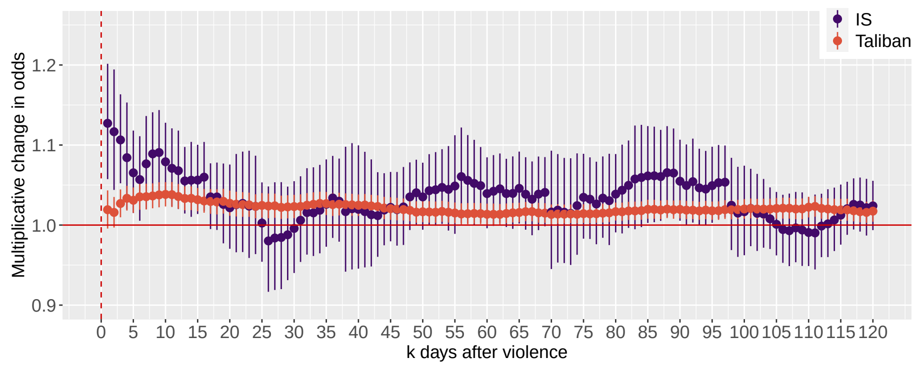


(a) Displacement for those present on the day of the violent event. Exponentiated regression coefficients, plotted on the y-axis, indicate the increase in the odds for individuals who were in the district on the day of violence, of being in a different district k days later. Bars indicate 95% confidence intervals. Estimates are based on 3354 violent events.

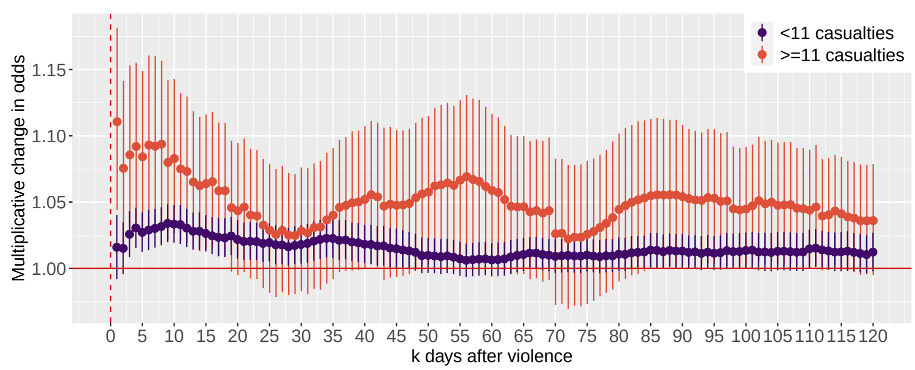


(b) 30-day displacement. Exponentiated regression coefficients indicate the increase in the odds that individuals are in a different district than they were 30 days prior. Negative x-axis values correspond to days preceding violence. Bars indicate 95% confidence intervals. Estimates are based on 3354 violent events.

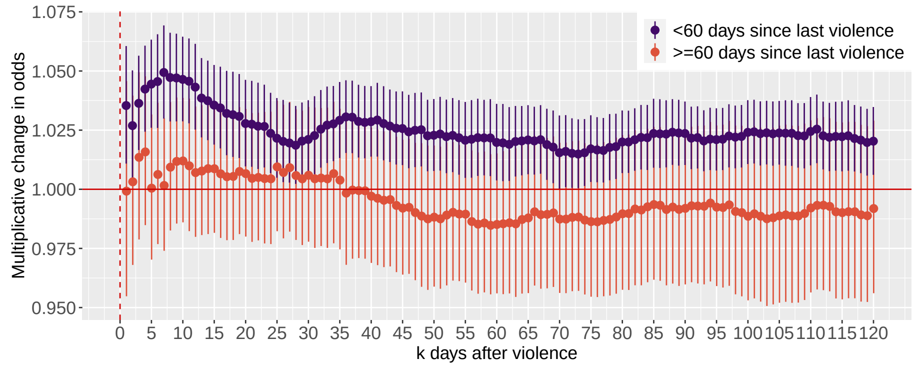
Figure 3: The effect of violence on displacement, disaggregated by type and location. Figures A-D show the impact of a violent day on individuals who were in the district on the violent day. Coefficients are estimated separately for (A) IS vs. Taliban violence (B) Events with 11 or more casualties vs. fewer than 11 casualties (C) Events following fewer than 60 days of peace vs. 60 or more days (D) Events in provincial capitals vs. non-capitals. (E) shows how the effect of each event on 30-day displacement depends on event and location characteristics. Bars indicate 95% confidence intervals.



