Pre-print version of

Amarasinghe, A., Nanlohy, S., Morgan, T., Hammond, D., Dahiya, Y., & Bailo, F. (2026). Mapping violence perceptions through YouTube comments: A new approach to real-time violence monitoring. *EPJ Data Science*.

https://doi.org/10.1140/epjds/s13688-026-00649-y

Mapping violence perceptions through YouTube comments: A new approach to real-time violence monitoring

Abstract

This paper introduces the Violence Perception Index (VPI), a novel methodology for quantifying violence-related discourse through geolocated YouTube comments. Utilizing the YouTube API and natural language processing techniques, the VPI measures public references to violence across 1.2 million unique geolocated videos in Mexico (2020–2024), extracting 14.8 million comments from over 500,000 videos with user engagement. This approach provides spatiotemporally granular data on violence-related discourse, which we treat as a proxy for violence perceptions, extending beyond traditional event-based datasets by capturing not only documented violence but also rumors, fears, and community discourse about violence, dimensions that influence community behavior and social stability independently of official records.

Violence scores are constructed using a weighted Spanish-language dictionary developed through semantic network expansion from violence-related seed terms. The dictionary-based scoring approach demonstrates moderate-to-substantial agreement with large language model classifications across 700 stratified comments (75-81% agreement), validating the method's capacity to systematically identify violence-related discourse at scale while maintaining computational efficiency for processing millions of comments.

The VPI is benchmarked against established violence indicators including ACLED fatalities and official municipal homicide statistics through panel regression specifications incorporating comprehensive spatial and temporal fixed effects. Analysis reveals systematic geographic heterogeneity: the VPI correlates strongly with ACLED data in high-population areas but exhibits stronger correlation with official homicide records in low-population contexts. Rather than constituting a methodological limitation, this pattern demonstrates the VPI's enhanced sensitivity in marginalized and remote regions where news-based datasets suffer from systematic reporting bias. The methodology is immediately scalable across languages and geographies, providing complementary intelligence for conflict monitoring, early warning systems, and policy interventions in precisely those underrepresented areas where traditional event-based monitoring systems provide incomplete coverage.

Keywords

violence perception; social media analysis; conflict monitoring; geospatial data; Mexico; social sensing; computational social science

Introduction

In an era where social and political stability is deeply entangled with digital communication, online platforms have become vital arenas for communities to perceive, discuss, and respond to violence. User-generated content, especially comments on social media, offers an unfiltered, ground-level perspective on unrest, whether documenting direct experiences, sharing

warnings, or circulating rumors. Yet the potential of this rich discourse to explain and quantify violence dynamics remains underexploited. Traditional datasets consistently report actual violence through fatality counts and event records, but they cannot capture perceptions of violence, the fear, uncertainty, and discourse that shape community behavior. These datasets also overlook unverified claims and rumors which, even when false, can provoke fear and destabilization in fragile contexts (Rohman & Ang, 2019). Critically, both news-based sources and official statistics suffer from systematic geographic bias (Weidmann, 2016): violence in urban centers receives extensive coverage while marginalized and remote areas remain systematically underreported (see Shaver et al., 2023). This paper addresses these issues by introducing the Violence Perception Index (VPI),[1] which quantifies violence-related discourse through geolocated YouTube comments across Mexico (2020-2024), providing spatially granular insights into how violence is perceived and discussed, particularly in those underrepresented regions where traditional monitoring fails.

The VPI leverages natural language processing and spatial modeling to transform user-generated YouTube comments into a quantitative spatiotemporal indicator of violence-related discourse, treated here as a proxy for violence perception. From 1.2 million unique geolocated videos, we extract comments from over 500,000 videos with user engagement, scoring 44 million comment-location pairs using a Spanish-language violence dictionary. We then aggregate these scores into monthly grid-cell measures (50km resolution) across Mexico. This approach captures not only where violence occurs, but *where people believe it is occurring*—a distinction critical for understanding how misinformation, institutional distrust, or localized fear can amplify instability, independent of actual events. Unlike retrospective event datasets, the VPI can provide near-real-time monitoring at unprecedented spatial granularity, enabling early detection of emerging violence hotspots through shifts in public discourse.

Our work contributes to growing efforts to quantify social and political behaviors through digital trace data. Recent studies use sentiment scores from news data (Amarasinghe, 2022, 2023), Wikipedia activity (Oswald & Ohrenhofer, 2022), and Twitter (Arthur et al., 2018) to predict conflict escalation, crisis detection, and public attitudes. We extend this literature by exploiting YouTube comments, a thus far untapped source offering geolocated, vernacular discourse that complements news-based datasets. Critically, we demonstrate that VPI correlates most strongly with official violence statistics in low-population, marginalized areas where news-based datasets show no significant relationship. This highlights VPI as a tool for monitoring precisely those regions where traditional systems fail, addressing long-standing concerns about reporting bias in conflict data (Shaver et al., 2023; Weidmann, 2016).

We validate the VPI against three established violence indicators: Armed Conflict Location and Event Data (ACLED) fatalities (Raleigh et al., 2023), Uppsala Conflict Data Program (UCDP) conflict fatalities (Davies et al., 2024), and official municipal homicide statistics. We also integrate socioeconomic and demographic data, including population density (WorldPop, 2020) and a composite marginalization index based on results from the 2020 Census that measures social and economic disadvantage across education, housing quality, income, and geographic isolation (Consejo Nacional de Población, 2021). Using panel regressions with comprehensive spatial and temporal controls, we examine whether realized violence predicts variation in violence perception as captured by the VPI. This validation strategy allows us to assess, first, if VPI meaningfully tracks actual violence dynamics and reveals systematic differences in how violence is discussed online versus documented across different geographic contexts and, second, if VPI captures dynamics in low-population and marginalized areas underrepresented by traditional conflict datasets.

This paper makes both methodological and conceptual contributions to conflict monitoring. Methodologically, we demonstrate that large-scale, systematic measurement of violence perception is feasible through geolocated social media discourse, offering a scalable framework applicable across languages and geographies. Conceptually, we establish that discourse about violence, which we treat as a proxy for perceptions, not just documented events, constitutes actionable conflict data. When communities fear violence, whether based on direct experience, news reports, or rumors, that fear shapes behavior, economic activity, and social stability. The VPI captures these dynamics in near-real-time at fine spatial

[1] The dataset is available for download at https://osf.io/fa493/overview?view_only=f7b43705605443c6a619b31cf66c60b5) (anonymised link).

resolution, providing complementary intelligence for early warning systems and policy interventions in contexts where community perceptions may diverge from or precede official violence records.

What does the VPI Measure?

The Violence Perception Index quantifies the intensity of violence-related discourse in geolocated online comments, but it is important to clarify what this captures and what it does not. The VPI aggregates all references to violence without distinguishing their underlying nature. A high VPI score in a given location-month may reflect several distinct phenomena:

1. **Direct experience:** Community members documenting violence they have witnessed or experienced firsthand.
2. **Perceived threat:** Expressions of fear or concern about potential violence, whether based on patterns, rumors, or general insecurity.
3. **News circulation:** Discussion of violence reported in the media, which may or may not be local.
4. **Rumors and speculation:** Unverified claims about violence that may be partially true, exaggerated, or false.
5. **Historical reference:** Commentary on past violence or long-standing patterns.

Each of these discourse types matters for understanding community responses to violence, but they are expected to exhibit different spatiotemporal dynamics. Direct experiences generate localized discourse spikes tied to specific events, while news circulation and rumors can spread across regions, creating perception increases in areas far from actual violence. National events, such as high-profile assassinations, may produce uniform discourse nationwide, while local events generate geographically concentrated responses. From a behavioral perspective, all contribute to the perception environment shaping voting behaviour, economic activity, and daily security practices, whether threats are immediate and local or diffuse and national.

In this sense, comment sections of geolocated videos may function as “third spaces”, non-political online spaces where political talk emerges organically (Wright, 2012). Because videos are geolocated and likely served to local audiences by YouTube's recommendation algorithms, their comment sections create localized gathering points where community members discuss local happenings independently of the video's actual content, much as informal conversation in a neighborhood café might drift toward local concerns regardless of the original topic.

In this iteration, we focus on establishing that violence-related discourse is systematically measurable at scale. The VPI does not differentiate between discourse types or spatial origins but it aggregates all violence-related language within each grid-cell. This represents both a limitation and a design choice. We cannot yet distinguish local eyewitness accounts from national news discussion, nor genuine fear from rumor-mongering. Different phenomena exhibit different correlation structures: locally experienced violence should correlate with same-grid violence over time, while nationally-driven discourse might correlate uniformly across grids during specific periods. Our design choice prioritizes demonstrating measurement feasibility over decomposing these signals.

The VPI quantifies what has traditionally been intangible: the intensity of violence in public consciousness, whether stemming from local experience, national news, or rumors, and whether reflecting in-grid events or concerns about violence elsewhere. By establishing that such measurement is feasible at scale and correlates meaningfully with documented violence patterns, we demonstrate the validity of treating violence perception as actionable conflict data.

A natural question is whether perceptions of violence as distinct from violence events correlate with actual violence levels. To examine this, we analyzed World Values Survey (WVS) Wave 7 (Haerpfer et al., 2024) data on violence-related attitudes across 62 countries. We find significant correlations between both mean attitudes and within-country attitude heterogeneity and national homicide rates, supporting the premise that perception-based measures track meaningfully with realized violence (see the Online Appendix Tables A23).

Why Mexico?

Mexico presents an ideal case study for deploying our violence perception methodology. The country exhibits high levels of both actual and perceived violence, creating an environment where digital discourse about violence is prevalent and consequential for social stability. Mexico's robust digital engagement provides the necessary methodological foundation: 78% of adults used social media in 2023 (Pew Research Center, 2023) and 79% relied on the Internet as their primary news source in 2024 (Newman et al., 2024), ensuring substantial user-generated content for analysis. Additionally, the June 2024 general election, the country's most violent on record, provided a critical temporal event for examining how violence perceptions fluctuate during periods of heightened political sensitivity.

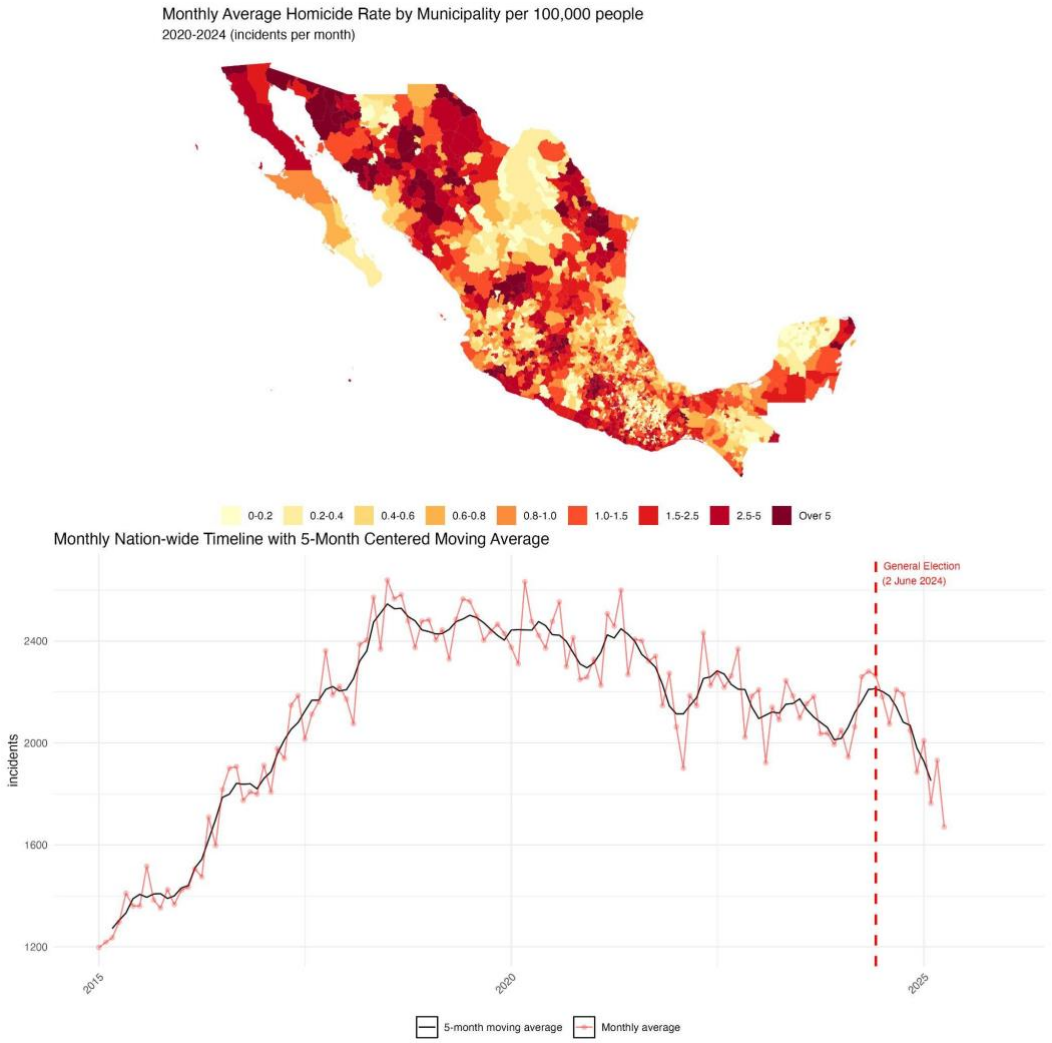


Figure 1. Geographic and temporal distribution of homicides in Mexico, Top row, spatial distribution of average monthly homicides across Mexican municipalities (2020–2024). Bottom row, monthly average of nation-wide homicide counts showing the 2019 peak and subsequent decline. Data source: Mexican Government (Gobierno de México, 2025).

More than 300,000 people have been murdered in Mexico over the past decade (Institute for Economics & Peace, 2025), with homicide rates peaking at 28.2 deaths per 100,000 people in 2019 before declining modestly to 23.3 in 2023-2024 (see Figure 1). Organized criminal groups drive approximately two-thirds of these killings and effectively control an estimated 30% of Mexico's territory (Murray & Stott, 2025). These groups act as powerful political actors: their turf wars and clashes with security forces generate extreme violence, while their extortion and illicit activities substantially undermine societal security and perceptions of safety. Violence concentrates in strategically valuable areas and in regions disputed by multiple criminal organizations.

Electoral violence exemplifies how organized crime shapes both actual violence and public perception. Political homicides rose from 51 in 2020 to 201 in 2024, with most targeting municipal-level officials whose control matters strategically to criminal groups (Data Cívica, 2024). Violence against candidates at the electoral cycle's start signals territorial control, warning that areas fall under criminal influence and deterring uncooperative candidates. Paradoxically, states dominated by single powerful cartels, such as Jalisco (the Jalisco New Generation Cartel stronghold) and Sinaloa (Sinaloa Cartel stronghold), experience less political violence, suggesting that criminal hegemony enables control without overt violence. This complex interplay between actual violence, territorial competition, and strategic signaling makes Mexico an especially rich context for examining how violence perceptions form and spread.