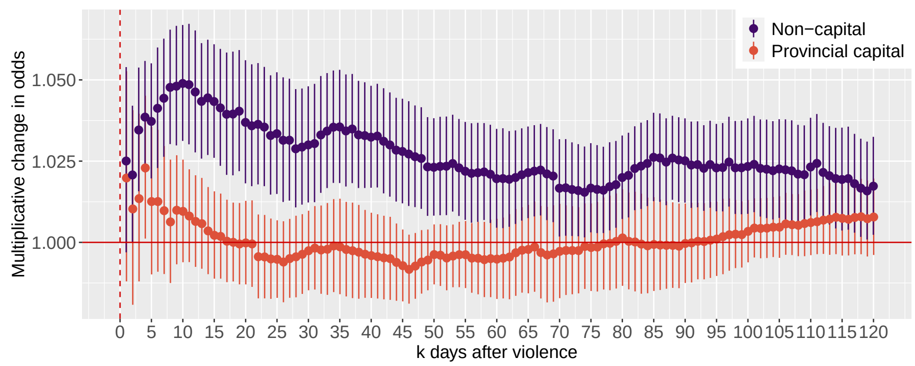
(a) Displacement for IS (N=185) vs. Taliban violence (N=3134).



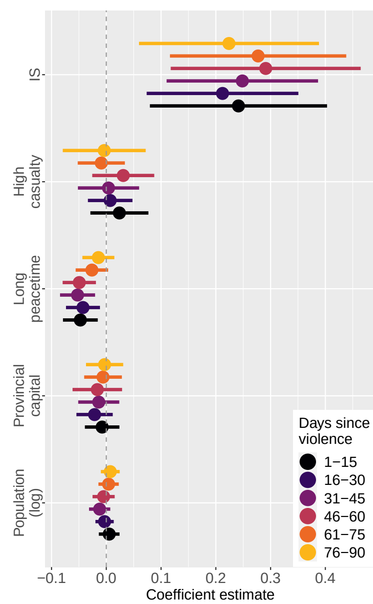
(b) Displacement for events with 11 or more casualties (N=397) vs. fewer than 11 casualties (N=2957).



(c) Displacement for events following fewer than 60 days of peace (N=2319) vs. 60 or more days (N=1035).



(d) Displacement for events in provincial capitals (N=2460) vs. non-capitals (N=894).



(e) How event and location characteristics affect the impact of each event on 30-day displacement. The different colors denote different outcome variables: the average effect of violence when it occurred 1-15, 16-30, 31-45, 46-60, 61-75, and 76-90 days in the past.

Figure 4: Migrant flows on days without violence. Diagram indicates the proportion of subscribers moving between locations of different types, where a move is defined as a change of home district over a 30-day period.