Data

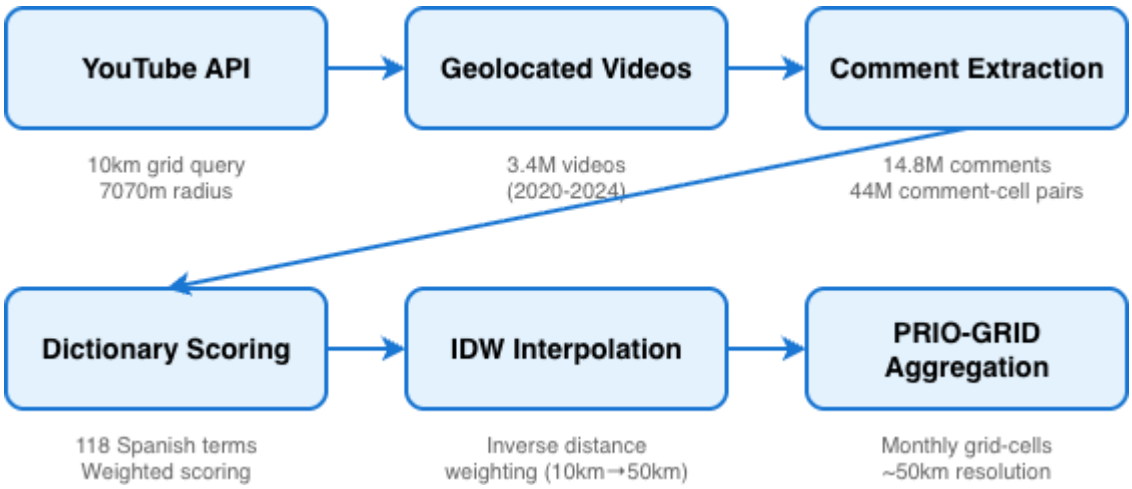


Figure 2: Data collection and processing pipeline

We construct the VPI through a multi-stage process involving data collection, natural language processing, and spatial aggregation. Figure 2 illustrates the complete workflow, which we detail below.

We use the YouTube API to extract comments and other related metadata of public videos associated with Mexico.[2] More details on the data collection and processing are provided in the Online Appendix and supplementary materials. First, we rasterize the geographic area of Mexico into 10 km grid-cells. A 10 km grid is created from the WorldPop grid with resident population estimates for 2020 after excluding grid-cells with less than 100 inhabitants (WorldPop, 2020). Second, we use the "search" endpoint of the YouTube API to fetch videos which indicate, in their metadata, a geographic location within a radius of 7070 m from each grid-cell centroid (a radius search being the only geographic query method supported by the API). Notably, we do not use any search term, and instead only query based on the video location. Based on a sample of 2,525 Mexican YouTube videos, we estimate that approximately 3.45% of YouTube videos indicate a location in their metadata. While this proportion may appear modest, it is comparable to geolocation rates on other platforms (Huang & Carley, 2020). Applied to YouTube's scale, this still yields a sample size substantially exceeding most survey-based or event-coded conflict datasets. The validation results presented below confirm that this geolocated sample captures meaningful variation in violence dynamics. For each fetched video, we extract metadata such as video id, title, description, video category and publication time. For our sample period between 1 Jan 2020 to 2 June 2024 (the day of federal election), we query 3,398,143 search results. Due to overlapping search radii (see Figure A1), many videos appear in multiple grid-cell queries; after deduplication, this yields 1,229,594 unique videos, of which 508,942 received at least one comment and form the basis of our analysis.

[2] Access to the YouTube API was provided as part of the YouTube Researcher Program.

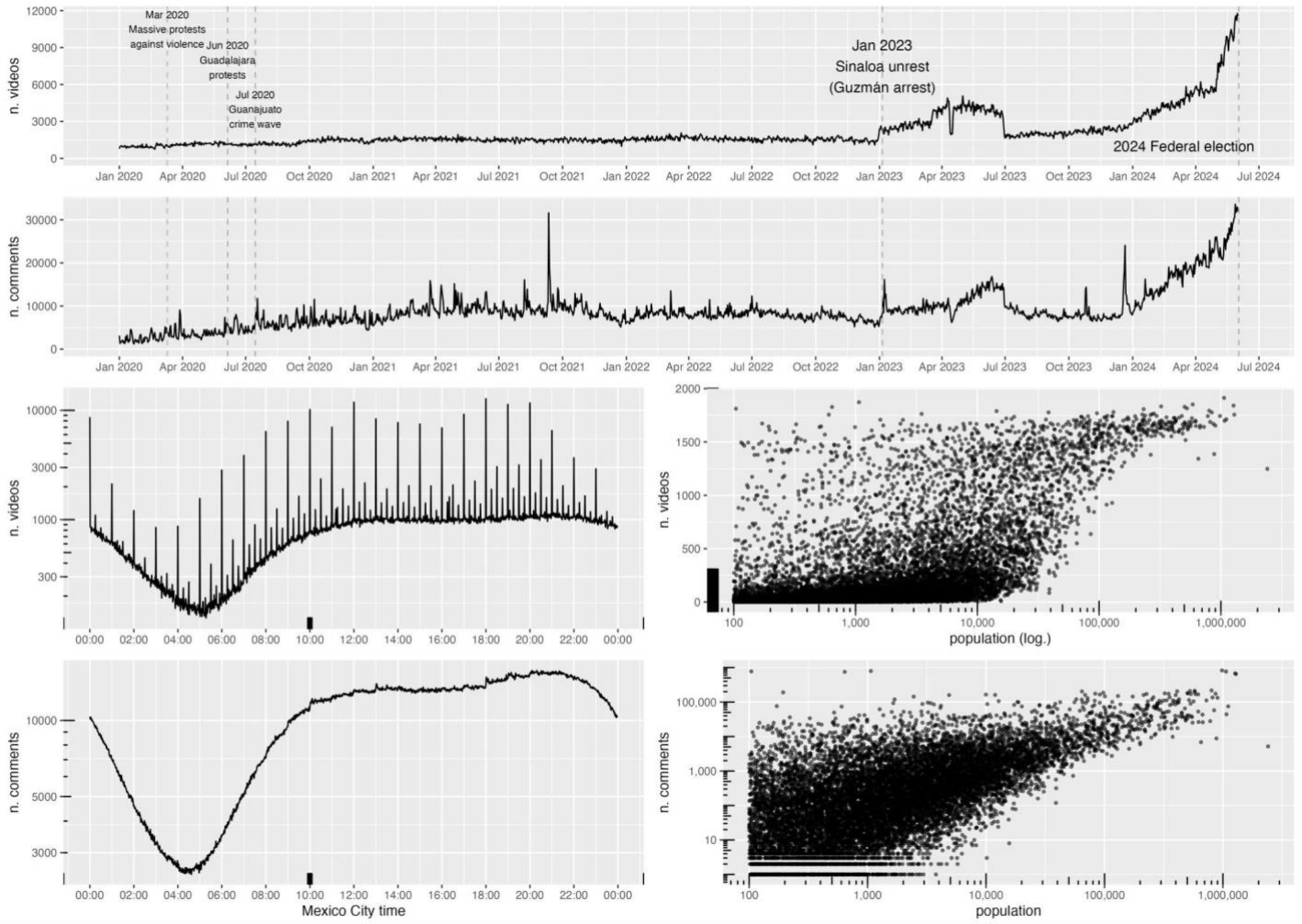


Figure 3. Distribution of videos and comments over time and by grid population. Top panel: Daily count of unique videos (n = 1,229,594, after removing duplicates from overlapping grid searches) and comments (n = 14,785,647). Vertical dashed lines mark notable events: the March 2020 nationwide women's protests against gender-based violence; the June 2020 protests in Guadalajara following the death of Giovanni López in police custody; the July 2020 presidential visit to Guanajuato amid escalating cartel violence and public criticism of security policy; the January 2023 unrest in Sinaloa following the arrest of Ovidio Guzmán López; and the June 2024 federal election. Bottom-left panel: Distribution of posting times throughout the day. Bottom-right panel: Relationship between grid-cell population density and the number of videos and comments associated with each grid-cell; the vertical line indicates the 100-person threshold for grid-cell inclusion.

Third, using the "video" endpoint of YouTube API, we fetch additional information for each video, including, notably, the number of comments received by the 1,229,594 unique videos. Fourth, for each video that received at least one comment, we query the "commentThreads" endpoint to fetch data and metadata of comments. This returns 14,785,647 unique comments to 508,942 videos. As video locations can span across multiple grid-cells, the same video is often associated with more than one grid-cell. This finally resulted in 43,900,491 unique comment-cell pairs.

At the top of Figure 3, we show the daily distribution of videos and comments based on the time of publication indicated in the metadata. We notice a significant increase in the number of videos and comments in the months preceding the 2024 federal election. The bottom-left panel shows posting times throughout the day, which validates that comments align with expected Mexican sleep patterns. Additionally, since video publications are typically scheduled in advance, most videos are posted at the top of the hour (e.g., 9:00, 10:00), reflecting automated scheduling practices. Finally, the bottom-right panel shows the close association between population density of the grid-cells used to query the YouTube API and the number of videos and comments associated with the same grid-cells. In this case, we also note the cutoff line of 100 people per grid-cell as explained above.

The geographic distribution of videos and comments (alongside actual population estimates) are mapped in Figure 4.

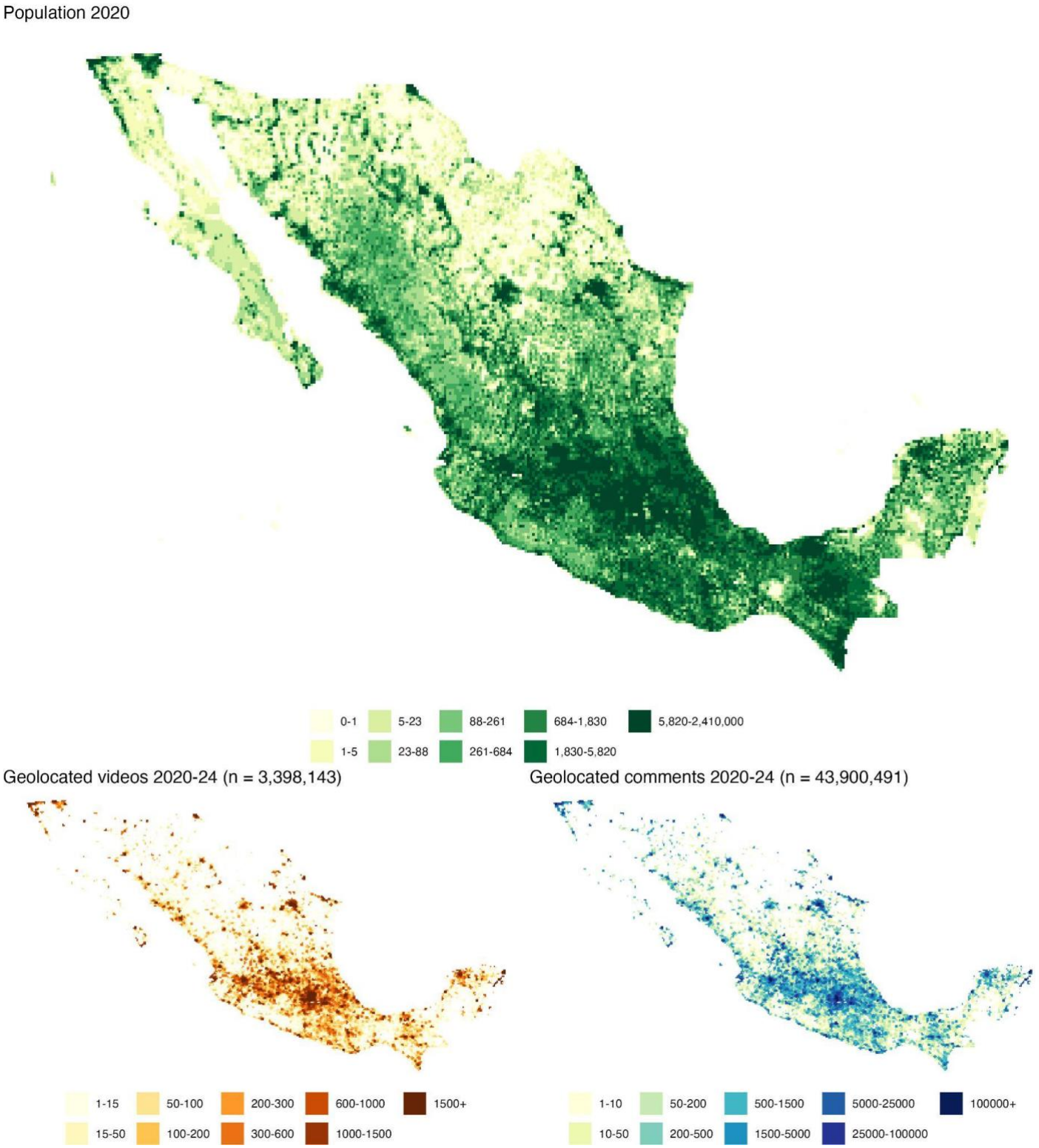


Figure 4. Geographic distribution of population and YouTube activity across Mexico at 10km grid resolution. Top row: Resident population from WorldPop 2020 estimates. Middle row: Number of unique geolocated YouTube videos per grid-cell. Bottom row: Number of comments per grid-cell. All panels use the same spatial extent. Note: Colors in all panels represent quantioles of the respective distributions; population panel uses a sequential (non-diverging) palette to distinguish it from the video/comment distributions.

To measure the perception of violence expressed by each comment we used a bag-of-words, dictionary-based approach. We leveraged the Open Multilingual Wordnet project (Bond & Paik, 2012) and the Spanish Wordnet (Gonzalez-Agirre et al., 2012) to build a dictionary of violence-related terms. Starting with 10 arbitrary seed words ("violencia," "asesinato,"

"homicidio," "tiroteo," "ataque," "enfrentamiento," "balacera," "secuestro," "narcotráfico," "delincuencia"), we extracted all words within a network distance of two. This process yielded a list of 118 terms (available in the Online Appendix in Table A3), which were then weighted at 1, 0.25, or 0.5, inversely proportional to their distance from the original seed words. Each comment is therefore assigned a scalar violence score based on the weighted frequency of dictionary terms appearing in the comment text. Of the 14.8 million comments in our dataset, approximately 7.6% received a positive violence score (i.e., contained at least one dictionary term).

The construction of word lists through expansion from seed terms is methodologically consequential, as different approaches can yield substantially different results (Di Natale & Garcia, 2024). We adopted the WordNet-based semantic network approach because multilingual WordNet resources enable consistent application across languages, supporting scalability beyond Mexico. Importantly, our analysis focuses on *relative* variation in violence discourse across time and space rather than absolute measurement. By capturing terms semantically proximate to core violence concepts, we detect meaningful spatiotemporal variation rather than estimate average levels.

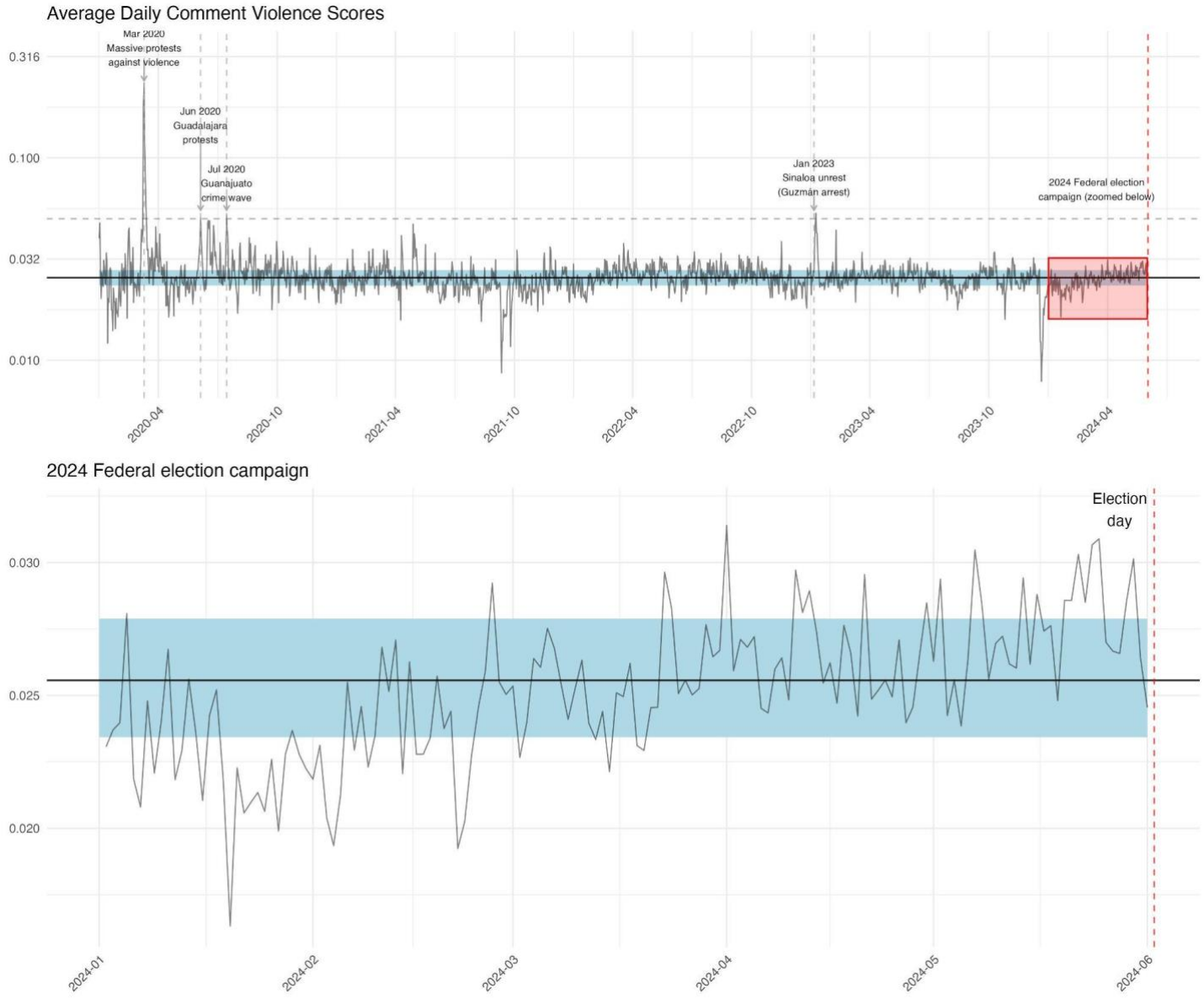


Figure 5. Average daily comment violence scores during the study period (January 2020–June 2024). The horizontal black line shows the long-term median, while the blue band represents the interquartile range (25th–75th percentiles) of daily averages. Vertical dashed lines mark events where daily scores exceeded 0.05: the March 2020 nationwide women's protests against gender-based violence; the June 2020 protests in Guadalajara following the death of Giovanni López in police custody; the July 2020 presidential visit to Guanajuato amid escalating cartel violence and public criticism of security policy; and the January

2023 unrest in Sinaloa following the arrest of Ovidio Guzmán López. The red shaded area highlights the 2024 federal election campaign period (March–June 2024), which is displayed in detail in the lower panel.

Beyond computational efficiency, the dictionary-based approach offers practical advantages for cross-linguistic scalability: WordNet resources exist for dozens of languages, enabling consistent application across geographies without requiring labelled training data. We explored a transformer-based alternative (fine-tuned BERT/BETO) and found it computationally impractical for real-time monitoring at scale and prone to systematic overclassification (more details are provided in the Online Appendix).

Finally, to produce the Violence Perception Index (VPI) at the PRIO-GRID 2.0 level (Tollefsen et al., 2012) (approximately 50km x 50km in size), we aggregate scores of comments published each day using inverse distance weighted interpolation (Pebesma, 2004) based on their location on the WorldPop 10 km grid. This method assigns each grid a weighted average of nearby comment scores, with weights inversely proportional to the distance between comment locations and grid-cell centroids, thereby transforming point-based social media data into a standardised gridded format compatible with other geospatial datasets. The VPI for each PRIO-GRID cell is calculated as the monthly average of these daily interpolated scores. The daily average for Mexico is plotted in Figure 5.

This approach addresses the challenge that individual comments are geolocated through their parent videos' coordinates, which may not align precisely with grid-cell centroids or boundaries. By using IDW interpolation, we create a smooth, continuous surface of violence perception intensity where each grid-cell's score reflects a distance-weighted average of all surrounding comments, with closer comments exerting greater influence. Notably, we interpolate the violence scores without covariates (using only the spatial coordinates). This method is particularly appropriate for our application because it preserves local variation in violence discourse whilst avoiding sharp discontinuities at arbitrary grid-cell boundaries, and it naturally handles the spatial uncertainty inherent in user-generated geographic data where the precise relationship between a video's recorded location and the geographic scope of its associated comments may be ambiguous.

Validation of Dictionary-Based Violence Scores

To validate our dictionary-based violence scoring approach, we compared it against classifications from four large language models (Qwen-Qwen3-vl-4b, IBM-Granite-3.2-8b, Google-Gemma-3n-E4b, and Meta-Llama-3-8b-Instruct) on a stratified random sample of 700 comments. We refer to the Online Appendix for more details on our approach. Each LLM independently classified whether comments discussed violence and rated violence frequency on a 0-10 scale. Despite fundamental methodological differences (our approach counts lexical occurrences while LLMs perform semantic interpretation) we find substantial convergent validity.

Agreement metrics demonstrate that both approaches identify the same underlying violence construct. Cohen's Kappa values between our binary classifications (i.e. with dictionary-based score of 1 or more) and LLM judgments range from $\kappa = 0.52$ to $\kappa = 0.62$, indicating moderate agreement, with overall Fleiss' Kappa of 0.80 ($p < 0.0001$) across all raters (Online Appendix Table A6). Raw agreement rates between our scores and LLMs range from 75-81% (Table A7), comparable to inter-LLM agreement of 87-94%. For continuous scores, Spearman rank correlations range from $\rho = 0.61$ to $\rho = 0.68$, and rank-based intraclass correlations from ICC = 0.61 to 0.68 (Tables A8-A9), indicating that our lexical approach and LLM semantic analysis produce highly consistent ordinal rankings of violence intensity. This convergence across complementary measurement approaches validates our dictionary-based scores as capturing meaningful variation in violence-related content suitable for large-scale analysis of over 14 million comments.

Validation of VPI Against Alternative Measures of Realized Violence

How well does our VPI reflect actual violence? The utility of any conflict indicator depends fundamentally on its correspondence with observable violence dynamics. This section presents a series of statistical analyses demonstrating that the VPI exhibits strong and robust correlations with realized violence across multiple established datasets. Beyond this

validation, we show that the VPI provides distinctive analytical value by capturing violence perceptions in remote and marginalized geographic contexts where news-based event datasets face systematic reporting limitations.

For the purpose of this empirical exercise, we use PRIO-GRID cells as the unit of analysis, while the sample period is January 2020 to May 2024.

Before commencing our statistical analyses, we explore the visual correlations of the VPI and alternative measures of realized violence. Figure 6 presents this visual representation. In Panel (a) we plot the distribution of average VPI over the sample period, in quantiles, across the set of grids covering Mexico. In Panels (b) and (c), we plot the distribution of the sum of ACLED fatalities and homicides, over the sample period, in quantiles. To ensure we are not simply capturing population dispersion patterns, all variables are weighted by a factor that expresses each grid's population as a share of the total country's population. To avoid confusion when VPI is weighted for the local population, we refer to it as VPI^w. Later we also use the VPI unweighted by population in our robustness exercises. There appears a striking correlation between the VPI^w and alternative measures of violence, with much of the realised as well as the perceived violence being concentrated in the southern parts of Mexico.

These descriptive patterns are suggestive but not definitive. Several confounding factors may generate spurious correlations between the VPI^w and violence measures. First, violence occurs where populations exist, meaning correlations could simply reflect population distribution rather than a genuine perception-violence relationship (Figure 4 shows this spatial correspondence). While we weight indices by grid-cell population share, this partial solution requires more rigorous controls. Second, geographic characteristics, distance to urban centers, proximity to coasts, terrain type, may independently influence both violence levels and digital discourse patterns. Third, temporal shocks such as economic downturns or political crises may simultaneously affect both realized and perceived violence.

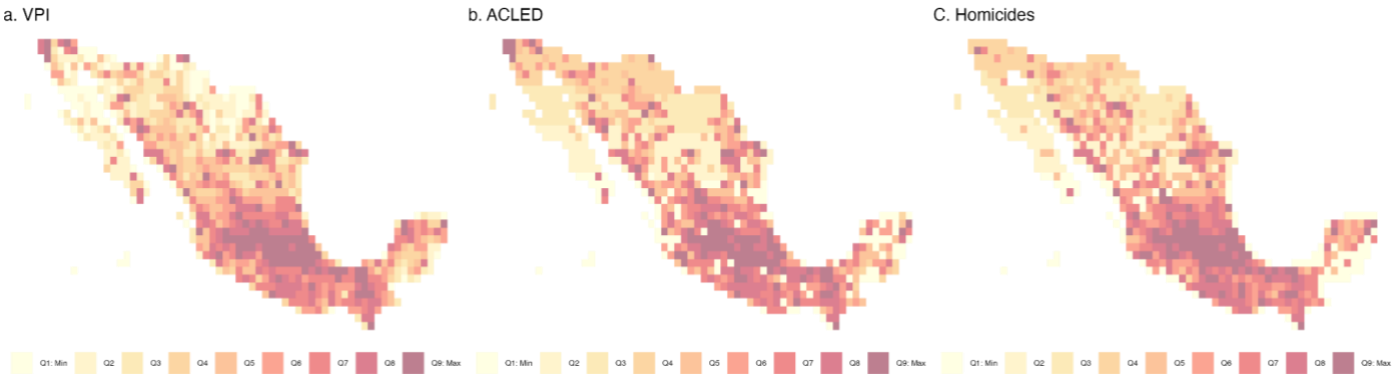
Our empirical strategy addresses these concerns through panel regression with comprehensive fixed effects. Grid fixed effects absorb all time-invariant characteristics including location, terrain, and baseline population density. Year and month fixed effects capture temporal and seasonal shocks common across all grids. This specification isolates the within-grid, over-time association between realized violence and violence perception, netting out confounding spatial and temporal factors.

Now we proceed to our formal statistical analysis. To examine the strength of the correlations between VPI^w, where VPI^w is weighted by each grid-cell's population share, and these alternative measures of realized violence, we utilise the following econometric specification:

$$VPI^w_{i,y,m} = \beta RV_{i,y,m} + \gamma_i + \theta_y + \lambda_m + \epsilon_{i,y,m} \quad (1)$$

where, VPI^w is the population-weighted Violence Perception Index, as observed in grid-cell i in month m of year y. RV is a measure of realised violence, again for grid i in month m of year y, based on alternative datasets, as discussed below. Depending on the dataset used for verification, this can be a measure of fatalities or violent events. γ is a vector of grid fixed effects, which captures all time-invariant unobservables attributable to a grid, such as its location within the country or its geographic landscape, which might affect the relationship between RV and VPI^w. Given the short sample period and the slowly evolving nature of population density, especially within the temporally granular "monthly" unit of analysis, the set of grid fixed effects also accounts for whether the grid has a high vs low level of population density. θ accounts for time-variant unobservables, such as global economic shocks, while λ accounts for seasonal unobservables, both of which might affect the relationship of interest. ϵ is the error term.

Figure 6. Geographic comparison of violence perception and realized violence indicators, aggregated at PRIO-GRID level (approximately 50km × 50km). All measures are averaged over the sample period (January 2020–May 2024) and population-weighted. (a) Violence Perception Index (VPI). (b) ACLED-reported fatalities. (c) Official homicide statistics. Colors represent quintiles of each distribution; darker shading indicates higher values. Colors represent quantiles.



Once all such unobservables are accounted for, the coefficient of interest β reflects the correlation between actual, realised violence and perceived violence, as quantified by VPI^w. We expect β>0, indicating that actual violence translates to perceived violence as expressed through public engagement online.

Table 1 presents the baseline estimates. Here, we present correlations for realized violence as per ACLED fatalities (Panel A), and Homicides (Panel B). For both panels, in Column (1) we present raw correlations, without accounting for any time-invariant or time-variant unobservables. In Column (2) we incorporate grid fixed effects, to tease out time-invariant grid-specific unobservables, while in Column (3) we incorporate year fixed effects, to tease out time-varying unobservables affecting all grids. Our preferred estimates appear in Column (4), where we additionally include month fixed effects to account for seasonal unobservables.

Across the two panels, we observe that realized violence, as proxied by $Fatalities$, is positively and statistically significantly correlated with violence perception, as proxied by VPI^w. The size of the coefficient is typically larger in Column (1), due to unobservables not being appropriately controlled for. However, importantly, the coefficient remains both economically and statistically significant when unobservables are accounted for, via the inclusion of comprehensive sets of fixed effects. Moreover, when fixed effects are incorporated to absorb unobservables, the reported R^2 improves, with $Fatalities$ explain up to 97% of the variation in VPI^w, on average. In terms of economic significance, the coefficients in Column (4) of Panels A and B, indicates that a 1-standard deviation increase in realized violence is associated with a 0.0028 and 0.0051 percentage point increase in VPI^w, respectively. These values translate to a meaningful 37% and 67% increase, relative to the average VPI^w within the sample period, for Panels A and B, respectively.

One key concern when examining the association between the VPI^w and realized violence indicators is population distribution: violence occurs where populations exist, and correlations could simply reflect this spatial correspondence. Our baseline specification addresses this concern through the inclusion of grid fixed effects, which capture all time-invariant grid-specific characteristics including population density. As such, the estimates in Table 1 report within-grid correlations over time, isolating how changes in realized violence within a given location predict changes in violence perception in that same location. This increases confidence that our estimates are not driven by between-grid differences in population.

We plot the distribution of the error term against realized violence in Figure 7 and geographically in Figure 8 to confirm that our estimates are not driven by outliers. These residual diagnostics reveal an important pattern: the error term from both models exhibits greater variance in areas with high levels of marginalization (bottom panel of Figure 7).

This pattern warrants careful interpretation. Larger residuals in marginalized areas could indicate model misspecification or measurement error. However, we argue it more likely reflects a distinctive strength of the VPI^w: enhanced sensitivity to violence discourse in precisely those contexts where traditional datasets face the greatest limitations. Marginalized areas, characterized by limited education, poor infrastructure, geographic isolation, and weak state presence, are systematically underrepresented in news-based conflict datasets like ACLED, which rely on media coverage concentrated in urban centers. Yet these same communities may actively discuss violence on social media platforms, generating signals the VPI^w captures but ACLED misses.

The wider error distribution in high-marginalization areas thus suggests the VPI detects violence-related discourse that does not fully correspond with officially documented events—not because the VPI is noisy, but because it captures local perceptions, rumors, and fears in communities where violence often goes unreported by formal channels. This interpretation finds support in our population split analysis (see Table 2) discussed below.