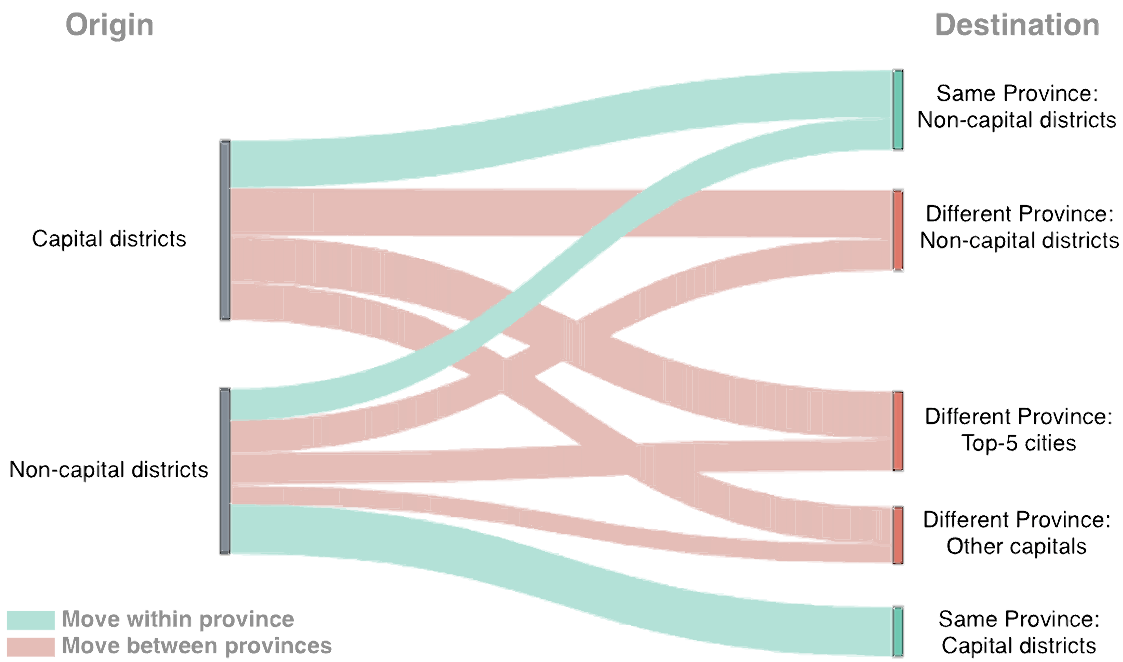


Figure 5: The effect of violence on destination choice. Figure shows how movement from districts on days of violence differs from movement from districts on days without violence, where a move is defined as being in a different district 7, 30, or 90 days after the reference date (see *Materials and Methods*). Left panel shows movement from non-capitals; right panel shows movement from capitals. Red denotes movement between provinces and green denotes movement within a province. Destinations are divided by the type of district: one of the five largest cities (Kabul, Kandahar, Hirat, Mazari Sharif, Jalalabad); another of the provincial capital districts; or a non-capital district. Bars indicate 95% confidence intervals.