Table 1: Baseline estimates - Correlation between VPI^w and alternative indicators of violence

	(1) $VPI^w_{i,y,m}$	(2) $VPI^w_{i,y,m}$	(3) $VPI^w_{i,y,m}$	(4) $VPI^w_{i,y,m}$
Panel A: ACLED				
$Fatalities_{i,y,m}$	0.2284***	0.0257***	0.0257***	0.0257***
	(0.0509)	(0.0060)	(0.0060)	(0.0060)
Observations	45,580	45,580	45,580	45,580
R-squared	0.5835	0.9740	0.9740	0.9741
Panel B: Crime				
$Homicides_{i,y,m}$	0.0836***	0.0143***	0.0143***	0.0142***
	(0.0026)	(0.0043)	(0.0043)	(0.0042)
Observations	45,580	45,580	45,580	45,580
R-squared	0.8312	0.9739	0.9739	0.9740
Grid FE	No	Yes	Yes	Yes
Year FE	No	No	Yes	Yes
Month FE	No	No	No	Yes

The outcome variable is the population-weighted VPI. In Panel A, $Fatalities$ is the number of violence- related fatalities in grid i in month m of year y, as reported by ACLED. In Panel B, $Homicides$ is the number of homicides in grid i in month m of year y, as reported by the Mexican Government (Gobierno de México, 2025). Standard errors, clustered at the grid level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

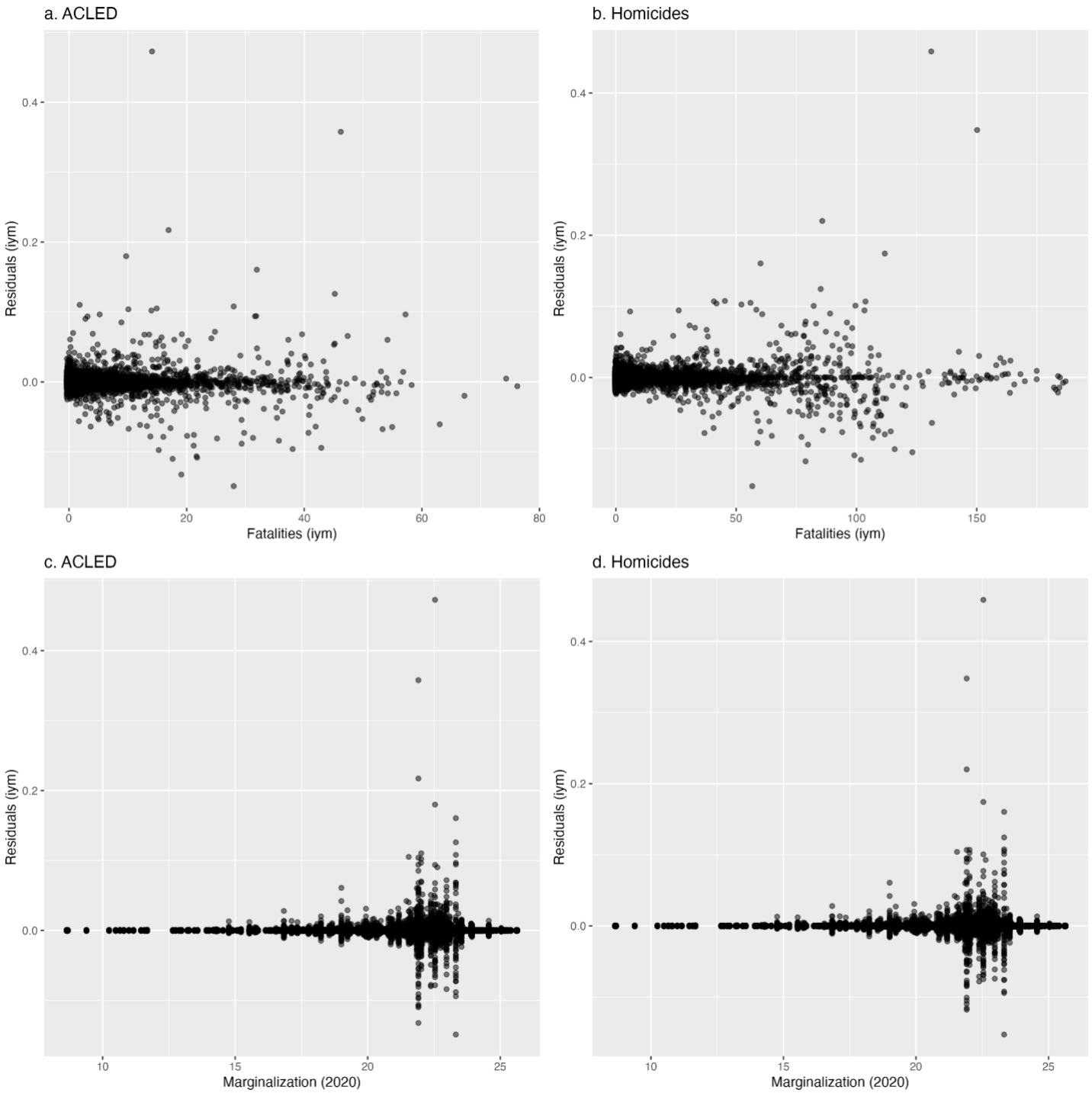
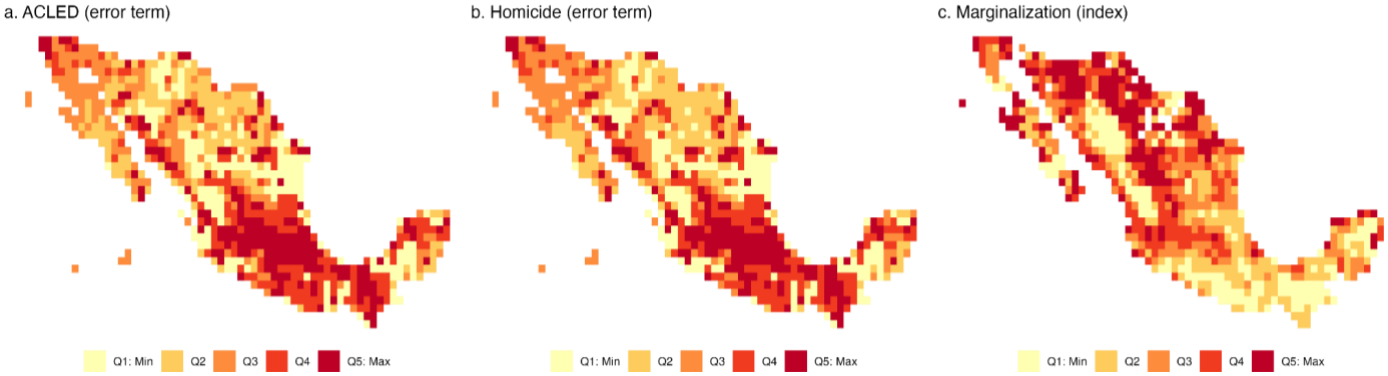


Figure 7: Residual analysis from baseline regression specifications (Equation 1). (a–b) Distribution of residuals ($\varepsilon_{i,y,m}$) against realized violence measures: ACLED fatalities (a) and official homicides (b). (c–d) Residuals plotted against the marginalization index, which measures social and economic disadvantage based on 2020 Census indicators including educational attainment, housing quality, income levels, and geographic isolation; higher values indicate greater marginalization. The wider residual variance in highly marginalized areas (panels c–d) suggests VPI^w captures violence discourse in communities where formal documentation channels are weakest.

Figure 8: Geographic distribution of the error term $\varepsilon_{i,y,m}$ and marginalisation index. Panels (a) and (b) show the distribution of the error term, aggregated at the grid level, estimated as per Equation (1), for ACLED fatalities and Homicides respectively. Panel (c) displays the marginalisation index averaged at the grid level based on the 2020 Census (higher values indicate greater marginalisation). Comparison across panels reveals that larger residuals tend to concentrate in highly marginalised areas,

particularly in southern Mexico, suggesting the VPI^w captures violence-related discourse in regions where news-based datasets face systematic reporting limitations. Colors represent quantiles.

To assess whether VPI^w primarily reflects nation-wide events that everyone discusses simultaneously, we decompose VPI^w variation into two components: variation across grids (spatial) and variation over time within grids (temporal). If VPI^w



mainly captures national discourse, such as widespread discussion of high-profile political assassinations or federal policy announcements; we would expect to see grids moving together over time, producing high temporal variation as all regions respond to the same national events. Conversely, if VPI^w captures localized violence perception, we would expect substantial differences between grids (violent vs. peaceful regions) but relatively stable patterns within each grid over time. Table A12 presents the variance decomposition, splitting total VPI^w variation into between-grid (spatial) and within-grid (temporal) components. The results indicate that 97.6% of VPI^w variance is between-grid (spatial heterogeneity) while only 2.7% is within-grid over time (temporal fluctuations). The between-grid standard deviation (0.0326) is approximately six times larger than the within-grid standard deviation (0.0054). This pattern indicates that geographic differences in violence perception vastly exceed common temporal shocks. VPI^w variation is primarily driven by "where" (which grid) rather than "when" (which time period). This is precisely the pattern expected if VPI^w captures localized or regional violence dynamics rather than uniform national discourse about historical events. The minimal within-grid temporal variation (2.7%) also validates our empirical strategy: our time fixed effects (year and month) successfully absorb common national shocks, leaving the identifying variation to come from differential grid-specific violence exposure. However, we acknowledge a remaining subtlety: if national events affect different grids heterogeneously (e.g., discussion intensity varies by proximity to the event or regional political alignment), time fixed effects won't fully absorb this variation. This could generate spurious correlation if such events cluster geographically with actual violence.

To examine temporal dynamics, we compared VPI^w and ACLED as 30-day moving averages normalized to their 2023 annual means (Online Appendix, Figure A13). Both indicators track upward together during the June–July 2020 violence escalation and the 2024 presidential campaign. Notably, VPI spikes in early 2020 while ACLED remains flat—capturing nationwide discourse on gender-based violence not reflected in fatality counts. These patterns suggest VPI is best suited for detecting localized perturbations rather than long-term trend analysis.

In Table 2, we identify high-population vs low-population grids, and estimate the correlations for these two sets of grids separately. Here too, the outcome variable is *VPI* (this time unweighted by population). The splitting of the sample in this manner is based on the distribution of population, with grids above the median level of population being classified as high-population grids, while those below are classified as low-population grids. As mentioned above, we observe an interesting pattern. In Panel A, when considering correlations between VPI and ACLED fatalities, there exists a statistically significant correlation for high-population grids, while no correlation is observed for low-population grids. ACLED fatalities are based on news reports, and news reports are more likely to capture violent events in locations with high population exists, which potentially underlies this observed correlation. In Panel B, when considering official homicide statistics, we observe that the VPI is strongly correlated with homicides in low-population grids. Official homicide statistics are less likely to suffer from reporting bias as in ACLED, and this therefore highlights a key contribution of this paper. The VPI developed in this

paper can accurately capture violence perceptions in locations where existing event datasets such as ACLED do not capture on-the-ground violence, and as such, it addresses a stark limitation in existing data sets.

Table 2: Split sample analysis based on high vs low population grid-cells

	(2) $VPI_{i,y.m}$	(3) $VPI_{i,y.m}$
	High Population	Low Population
Panel A: ACLED		
$Fatalities_{i,y,m}$	0.0004***	0.0009
	(0.0001)	(0.0010)
Observations	22,790	22,790
R-squared	0.3613	0.3076
Panel B: Crime		
$Homicides_{i,y,m}$	0.0001	0.0004**
	(0.0001)	(0.0002)
Observations	22,790	22,790
R-squared	0.3610	0.3078
Grid FE	Yes	Yes
Year FE	Yes	Yes
Month FE	Yes	Yes

The outcome variable is *VPI,* which is VPI unweighted by population. In Panel A, $Fatalities$ is the number of violence- related fatalities in grid i in month m of year y, as reported by ACLED. In Panel B, $Homicides$ is the number of homicides in grid i in month m of year y, as reported by the Mexican Government. Standard errors, clustered at the grid level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Beyond examining residual patterns, we conduct additional robustness tests using alternative specifications and data transformations.

First, in Table A13, we present estimates excluding grid fixed effects, but including grid-level population as a control. Here we use the VPI unweighted by population, which we denote as *VPI,* as the outcome variable. We observe that the correlations remain statistically significant for both ACLED fatalities and homicides, although the size of the coefficient is smaller due to the exclusion of grid fixed effects.

Recall that the VPI^w is weighted by the share of population in each grid. In Table A14, we redo our baseline estimates, but with VPI unweighted by population as the outcome variable. We observe that the direction, magnitude and statistical significance of the relationship does not change drastically. However, the reported R^2 is lower than that observed in Table 1, which indicates that the population-weighted VPI, as used in baseline estimates, is a more accurate representation of violence perception, accounting for between-grid differences in population.

In Table A15, we present estimates when controlling for grid-level marginalization, and observe that the correlations between VPI^w and ACLED fatalities as well as homicides remain strong. (In Figure 8, we illustrate the mapping of census localities onto the PRIO-GRID by averaging the marginalization index at the grid-cell level.) In Table A16, we use the Inverse Hyperbolic Sine (IHS) transformation of the treatment variable, and results remain robust.

One concern related to the VPI^w is that rumors and speculation, phenomena VPI^w likely captures, can exhibit power law characteristics in their spread patterns. An examination of VPI^w's distribution confirms substantial right-skewness (skewness = 12.45) and heavy tails (kurtosis = 198.53), with 75% of observations below 0.0047. This pattern is consistent with information cascade dynamics where most grids experience modest violence discourse while a few experience intense discussion, potentially amplified by viral rumor spread. Our baseline estimates already partially address these distributional characteristics through standardization and population weighting to eliminate scaling issues. All estimates use standard errors clustered at the grid level, which are robust to heteroskedasticity and within-cluster correlation arising from heavy-tailed distributions. Additionally, grid fixed effects absorb time-invariant heterogeneity, including grids that persistently show extreme VPI^w values due to structural factors. This ensures coefficients identify within-grid variation rather than being driven by cross-sectional extremes. To further address this concern, we conduct an Inverse Hyperbolic Sine (IHS) transformation on the outcome variable, to generate asinh(VPI^w), and use this transformed variable as the outcome in a robustness test. Table A17 of the Appendix shows that our estimates are robust to this transformation as well.

Another key concern that can arise in this context is on the geographic scope of violence discourse: does VPI^w in remote areas primarily reflect local violence or discourse about distant events? To address this concern, we conduct a spatial spillover analysis examining whether VPI^w in grid cell i responds to violence in neighboring grid cells j. Table A18 presents results incorporating both own-grid violence and spatially-weighted neighboring violence at distance cutoffs of 100km, 200km, and 300km. We use Conley (1999) clustered standard errors to account for spatial autocorrelation in the error term. ACLED fatalities show modest spillover effects within 100km (Panel A, Column 1), but these dissipate beyond that distance. Homicides show no significant spillovers at any distance (Panel B). Critically, own-grid violence remains statistically significant across all specifications, indicating VPI responds primarily to local conditions.

Next we examine whether the spatial dynamics underlying the VPI^w differ for low vs high population areas. We therefore examine spillover effects in low-population and high-population grids separately. The results reveal striking heterogeneity. In low-population grids (Table A19), spillovers dominate. VPI^w responds primarily to violence in neighboring grids, with weaker own-grid effects. By contrast in high-population grids (Table A20), own-grid violence dominates, with minimal spillover effects. This heterogeneity reflects differences in information environments. In low-population areas, regional spillovers are substantively meaningful. Remote communities lack hyperlocal news infrastructure and often form functional economic regions spanning multiple grids. Violence in neighbouring areas affects travel safety, market access, and security perceptions across the entire region. In high-population areas, hyperlocal discourse dominates. Urban areas generate abundant neighborhood-specific content with geographically precise violence discussion, creating information saturation that overwhelms spillover effects. The heterogeneity demonstrates that VPI^w captures how violence perception actually operates across Mexico's urban-rural continuum.

ACLED reports violence related to battles, explosions/remote violence, protests, riots and violence against civilians. (It also reports “strategic violence” as an event category, but within our sample period, no fatalities are reported for this category) We dissect the ACLED data set to identify the number of fatalities for each event category, and use these as the predictors in Table A21 with the objective of understanding which event category generates the highest level of perceived violence. Interestingly, we observe that event categories “battle” and “violence against civilians” are those that have the highest and statistically significant impact. This aligns well with the idea that public violence perception, as captured by VPI, is particularly sensitive to violent events that directly affect the public.

Finally, in Table A22, we use an alternative dataset to identify violent events: the UCDP. UCDP records data on armed conflicts across the world, along with detailed information on date, location, actors involved and the number of fatalities, among others. Based on the actors involved, it classifies violence as state-based violence, nonstate violence and one-sided. We observe that there are no state-based violent events reported for Mexico for the sample period. 99% of the reported events are classified as non-state violence, with one-sided violence constituting only 1% of reported violent events. Critically, we observe that UCDP employs substantially higher fatality thresholds than ACLED, focusing exclusively on organized armed conflict with battle-related deaths meeting specific criteria. For our study context and period, this produces an extreme data sparsity challenge: 94% of grid-month observations (42,976 of 45,580) report zero UCDP fatalities, with

only 2,604 observations (6%) containing any recorded violence. This sparsity is even more pronounced when disaggregating by conflict type, making type-specific estimation statistically infeasible. Due to these reasons, we do not observe a statistically significant correlation between VPI and UCDP fatalities in Table A22..

Conclusion

This paper introduces the Violence Perception Index (VPI), which quantifies violence-related discourse, treated as a proxy for violence perceptions, through geolocated YouTube comments across Mexico (2020-2024). Systematic validation demonstrates strong correlations with established violence measures including ACLED fatalities and official homicide statistics. Critically, the VPI captures violence discourse with particular effectiveness in remote and marginalized areas systematically underrepresented by news-based conflict datasets.

Our analysis reveals that while VPI correlates with ACLED in high-population areas, it correlates with official violence statistics in low-population contexts where ACLED shows no relationship. Residual analysis reinforces this pattern: larger prediction errors in highly marginalized areas reflect the VPI's capacity to detect violence discourse in communities where formal documentation channels are weakest. This provides ground-level intelligence about violence perceptions in precisely those contexts—remote municipalities, marginalized regions, areas with weak state presence—where traditional monitoring systems fail.

The methodology is immediately scalable across languages and geographies, applicable to conflict-affected contexts globally. Future implementations could integrate multiple platforms for cross-platform validation or incorporate local languages to capture discourse in linguistically diverse settings.

However, a few limitations are worthy of discussion. The VPI captures discourse from social media users, who may differ systematically from general populations in age, urban residence, and digital literacy. Platforms are vulnerable to coordinated manipulation: organized groups may launch propaganda campaigns distorting violence perceptions. Government censorship may suppress violence discussion, leading the VPI to underreport in controlled information environments. While our spatial and temporal fixed effects partially address these concerns, users should interpret VPI patterns with awareness of potential distortions.

The VPI aggregates all violence-related discourse without distinguishing underlying sources. High VPI scores may reflect: (1) direct eyewitness accounts of local violence, (2) regional spillover discourse about violence in neighboring areas (as evidenced in Tables A18-A19), (3) discussion of nation-wide events or historical violence covered by mainstream media, or (4) rumors and speculation disconnected from actual events. We expect the balance between local and non-local discourse to shift depending on the salience of local versus national events at any given time. Our variance decomposition analysis (Table A12) suggests nation-wide events contribute relatively little to VPI variance: 97.6% of variation is spatial (between-grid) while only 2.7% is temporal (within-grid over time), indicating VPI primarily reflects grid-specific and regional dynamics rather than uniform national discourse. While geolocated videos are likely served by YouTube's recommendation algorithms to local audiences, suggesting comment sections function as localized "third spaces" (Wright, 2012)—we cannot fully decompose these signals in the current implementation, nor rule out that some violence discourse in remote areas references events occurring elsewhere. Disambiguating the geographic scope of violence references represents a natural extension of this work. Future implementations could employ large language models to classify whether comments reference local, regional, or national events, enabling separate indices that distinguish violence experienced locally from mediated discussion of distant events. Network analysis of comment patterns across grids might also reveal whether discourse spikes reflect coordinated responses to national news versus independent local reactions. These refinements would enable more targeted policy applications: distinguishing communities reacting to documented violence from those destabilized by disinformation.

Future work could employ supervised machine learning or LLM-based classification to separate discourse types. For instance, temporal proximity analysis (does VPI spike immediately after documented local events?), linguistic markers

(first-person experiential vs. reportorial language), or network analysis (organic discourse vs. coordinated campaigns) could distinguish local violence experience from mediated discussion of distant events. The recent work by Kushwaha et al. (2025) demonstrates that conflict classification can reveal shared drivers and correlations among events; similar approaches could help identify whether regions with similar VPI patterns share common underlying violence drivers (institutional weakness, criminal group presence, economic conditions) rather than merely responding to the same individual events. This would enable analysis of whether violence perception clusters align with substantive conflict groupings rather than administrative boundaries.

Finally, following the Total Error Framework for Digital Traces of Human Behavior on Online Platforms (Sen et al., 2021), several systematic errors warrant acknowledgment. However, the "third space" conceptualization (Wright, 2012) suggests that some apparent measurement concerns may be less problematic than they first appear. Because geolocated videos accumulate audiences connected by shared geographic locality rather than topical interest, comment sections function as localized gathering points where violence discourse emerges independently of video content. Nevertheless, platform coverage error remains: YouTube users skew younger and more urban than the general population, while commenters represent a further self-selection toward those willing to engage publicly. Our geographic query approach may also introduce trace selection error through ambiguity in video geolocation and poorly documented API return limits in densely populated areas. These errors are partially mitigated by population weighting and validation against survey-based perception measures, but the VPI should be understood as capturing violence discourse among digitally engaged populations rather than representative samples of all community members.

By demonstrating that systematic measurement of violence perception is feasible at scale, this work enables conflict monitoring that captures dynamics in remote and marginalized areas where traditional event-based datasets face known limitations. The VPI provides complementary intelligence for early warning systems and policy interventions, particularly in contexts where community perceptions may diverge from official violence records.

References

Amarasinghe, A. (2022). Diverting domestic turmoil. *Journal of Public Economics*, *208*, 104608. https://doi.org/10.1016/j.jpubeco.2022.104608

Amarasinghe, A. (2023). Public sentiment in times of terror. *Journal of Development Economics*, *162*, 103058. https://doi.org/10.1016/j.jdeveco.2023.103058

Arthur, R., Boulton, C. A., Shotton, H., & Williams, H. T. P. (2018). Social sensing of floods in the UK. *PLOS ONE*, *13*(1), e0189327. https://doi.org/10.1371/journal.pone.0189327

Bond, F., & Paik, K. (2012). A Survey of Wordnets and their Licenses. *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, 64–71. https://www.academia.edu/download/31907561/gwc2012.pdf#page=69

Consejo Nacional de Población. (2021). *Índices de marginación 2020* [Data set]. http://www.gob.mx/conapo/documentos/indices-de-marginacion-2020-284372

Conley, T.G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics*, *92*(12), 1-45. https://doi.org/10.1016/S0304-4076(98)00084-0

Data Cívica. (2024). *Votar entre balas* [Data set]. https://votar-entre-balas.datacivica.org/datos-votar-entre-balas

Davies, S., Engström, G., Pettersson, T., & Öberg, M. (2024). Organized violence 1989–2023, and the prevalence of

organized crime groups. *Journal of Peace Research*, *61*(4), 673–693. https://doi.org/10.1177/00223433241262912

Di Natale, A., & Garcia, D. (2024). LEXpander: Applying colexification networks to automated lexicon expansion. *Behavior Research Methods*, *56*(2), 952–967. https://doi.org/10.3758/s13428-023-02063-y

Gobierno de México. (2025). *Datos Abiertos de Incidencia Delictiva* [Data set]. https://www.gob.mx/sesnsp/acciones-y-programas/datos-abiertos-de-incidencia-delictiva

Gonzalez-Agirre, A., Laparra, E., & Rigau, G. (2012). Multilingual Central Repository version 3.0. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 2525–2529. https://adimen.ehu.eus/~rigau/publications/lrec12-glr.pdf

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., & Puranen, B. (2024). *World Values Survey Wave 7 (2017-2022) Cross-National Data-Set* (Version 6.0.0) [Data set]. World Values Survey Association. https://doi.org/10.14281/18241.24

Huang, B., & Carley, K. M. (2020). A large-scale empirical study of geotagging behavior on Twitter. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '19*, 365–373. https://doi.org/10.1145/3341161.3342870

Institute for Economics & Peace. (2025). *Mexico Peace Index 2025: Identifying and Measuring the Factors That Drive Peace* [Data set]. https://www.visionofhumanity.org/resources/

Kushwaha, N., Oh, W. S., Shah, S., & Lee, E. D. (2025). Data-driven conflict classification exposes weak predictive indicators. *Royal Society Open Science*, *12*(12), 250897. https://doi.org/10.1098/rsos.250897

Murray, C., & Stott, M. (2025, February 9). Mexico tries to tame 'monster' cartels to please Donald Trump. *Australian Financial Review*. https://www.afr.com/world/north-america/mexico-tries-to-tame-monster-cartels-to-please-donald-trump-20250209-p5lap3

Newman, N., Fletcher, R., Robertson, C. T., Ross Arguedas, A., & Nielsen, R. K. (2024). *Reuters Institute digital news report 2024*. Reuters Institute for the Study of Journalism. https://doi.org/10.60625/RISJ-VY6N-4V57

Oswald, C., & Ohrenhofer, D. (2022). Click, click boom: Using Wikipedia data to predict changes in battle-related deaths. *International Interactions*, *48*(4), 678–696. https://doi.org/10.1080/03050629.2022.2061969

Pebesma, E. J. (2004). Multivariable geostatistics in S: The gstat package. *Computers & Geosciences*, *30*(7), 683–691. https://doi.org/10.1016/j.cageo.2004.03.012

Pew Research Center. (2023). *Spring 2023 Survey Data* [Data set]. https://www.pewresearch.org/dataset/spring-2023-

survey-data/

Raleigh, C., Kishi, R., & Linke, A. (2023). Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. *Humanities and Social Sciences Communications*, *10*(1), 74. https://doi.org/10.1057/s41599-023-01559-4

Rohman, A., & Ang, P. H. (2019). Communication, Culture, and Governance in Asia| Truth, Not Fear: Countering False Information in a Conflict. *International Journal of Communication*, *13*, 16–16.

Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. *Public Opinion Quarterly*, *85*(S1), 399–422. https://doi.org/10.1093/poq/nfab018

Shaver, A., Kazis-Taylor ,Hannah, Loomis ,Claudia, Bartschi ,Mia, Patterson, Paul, Vera ,Adrian, Abad ,Kevin, Alqarwani ,Saher, Bell ,Clay, Bock ,Sebastian, Cabezas ,Kieran, Felix ,Heidi, Gonzalez ,Jennifer, Hoeft ,Christopher, Ibarra Martinez ,Aileen, Keltner ,Kai, Moroyoqui ,Jessica, Paman ,Kieko, Ramirez ,Ethan, … and Eskander, M. (2023). Expanding the Coverage of Conflict Event Datasets: Three Proofs of Concept. *Civil Wars*, *25*(2–3), 367–397. https://doi.org/10.1080/13698249.2023.2254988

Tollefsen, A. F., Strand, H., & Buhaug, H. (2012). PRIO-GRID: A unified spatial data structure. *Journal of Peace Research*, *49*(2), 363–374. https://doi.org/10.1177/0022343311431287

Weidmann, N. B. (2016). A Closer Look at Reporting Bias in Conflict Event Data. *American Journal of Political Science*, *60*(1), 206–218. https://doi.org/10.1111/ajps.12196

WorldPop. (2020). *Global high resolution population denominators Project Funded by The Bill and Melinda Gates Foundation* (No. OPP1134076) [Data set]. https://doi.org/10.5258/SOTON/WP00660

Wright, S. (2012). From ‘Third place’ to ‘Third space’: Everyday political talk in non-political online spaces. *Javnost - The Public*, *19*(3), 5–20. https://doi.org/10.1080/13183222.2012.11009088

Online Appendix

Geographic framework

Grid construction

The data collection process begins with WorldPop population estimates for 2020 at approximately 1-kilometre resolution (0.0083 degrees, ~925 metres) (WorldPop, 2020). These data are aggregated to approximately 10-kilometre resolution (0.0833 degrees, ~9250 metres) by summing population counts across 10×10 cell blocks, creating grid-cells that align with our spatial sampling framework. To focus on areas with substantial human activity, only grid-cells containing more than 100 inhabitants are retained for data collection. This approach ensures computational efficiency whilst maintaining coverage of populated areas where meaningful discourse occurs.

The aggregated raster yielded 26,227 grid-cells covering Mexico's geographic extent. Applying the population threshold of 100 inhabitants, 14,266 cells (54.4%) were retained for data collection, whilst 11,961 cells (45.6%) with sparse population were excluded. Despite excluding nearly half of all grid cells, the selected cells captured 140,343,624 inhabitants, representing 99.8% of Mexico's total population (140,555,068) in the dataset. For each selected grid-cell, a centroid coordinate was calculated and used to generate a circular search radius of 7,070 metres (covering approximately 157.0 km² per query) for querying the YouTube Data API. The resulting 14,266 search circles provide systematic geographic coverage of Mexico's inhabited regions, with overlapping coverage ensuring comprehensive video capture across populated areas.

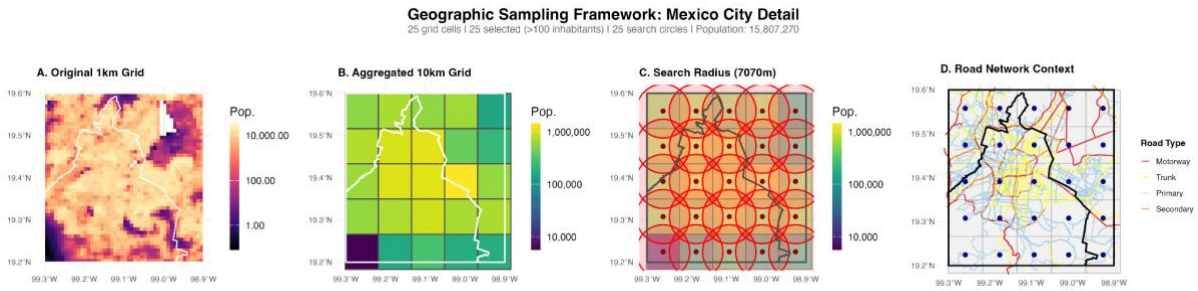


Figure A1: Geographic sampling framework

Spatial query parameters

For each retained grid cell, videos are queried within a 7070-metre radius from the cell centroid.

Three-stage API collection process

Stage 1: Video discovery via "search" endpoint

For each query, we submitted a request to the YouTube Data API v3 "search" endpoint (https://www.googleapis.com/youtube/v3/search) specifying only the following parameters (and notably no search term):

- *location* and *locationRadius*: Specified the circular geographic search area using the grid-cell centroid coordinates (e.g., "19.4326,-99.1332") and a fixed radius of 7070 metres.

- *publishedAfter* and *publishedBefore*: Defined the temporal boundaries of the search query, limiting results to videos published within the specified timeframe (1 January 2020 to 30 June 2024 for the Mexico dataset).

Each query returned videos whose metadata indicated a geographic location within the specified circular area, enabling systematic spatial coverage across all retained grid-cells.

This endpoint returns essential video identifiers and basic metadata including:

- *videoId* (unique identifier for subsequent queries)
- *title* (video title as provided by uploader)
- *description* (user-provided video description)
- *publishedAt* (video upload timestamp)

For the Mexico pilot study covering January 2020 to June 2024, this process yielded 3,398,143 search results across all grid-cells.

Stage 2: Enhanced metadata via "videos" endpoint

Subsequently, the "videos" endpoint (https://www.googleapis.com/youtube/v3/videos) extracts comprehensive metadata for each video identified in Stage 1. This endpoint provides enriched information essential for analysis including:

- *viewCount* (total video views for engagement metrics)
- *commentCount* (total comments for discourse volume assessment)
- Additional metadata fields for content categorisation

After deduplicating overlapping content (videos appearing in multiple grid-cells) and filtering restricted content, the Mexico pilot extracted 1,229,594 unique videos. This deduplication results either from the overlap of circular areas (see Figure A1 above) or for the association of a video with more than one circular area.

Stage 3: Comment Extraction via "commentThreads" Endpoint

For videos receiving at least one comment, the "commentThreads" endpoint ("https://www.googleapis.com/youtube/v3/commentThreads") retrieves all associated comments and their metadata. This endpoint captures:

- Complete comment text for natural language processing
- Commenter metadata (where available and permissible)
- Comment timestamps for temporal analysis
- Reply thread structures

The Mexico pilot yielded 14,785,647 unique comments across 508,942 videos, generating 43,900,491 unique comment-cell spatial associations due to videos spanning multiple grid-cells.

Validation of geospatial metadata frequency

Methodology for Geotagging Frequency Analysis

To estimate the prevalence of explicit geographic metadata in user-generated content, we conducted a systematic audit using the YouTube Data API v3. The sampling strategy utilized a "Stopword-Query" approach designed to capture a representative cross-section of general activity within the target region without biasing the sample toward specific news events, viral trends, or professional media outlets.