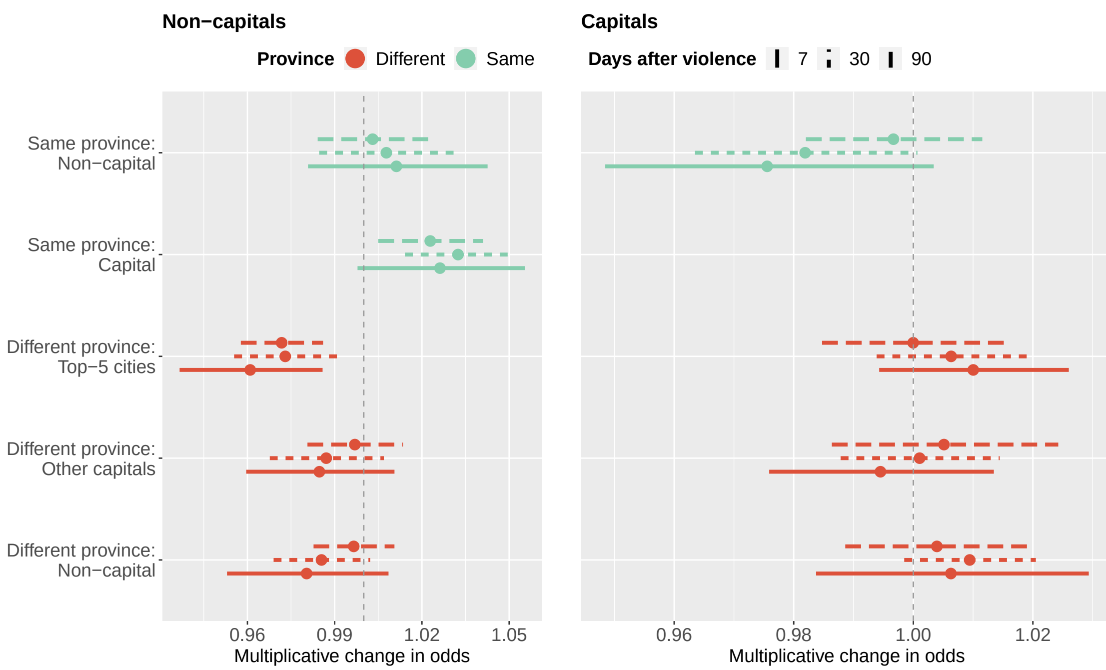
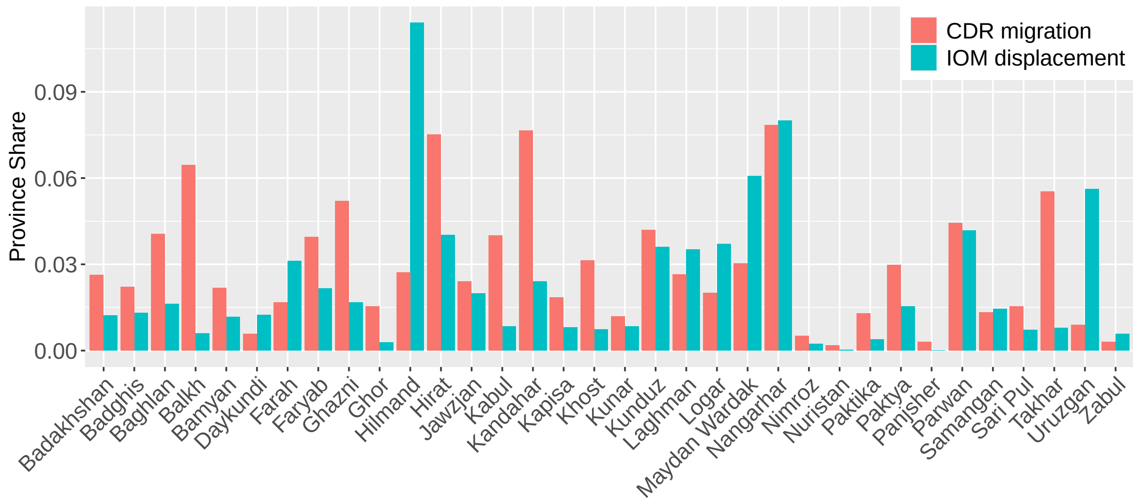
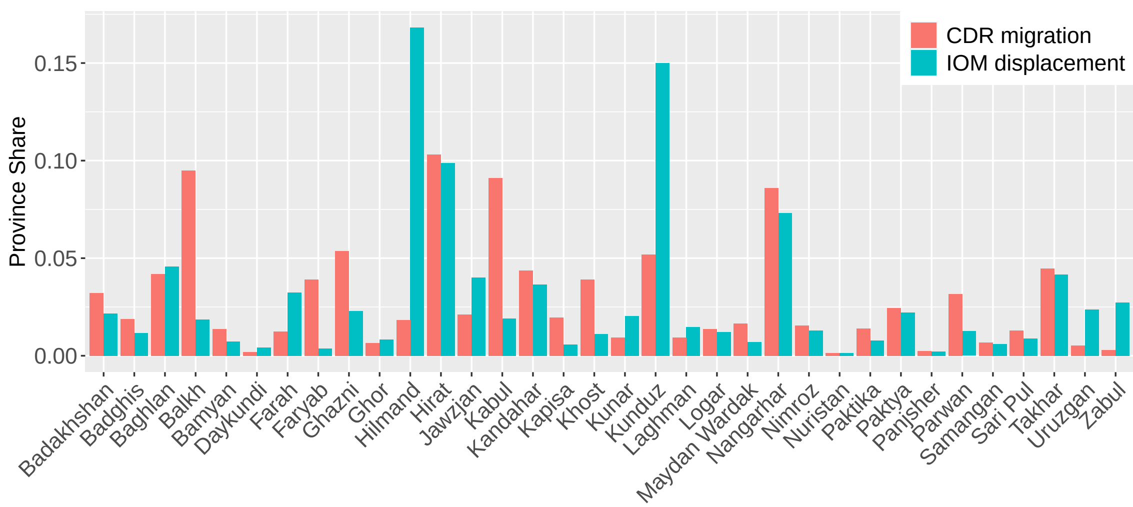


Figure 6: Comparison of IOM-based estimates of displacement (official data) and CDR-based estimates of migration (mobile phone data). Kabul district is not included as it distorts the y-scale of the axes.



(a) Province shares of total outgoing individuals, by origin province.



(b) Province shares of total incoming individuals, by destination province.