The search queries consisted of nine ubiquitous Spanish-language functional words (stopwords): de, la, que, el, en, y, a, los, and del. By querying these words, we retrieved a broad spectrum of content relevant to the Mexican geographic context, restricted by the API's *regionCode="MX"* parameter.

A two-stage verification process was implemented to identify geotagged content:

1. Search Stage: For each keyword, the *search().list* endpoint was used with pagination (utilizing the nextPageToken) to retrieve 100 video IDs per word per iteration.
2. Verification Stage: The resulting IDs were passed to the *videos().list* endpoint to parse the recordingDetails.location object. A video was categorized as "Geotagged" only if it contained valid latitude/longitude coordinate metadata provided by the uploader.

To mitigate temporal bias, we executed the search protocol in four discrete runs. The results demonstrate that explicit coordinate-level geotagging is an infrequent phenomenon in the Mexican YouTube ecosystem.

Table A1: Systematic Audit of Geotagged Content Frequency

Run Sequence	Total Videos Analyzed (N)	Geotagged Videos (n)	Geotagging Rate (%)
Run 1	825	31	3.76%
Run 2	425	14	3.29%
Run 3	650	20	3.08%
Run 4	625	22	3.52%
Grand Total	**2,525**	**87**	**3.45%**

Video category analysis

Although this information is not used in computing the VPI, we still present in Table A2 and Figure A2 the descriptive statistics for YouTube videos by category across the Mexico dataset (January 2020–May 2024). Categories are either set by the video creators or automatically identified by YouTube once the video is uploaded. The table summarises the distribution of 3,398,143 search results and 1,229,594 unique videos across YouTube's content categories. For each category, we report the number of videos, percentage of total videos, total comments received, and comment engagement metrics including mean, median, and standard deviation.

The data reveal substantial variation in both video prevalence and engagement across categories, with People & Blogs, Music, and Entertainment comprising the largest shares of geolocated content. Notably, News & Politics videos receive on average by far the highest number of comments per video, although this category also exhibits the largest standard deviation, indicating considerable heterogeneity in engagement levels. This pattern suggests that whilst some news-related videos generate intensive discussion and debate, others receive minimal interaction.

Table A2: Videos and Comments by Video Category

Category	N Videos	% of Total	Total Comments	Mean Comments	Median Comments	SD Comments
People & Blogs	655253	53.30%	4358853	6.7	0	136.2
Music	140929	11.50%	1556861	11	1	311.8
Entertainment	113675	9.20%	3376559	29.7	1	440.3
Travel & Events	92153	7.50%	1812657	19.7	1	137.1
Education	42980	3.50%	782310	18.2	0	195.7
Sports	37794	3.10%	509692	13.5	0	401.4
Gaming	33893	2.80%	542068	16	1	106.2
News & Politics	27639	2.20%	1437948	52	0	401.2
Autos & Vehicles	17934	1.50%	389250	21.7	2	100.9
Howto & Style	15396	1.30%	377421	24.5	2	221.3
Film & Animation	13931	1.10%	95587	6.9	0	58.9
Science & Technology	11252	0.90%	257091	22.8	0	206.7
Comedy	11146	0.90%	163694	14.7	0	130.9
Nonprofits & Activism	8718	0.70%	74300	8.5	0	113.1
Pets & Animals	6901	0.60%	125726	18.2	1	140.4

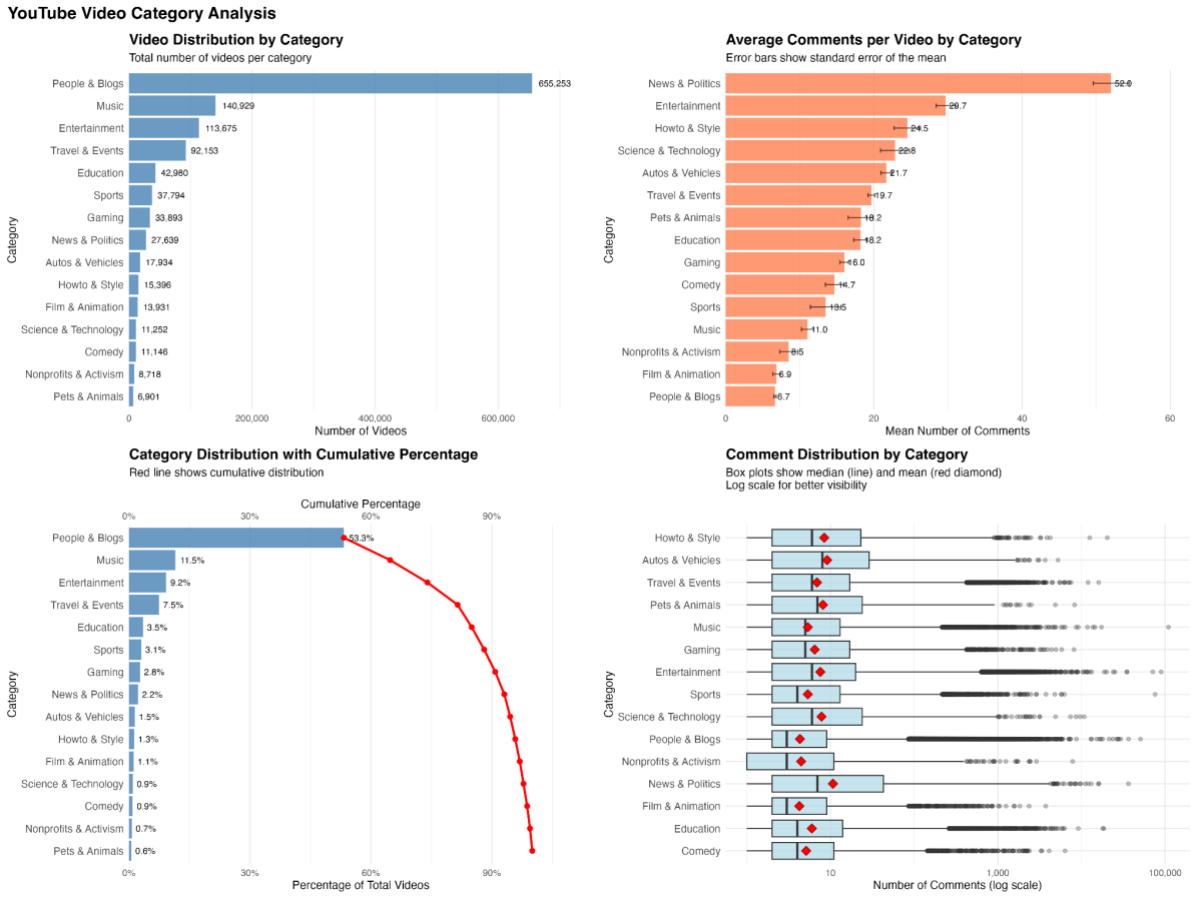


Figure A2: Videos and Comments by Video Category

Violence Perception Index construction

To quantify violence intensity in user comments, we constructed a weighted Spanish-language dictionary using a seed-based expansion approach with the NLTK WordNet interface and Open Multilingual Wordnet (OMW-1.4) (Bond & Paik, 2012; Gonzalez-Agirre et al., 2012). Beginning with ten manually selected seed words representing core violence concepts ("violencia", "asesinato", "homicidio", "tiroteo", "ataque", "enfrentamiento", "balacera", "secuestro", "narcotráfico", "delincuencia"), we systematically expanded the dictionary by traversing WordNet's semantic network of synonyms and related terms. The expansion proceeded iteratively to a maximum depth of two levels, with each successive level assigned exponentially decaying weights to reflect decreasing semantic proximity to core violence concepts. Seed words received a weight of 1.0, direct synonyms and related terms at depth 1 received 0.5, and terms at depth 2 received 0.25. This weighted approach ensures that terms more closely related to explicit violence contribute more strongly to the violence measure whilst capturing peripheral violence-related discourse. The resulting dictionary is presented in Table A3.

Table A3: Weighted Spanish keyword dictionary for Violence Perception Index (VPI) calculation

Word	Weight	Word	Weight	Word	Weight
violencia	1	eficacia	0.25	vía_de_acceso	0.25
asesinato	1	potencia	0.25	arranque	0.25

homicidio	1	energía	0.25	arrebato	0.25
tiroteo	1	vigor	0.25	chorreo	0.25
ataque	1	influencia	0.25	ebullición	0.25
enfrentamiento	1	peso	0.25	efusión	0.25
balacera	1	asesinato_mafioso	0.5	estallido	0.25
secuestro	1	golpe	0.5	explosión	0.25
narcotráfico	1	exitazo	0.25	robo_armado	0.25
delincuencia	1	triunfar	0.25	encuentro	0.5
fuerza	0.5	triunfo	0.25	escaramuza	0.5
agresividad	0.5	éxito	0.25	roce	0.5
furia	0.5	amenazador	0.25	cara_a_cara	0.5
rabia	0.25	codazo	0.25	conflicto	0.5
cólera	0.25	empuje	0.25	confrontación	0.5
ira	0.25	empujón	0.25	pugna	0.5
coraje	0.25	movimiento_brusco	0.25	batalla_campal	0.25
valor	0.25	que_pincha	0.25	refriega	0.25
ánimo	0.25	golpe_de_estado	0.25	afán	0.25
intensidad	0.25	golpetazo	0.25	esfuerzo	0.25
volumen	0.25	impacto	0.25	esmero	0.25

fuerza_muscular	0.25	golpe_de_efecto	0.25	acción	0.25
músculo	0.25	conmoción	0.25	batalla	0.25
nervio	0.25	disgusto	0.25	combate	0.25
poder	0.25	choque	0.25	lucha	0.25
resistencia	0.25	porrazo	0.25	aparentemente	0.25
topetazo	0.25	rapto	0.5	ostensiblemente	0.25
toque	0.25	ofensa	0.25	reunión	0.25
cinturón	0.25	momento_decisivo	0.25	acercamiento	0.25
derribo	0.25	entrada	0.25	partido	0.25
golpe_fuerte	0.25	paso	0.25	altercado	0.25
leche	0.25	portal	0.25	discusión	0.25
tortazo	0.25	puerta	0.25	encontrón	0.25
cardenal	0.25	boca	0.25	palabras	0.25
moratón	0.25	ingreso	0.25	pelea	0.25
asalto	0.5	pelotera	0.25	riña	0.25
atentado	0.5	pleito	0.25	contacto	0.25
acceso	0.5	pincelada	0.25	lucha_libre	0.25
crisis	0.5	tacto_ligero	0.25	oposición	0.25
ictus	0.5				

Text Processing and Lemmatisation

For each comment in the dataset, we applied natural language processing to compute a scalar violence score. Using the SpaCy[3] Spanish language model (es_core_news_sm), we lemmatised each comment to reduce inflected words to their base forms, enabling consistent matching with dictionary terms regardless of grammatical variations (e.g., "mataron", "mató", "matando" all reduce to "matar"). The lemmatisation process converted all text to lowercase and processed each comment through SpaCy's linguistic pipeline, extracting lemmas for all tokens. This normalisation step is crucial for Spanish text analysis given the language's rich morphological variation, ensuring that different conjugations, declensions, and grammatical forms of the same root concept are properly identified and weighted.

Scalar score calculation

The scalar violence score for each comment was calculated by identifying all lemmatised tokens present in the weighted dictionary and summing their associated weights, with repeated occurrences of the same term counted multiple times to reflect intensity. For example, a comment containing "violencia" (weight 1.0) twice and "tiroteo" (weight 1.0) once would receive a scalar score of 3.0. This process was parallelised across multiple CPU cores using Python's multiprocessing library to efficiently handle large comment volumes. Only comments containing at least one dictionary term received a score; comments with no matches were excluded from subsequent analysis. The resulting dataset comprised comment identifiers, associated video metadata, publication timestamps, individual word counts with weights, and the aggregated scalar sum, the comment's violence score, enabling both granular analysis of specific violence terminology and overall intensity assessment across the corpus.

Geographic Aggregation of Scores

To construct spatially continuous violence perception measures across the study area, we employed Inverse Distance Weighting (IDW) interpolation (Pebesma, 2004) to aggregate comment-level scalar scores to a regular 10-kilometre grid. IDW is a deterministic spatial interpolation method that estimates values at unsampled locations by computing a weighted average of known values from nearby sampled points, with weights inversely proportional to the distance between the interpolation point and the sampled locations. Specifically, for each grid-cell centroid, we calculated the interpolated score score using the formula:

$$VPI_c = \frac{\sum_{(i=1)}^{n} \frac{(comment\ score_i)}{d_i^p}}{\sum_{(i=1)}^{n} \frac{1}{d_i^p}}$$

where VPI_c represents the violence perception index in grid-cell c, d the distance from the grid-cell c centroid to the comment's geographic location (via its parent video coordinates), n is the number of comments for a given time window, and p serving as the power parameter. We set the power parameter to 2 (idp = 2), meaning that a comment's influence on a grid cell's score decreases with the square of the distance, ensuring that nearby comments contribute more strongly to a cell's violence perception index than distant ones whilst still incorporating information from the broader spatial context.

This approach addresses the challenge that individual comments are geolocated through their parent videos' coordinates, which may not align precisely with grid-cell centroids or boundaries. By using IDW interpolation, we create a smooth, continuous surface of violence perception intensity where each grid-cell's score reflects a distance-weighted average of all surrounding comments, with closer comments exerting greater influence. Notably, we interpolate the violence scores without covariates (using only the spatial coordinates). This method is particularly appropriate for our application because it preserves local variation in violence discourse whilst avoiding sharp discontinuities at arbitrary grid-cell boundaries, and

[3] https://spacy.io/

it naturally handles the spatial uncertainty inherent in user-generated geographic data where the precise relationship between a video's recorded location and the geographic scope of its associated comments may be ambiguous.

Figure A3 illustrates the different steps of the interpolation process. In Step 1, we identify a number of comments with their corresponding score based on the weighted-term frequency approached detailed above. Step 2, visualise the distance from all the comments to a selected grid-cell. Distances will also be calculated for all the other grid-cells to map. The result of the interpolation for all the grid-cells is shown in Step 3.

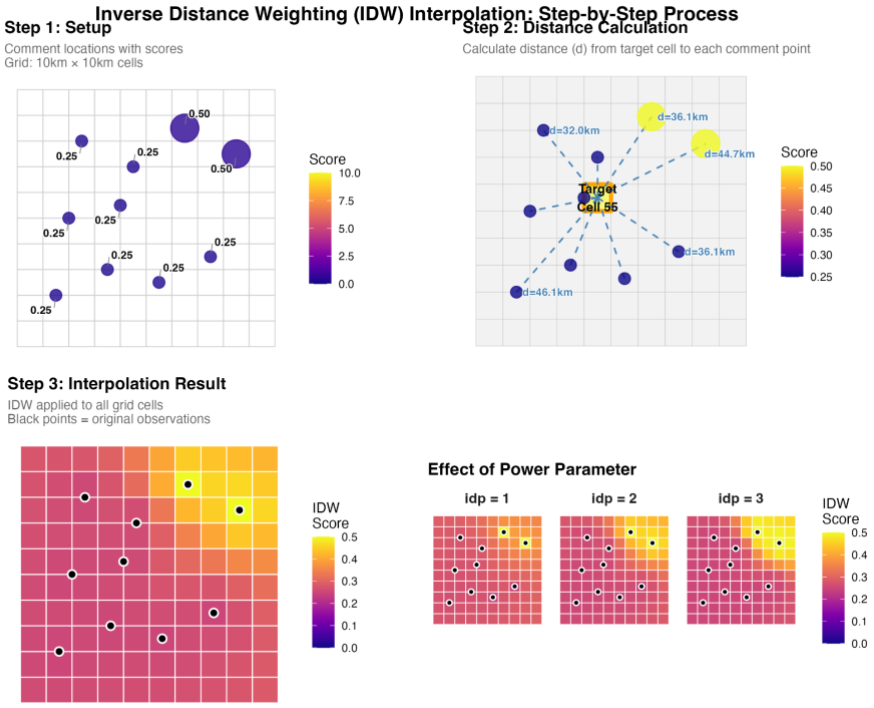


Figure A3: Step by step visualisation of the geographic interpolation process.

We note that our approach aggregates continuous violence scores rather than binary classifications, which preserves gradations in intensity. However, formal quantification methods that explicitly account for classification uncertainty may further improve geographic estimates (Schumacher et al., 2025), representing an avenue for future methodological refinement.[4]

Length of Comments and Association with Violence Score

Based on the quantile distribution of comment word counts, YouTube comments in this dataset are predominantly brief, with a median length of just 9 words, while 95% of comments contain 47 words or fewer, and 99% stay under 97 words.

As shown in Table A4, comments containing violence-related terms (positive scalar scores) are longer than those without such terms, with median lengths of 23 versus 8 words respectively. This positive correlation between comment length and violence score is to be expected, as the probability of encountering violence-related terms naturally increases with the

[4] Schumacher, T., Strohmaier, M., & Lemmerich, F. (2025). A Comparative Evaluation of Quantification Methods. *Journal of Machine Learning Research*, *26*(55), 1–54.

total number of words in a comment, simply due to greater lexical coverage. The word count across the two groups is also visualised in Figure A4.

			Word count				
Positive score	**n**	**%**	**mean**	**median**	**SD**	**min**	**max**
TRUE	1,125,129	7.6096	37.23874	23	56.66629	1	2198
FALSE	13,660,518	92.3904	13.26367	8	18.22126	0	3402

Table: A4 Number of words in comments for comment with a positive and nill violence scores.

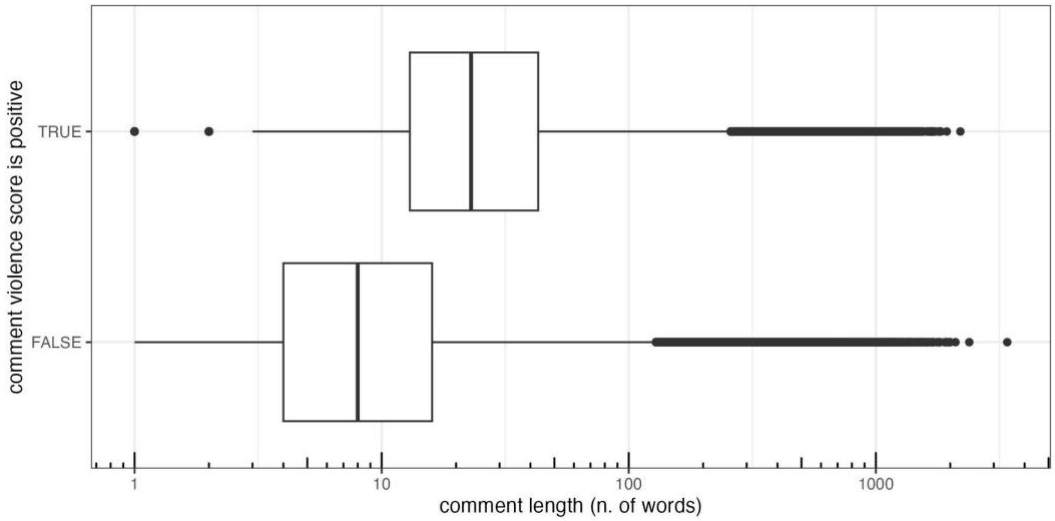


Figure A4. Number of Words per Comment (n = 14,785,647)

Within the group of comments assigned to a positive scalar score, a legitimate concern is about a possible linear and strong correlation between word count and scalar score. Potentially, instead of measuring the intensity of a comment reference to a violent episode or about a conversation on violence, the score could simply measure the length of the comment. We address this concern with a visual analysis and then with correlation analysis. In Figure A5 we note that indeed the two measures are positively correlated and yet not linearly. Correlation appears significantly stronger only as the number of words passes 100.

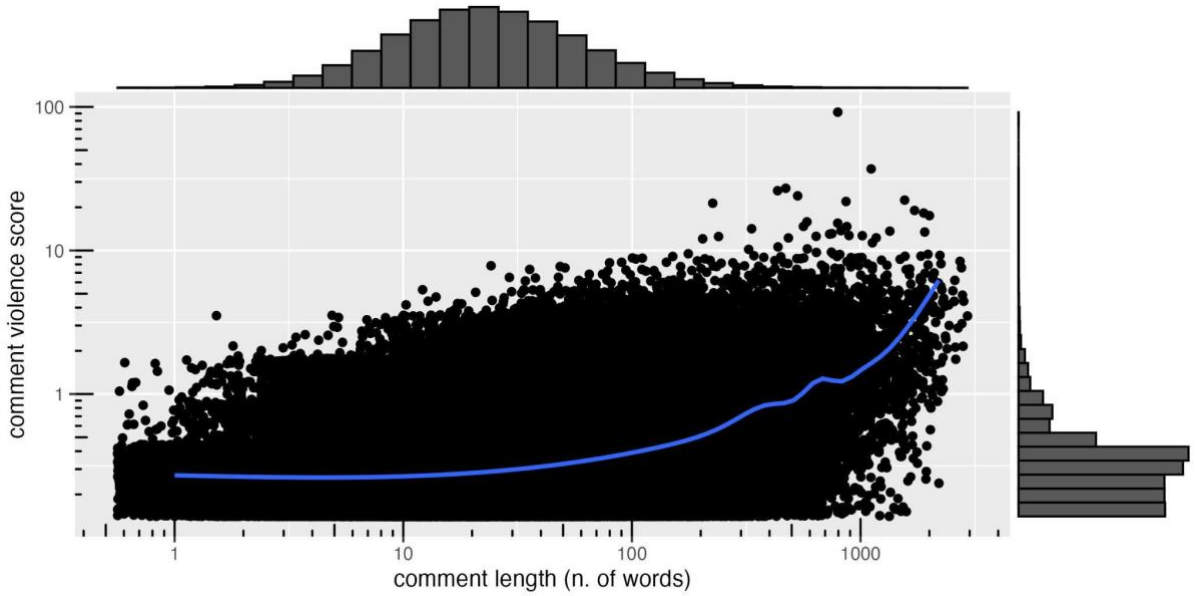


Figure A5. Correlation between word count and violence score (comment position was jittered to improve visibility).

We test for non-linearity in the relationship between word count and violence score by estimating correlations across different comment length ranges (Table A5). While the correlation remains consistently positive and significant across all length categories, its magnitude varies substantially: for comments under 25 words (comprising the majority of our data), the correlation is only 0.1, increasing to 0.23 for comments under 100 words (covering approximately 95% of all comments). This weak to modest correlation indicates that word count explains only a small fraction of variance in violence scores, meaning that our scalar scoring approach, which sums violence terms without normalizing by length, does not systematically overweight longer comments in practice.

Word count	**Correlation**	**% of comments**
Any	0.418753	100
< 25	0.096625	52.55
< 50	0.161426	79.82
< 100	0.225333	94.21
100+	0.401484	5.79

Table A5: Correlation between word count and scalar score for different comment lengths.

Validation of Violence Score with LLM labelling

To validate our violence scoring approach, we drew a stratified random sample of comments based on their scalar violence scores. This sampling strategy ensured adequate representation across the full spectrum of content, from comments with no violence references to those with predominant violence themes.

We first created an initial binning structure with ten categories: one bin for comments with zero violence score, followed by nine equal-width bins spanning the range of positive scores. However, the three highest bins (bins 7, 8, and 9) contained relatively few comments due to the right-skewed distribution of violence scores. To ensure adequate sample sizes while maintaining representation of high-violence content, we collapsed these three bins into a single category, resulting in a final seven-bin structure:

- **Bin 0**: Score = 0 (no violence detected)
- **Bins 1-5**: Equal-width intervals covering low-to-moderate violence scores
- **Bin 6**: Combined high-violence category (collapsed from original bins 7-9)

From each of the seven bins, we randomly sampled 100 comments, yielding a total validation sample of 700 comments. This stratified approach served two important purposes. First, it ensured sufficient representation of high-violence comments, which are relatively rare in the full dataset but theoretically important for validation. Second, it provided adequate statistical power to assess agreement across the entire range of violence intensity, rather than being dominated by the most common (low-violence) cases.

The stratified design allows us to examine whether agreement between our manual coding and LLM classifications remains consistent across different levels of violence intensity, or whether certain types of content (e.g., subtle violence references versus explicit violent language) present greater challenges for automated detection. This approach provides a more rigorous validation than simple random sampling would offer, as it explicitly tests the measurement system's performance across theoretically relevant subgroups.

LLM Prompting Strategy

All four LLM models received identical prompts to ensure comparability of results. The prompting approach was designed to elicit structured, consistent responses that directly parallel a manual coding approach while minimizing model-specific interpretation biases.

Each comment was submitted to the LLM with a two-part prompt structure. The system message established the role: "You are a research assistant analyzing social media comments. Always respond with valid JSON only." This framing emphasized the analytical nature of the task and the requirement for structured output.

The user prompt provided the comment text and explicit instructions for two classification tasks matching our validation objectives:

1. **Binary classification**: "Does the comment discuss violence or a violent episode? (true/false)"
2. **Frequency rating**: "How frequent is the reference to violence from 0 (not at all) to 10 (predominant)"

The prompt explicitly requested responses in JSON format with three required fields: *discusses_violence* (boolean), *frequency_score* (0-10 integer), and *brief_explanation* (one sentence justification). This structured format enabled automated parsing and eliminated ambiguity in response interpretation.

To maximize response consistency, we set the temperature parameter to 0.2 across all models, favoring deterministic outputs over creative variation. We limited responses to 200 tokens to encourage concise classifications and prevent lengthy elaborations that might complicate parsing.

The prompt emphasized responding "ONLY with valid JSON" and "nothing else," explicitly instructing models not to add conversational elements, explanations outside the JSON structure, or markdown formatting. When models occasionally deviated from this format (e.g., wrapping JSON in markdown code blocks), our processing pipeline included fallback parsing logic to extract the structured data.

Limitations of Cross-Method Comparison

It is important to acknowledge fundamental differences between our scalar scoring approach and the LLM-based validation that limit direct comparability. Our scalar score is constructed as a simple sum of violence-related terminology instances identified through dictionary-based keyword matching. This approach, which is computationally efficient given the task of scoring more than 14 million comments, counts discrete lexical occurrences without considering semantic context, rhetorical function, or the overall discourse structure of comments. In contrast, LLM models perform holistic semantic analysis, interpreting violence references within their linguistic and contextual environment. LLMs can distinguish between reporting violence, advocating violence, or using violent metaphors—nuances our dictionary-based approach does not capture.

Consequently, the validation exercise compares two fundamentally different measurement philosophies: frequency-based lexical detection versus contextual semantic interpretation. Our scalar score may register multiple violence terms in a single comment as indicating high violence intensity, while an LLM might recognize these as metaphorical or rhetorical rather than literal violence discussion. Conversely, an LLM might identify implicit violence themes that lack explicit violence terminology, which our dictionary approach would miss entirely.

Binary Classification Agreement

We first compared agreement on the binary classification task (whether a comment discusses violence or not). To enable comparison with LLM binary classifications, we converted our scalar scores to binary indicators using a threshold of 1.0: comments scoring above 1.0 were classified as discussing violence, while those at or below 1.0 were classified as not discussing violence. Table A6 presents Cohen's Kappa coefficients measuring pairwise agreement between all raters, including our scalar score.

Qwen-Qwen3-vl-4b	Ibm-Granite-3.2-8b	Google-Gemma-3n-E4b	Meta-Llama-3-8b-Instruct	Scalar score
1	0.8368012	0.86796383	0.75587508	0.62405929
0.8368012	1	0.81760134	0.78454929	0.58863899
0.86796383	0.81760134	1	0.73660206	0.61804424
0.75587508	0.78454929	0.73660206	1	0.52097916
0.62405929	0.58863899	0.61804424	0.52097916	1

Table A6: Kappa matrix comparing agreement on binary coding of comment (discuss/don't discuss violence).

The results show substantial to almost perfect agreement across all model pairs. Inter-model agreement ranges from $\kappa = 0.74$ (Google-Gemma vs Meta-Llama) to $\kappa = 0.87$ (Qwen vs Google-Gemma), indicating that different LLMs identify

violence with high consistency. Critically, agreement between our scalar score and the LLM models ranges from $\kappa = 0.52$ (Meta-Llama) to $\kappa = 0.62$ (Qwen), suggesting moderate agreement. The overall Fleiss' Kappa across all five raters (four LLMs plus our scalar score) is 0.800 ($p < 0.0001$), indicating substantial overall agreement and validating the consistency of violence detection across human and automated coding methods.

Table A7 complements these findings by presenting raw agreement rates. LLM models agree with each other 87-94% of the time, while agreement between LLMs and our scalar score ranges from 75-81%.

Model 1	Model 2	Agreement %
Qwen-Qwen3-vl-4b	Ibm-Granite-3.2-8b	92.2137405
Qwen-Qwen3-vl-4b	Google-Gemma-3n-E4b	93.5877863
Qwen-Qwen3-vl-4b	Meta-Llama-3-8b-Instruct	88.3969466
Ibm-Granite-3.2-8b	Google-Gemma-3n-E4b	91.2977099
Ibm-Granite-3.2-8b	Meta-Llama-3-8b-Instruct	90.0763359
Google-Gemma-3n-E4b	Meta-Llama-3-8b-Instruct	87.480916
Qwen-Qwen3-vl-4b	Scalar score	80.9160305
Ibm-Granite-3.2-8b	Scalar score	78.9312977
Google-Gemma-3n-E4b	Scalar score	80.610687
Meta-Llama-3-8b-Instruct	Scalar score	75.4198473

Table A7: Raw pairwise agreement rates for all models

Continuous Score Agreement

Beyond binary classification, we examined agreement on the continuous violence frequency scores. This analysis required addressing the scale difference between our scalar sum (which is unbounded and derived from term counts) and the LLM scores (bounded 0-10 reflecting subjective intensity judgments). We therefore applied three transformation methods: z-score standardization (mean=0, sd=1), min-max scaling (normalizing our score to 0-10), and rank transformation (converting to percentile rankings).

Table A8 presents correlation coefficients between the continuous scores under different transformations. Pearson correlations, which measure linear relationships, range from $r = 0.35$ (Qwen and IBM-Granite) to $r = 0.44$ (Google-Gemma and Meta-Llama), indicating moderate positive associations. Critically, these values remain constant across original, z-score, and min-max transformations because Pearson correlation is invariant to linear transformations. However, Spearman rank correlations show substantially stronger relationships, ranging from $\rho = 0.61$ (IBM-Granite) to $\rho = 0.68$ (Google-Gemma). This 20-30 percentage point increase suggests that while the absolute value mappings differ between our lexical counting approach and LLM semantic judgments, the ordinal relationships, which comments contain more or less violence, are well preserved across methods.

Model	Method	Pearson_r	Spearman_rho

Qwen-Qwen3-vl-4b	Original	0.35461691	0.66276043
Qwen-Qwen3-vl-4b	Z-score	0.35461691	0.66276043
Qwen-Qwen3-vl-4b	Min-Max	0.35461691	0.66276043
Qwen-Qwen3-vl-4b	Rank	0.66276043	0.66276043
Ibm-Granite-3.2-8b	Original	0.35375722	0.61407855
Ibm-Granite-3.2-8b	Z-score	0.35375722	0.61407855
Ibm-Granite-3.2-8b	Min-Max	0.35375722	0.61407855
Ibm-Granite-3.2-8b	Rank	0.61407855	0.61407855
Google-Gemma-3n-E4b	Original	0.43735454	0.6779434
Google-Gemma-3n-E4b	Z-score	0.43735454	0.6779434
Google-Gemma-3n-E4b	Min-Max	0.43735454	0.6779434
Google-Gemma-3n-E4b	Rank	0.6779434	0.6779434
Meta-Llama-3-8b-Instruct	Original	0.433003	0.64179281
Meta-Llama-3-8b-Instruct	Z-score	0.433003	0.64179281
Meta-Llama-3-8b-Instruct	Min-Max	0.433003	0.64179281
Meta-Llama-3-8b-Instruct	Rank	0.64179281	0.64179281

Table A8: Correlation coefficients between LLM continuous scores and our scalar score

The Intraclass Correlation Coefficient (ICC) provides a more stringent test of absolute agreement between continuous measures (Table A9). ICC values vary dramatically depending on the transformation method, revealing important insights about the nature of disagreement between approaches. With z-score standardization, ICC values range from 0.35 to 0.44, indicating fair to moderate agreement. Min-max scaling produces very low ICC values (0.07-0.10), suggesting that even after aligning the numeric ranges, the distributions remain fundamentally different—our scores likely show a highly right-skewed distribution with many low values and few extreme values, while LLM scores may be more uniformly distributed across the 0-10 scale. However, rank-based ICC values are substantially higher (0.61-0.68), nearly identical to the Spearman correlations and confirming that the relative ordering of comments by violence intensity is highly consistent between our dictionary-based coding and all four LLM approaches. This pattern indicates that distributional differences, rather than disagreement about relative violence levels, account for most of the apparent disagreement. Both measurement approaches effectively identify which comments are more or less violent; they simply quantify that "more-ness" on different numerical scales with different distributional properties.

Model	**ICC zscore**	**ICC minmax**	**ICC rank**
Qwen-Qwen3-vl-4b	0.35496667	0.0659397	0.66035087

Ibm-Granite-3.2-8b	0.35410659	0.07395248	0.60790836
Google-Gemma-3n-E4b	0.43773055	0.08085125	0.67820294
Meta-Llama-3-8b-Instruct	0.43337815	0.09895366	0.6370184

Table A9: Interclass correlation between LLM continuous scores and our scalar score after three transformations: z-score, min-max and ranking.

Moreover we note that methodological work suggests that direct pointwise LLM scoring can suffer from bunching around arbitrary numbers, with pairwise comparisons or token-probability-weighted approaches potentially yielding more reliable scalar measurements (Licht et al., 2025).[5]

The full set of 700 comments annotated by the LLMs is available upon request in the replication package.

Comparison with Fine-Tuned BERT Classification

To evaluate whether transformer-based classification would improve upon our dictionary approach, we fine-tuned a Spanish BERT model (BETO: dccuchile/bert-base-spanish-wwm-cased; Cañete et al., 2023)[6] using the LLM-consensus annotations from our 700-comment validation sample. We trained separate models for binary classification (violence/no violence) and continuous intensity prediction (0–10 scale).

Classification of the full corpus (14.8 million comments) required over one week of computation on an Apple M1 processor, raising concerns about the approach's feasibility for near-real-time monitoring applications.

More critically, BERT exhibited systematic overclassification. Table A10 presents the comparison between dictionary and BERT binary classifications.

	BERT: No Violence	**BERT: Violence**
Dictionary: No Violence	14,183,376	555,117
Dictionary: Violence	15,624	32,336

Table A10: Confusion Matrix—Dictionary vs. BERT Classification

The dictionary approach flagged 0.3% of comments as violence-related, while BERT flagged 4.0%. Agreement metrics indicate low concordance: Cohen's $\kappa = 0.096$, though raw agreement was high (96.1%) due to the predominance of non-violent comments.

[5] Licht, H., Sarkar, R., Wu, P. Y., Goel, P., Stoehr, N., Ash, E., & Hoyle, A. M. (2025). Measuring scalar constructs in social science with LLMs. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 32144–32171). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1635

[6] Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2023). *Spanish Pre-trained BERT Model and Evaluation Data* (arXiv:2308.02976). arXiv. https://doi.org/10.48550/arXiv.2308.02976

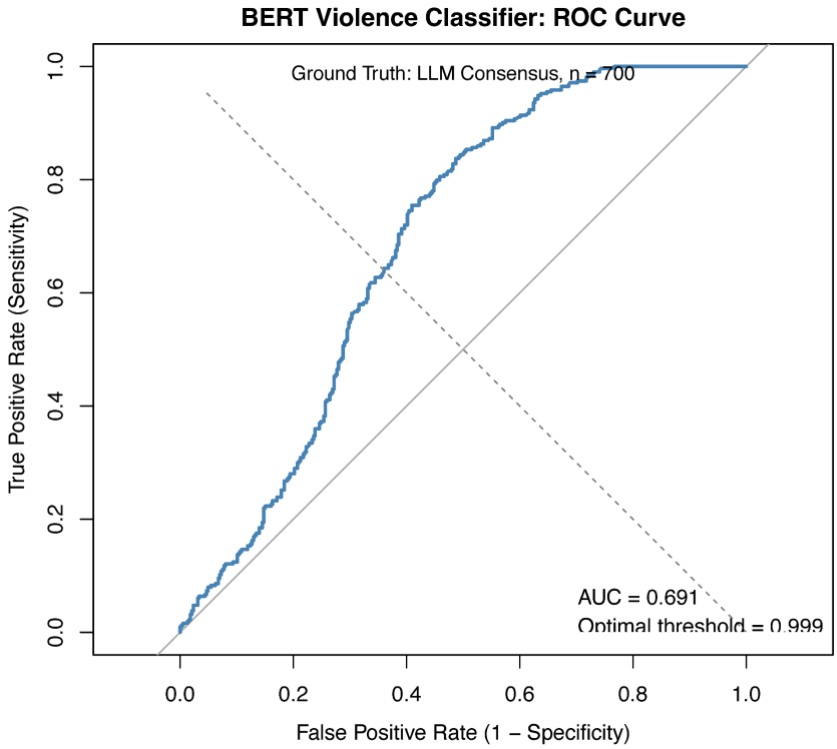


Figure A6. ROC curve for the BERT-based violence classifier. Ground truth labels derived from majority agreement across four LLMs (n = 700).

To determine whether BERT was detecting genuine violence missed by the dictionary or systematically overshooting, we randomly sampled 100 comments classified as violent by BERT but not by the dictionary. We submitted these to the same four LLMs used in our validation study. If BERT were capturing subtle semantic violence that the dictionary missed, we would expect high LLM agreement with BERT's classifications. Instead, LLMs flagged only 37–54% of these comments as violence-related (Google-Gemma: 53.5%, IBM-Granite: 37.0%, Meta-Llama: 48.5%, Qwen: 53.0%), with strong inter-rater agreement (ICC = 0.866, 95% CI: 0.801–0.910). This suggests that approximately half of BERT's additional positive classifications represent false positives rather than violence discourse undetected by the dictionary.

These findings indicate that for large-scale perception of violence measurement, the dictionary-based approach provides a better balance of precision, computational efficiency essential for real-time applications and cross-linguistic scalability. The BERT approach may be valuable for future work focused on fine-grained comment-type differentiation where computational constraints are less binding.

Additional Descriptive Statistics

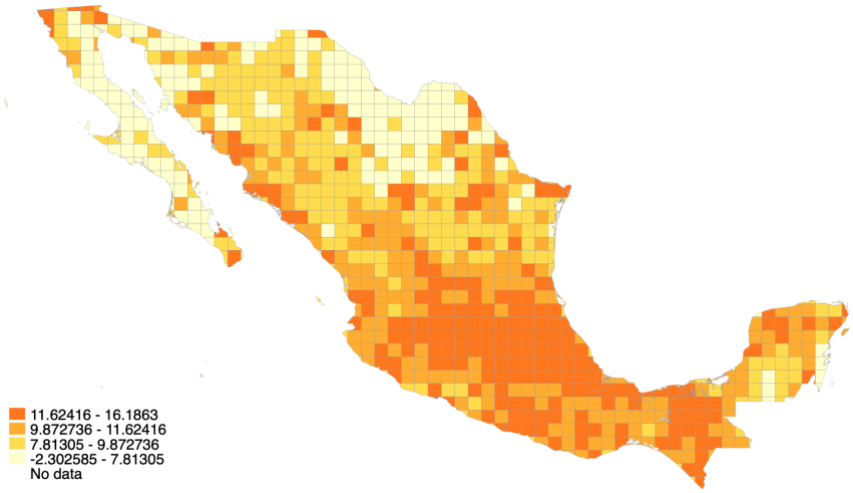


Figure A7. Geographic distribution of population at the PRIO-GRID cell level.

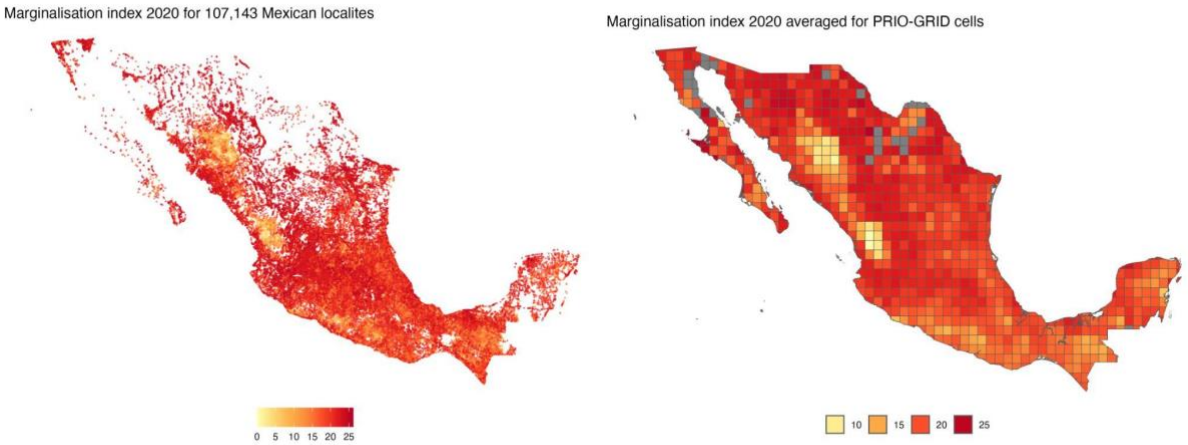


Figure A8. Marginalisation index measured in 2020 for 107,143 census localities and averaged at the PRIO-GRID cell level.

Table A11: Descriptive statistics

Variable	No.of Observations	Mean	Std. Deviation	Min	Max
$VPI^{w}{}_{i,y,m}$ (weighted by population)	45,580	0.0076	0.0330	0	1
$VPI_{i,y,m}$ (not weighted by population)	45,580	0.1927	0.0425	0	1
$ACLED\ Fatalities_{i,y,m}$ (weighted by population)	45,580	0.0096	0.1104	0	3.9736
$ACLED\ Fatalities_{i,y,m}$ (not weighted by population)	45,580	0.7679	3.4986	0	76
$ACLED\ Battle\ Fatalities_{i,y,m}$(weighted by population)	45,580	0.0011	0.0153	0	0.8199
$ACLED\ Battle\ Fatalities_{i,y,m}$ (not weighted by population)	45,580	0.1356	0.9247	0	35
$ACLED\ Explosions\ Fatalities_{i,y,m}$(weighted by population)	45,580	0.0001	0.0002	0	0.0334
$ACLED\ Explosions\ Fatalities_{i,y,m}$(not weighted by population)	45,580	0.0005	0.0383	0	6
$ACLED\ Protest\ Fatalities_{i,y,m}$(weighted by population)	45,580	0.0000	0.0000	0	0.0061
$ACLED\ Protest\ Fatalities_{i,y,m}$ (not weighted by population)	45,580	0.0001	0.0094	0	1
$ACLED\ Riot\ Fatalities_{i,y,m}$(weighted by population)	45,580	0.0001	0.0027	0	0.3296
$ACLED\ Riot\ Fatalities_{i,y,m}$ (not weighted by population)	45,580	0.0063	0.2204	0	39
$ACLED\ Civ.Violence\ Fatalities_{i,y,m}$(weighted by population)	45,580	0.0084	0.0982	0	3.8474
$ACLED\ Civ.Violence\ Fatalities_{i,y,m}$(not weighted by population)	45,580	0.6254	3.1021	0	76
$Homicides_{i,y,m}$ (weighted by population)	45,580	0.0339	0.3600	0	11.4600
$Homicides_{i,y,m}$	45,580	2.6105	10.6104	0	187

(not weighted by population)

Additional Estimates

Table A12: Variance Decomposition of VPI^w

Component	Mean	Std. Dev.	Variance	%of Total	Observations
Overall	0.0076	0.0330	0.00109	100%	N=45,580
Between-grid		0.0326	0.00106	97.6%	n=860
Within-grid		0.0054	0.000032.7%	2.7%	T=53

Between variance represents spatial variation across grid cells. Within variance represents temporal variation within grid cells over time. The decomposition shows that 97.6% of VPI variation is geographic (between grids) while only 2.7% reflects common temporal fluctuations (within grids over time), indicating VPI primarily captures localized violence perception rather than nation-wide discourse.

Table A13: Estimates controlling for population

	(1) $VPI_{i,y.m}$	(2) $VPI_{i,y.m}$	(3) $VPI_{i,y.m}$
Panel A: ACLED			
$Fatalities_{i,y,m}$	0.0008***	0.0008***	0.0008***
	(0.0001)	(0.0001)	(0.0001)
Observations	45,580	45,580	45,580
R-squared	0.0342	0.0560	0.0756
Panel B: Crime			
$Homicides_{i,y,m}$	0.0003***	0.0003**	0.0003**
	(0.0001)	(0.0001)	(0.0001)
Observations	45,580	45,580	45,580
R-squared	0.0372	0.0589	0.0785
Grid FE	No	No	No
Year FE	No	Yes	Yes
Month FE	No	No	Yes

The outcome variable is VPI . In Panel A, $Fatalities$ is the number of violence- related fatalities in grid i in month m of year y, as reported by ACLED. In Panel B, $Homicides$ is the number of homicides in grid i in month m of year y, as reported by the Mexican Government. Standard errors, clustered at the grid level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A14: Estimates with no population weighting

	(1) $VPI_{i,y.m}$	(2) $VPI_{i,y.m}$	(3) $VPI_{i,y.m}$	(4) $VPI_{i,y.m}$
Panel A: ACLED				
$Fatalities_{i,y,m}$	0.0013***	0.0004***	0.0004***	0.0004***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Observations	45,580	45,580	45,580	45,580
R-squared	0.0115	0.3063	0.3281	0.3478
Panel B: Crime				
$Homicides_{i,y,m}$	0.0005***	0.0003**	0.0002**	0.0002
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Observations	45,580	45,580	45,580	45,580
R-squared	0.0157	0.3063	0.3280	0.3477
Grid FE	No	Yes	Yes	Yes
Year FE	No	No	Yes	Yes
Month FE	No	No	No	Yes

The outcome variable is VPI. In Panel A, $Fatalities$ is the number of violence- related fatalities in grid i in month m of year y, as reported by ACLED. In Panel B, $Homicides$ is the number of homicides in grid i in month m of year y, as reported by the Mexican Government. Standard errors, clustered at the grid level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A15: Adding grid-level marginalization index as a control

	(1) $VPI^{w}_{i,y,m}$	(2) $VPI^{w}_{i,y,m}$	(3) $VPI^{w}_{i,y,m}$
Panel A: ACLED			
$Fatalities_{i,y,m}$	0.0049***	0.0049***	0.0049***
	(0.0000)	(0.0000)	(0.0000)
Observations	42,400	42,400	42,400
R-squared	0.2722	0.2722	0.2723
Panel B: Crime			
$Homicides_{i,y,m}$	0.0020***	0.0020***	0.0020***
	(0.0000)	(0.0000)	(0.0000)
Observations	42,400	42,400	42,400
R-squared	0.4041	0.4042	0.4042
Grid FE	No	No	No
Year FE	No	Yes	Yes
Month FE	No	No	Yes
Marginalization index	Yes	Yes	Yes

The outcome variable is VPI^{w}. In Panel A, $Fatalities$ is the number of violence- related fatalities in grid i in month m of year y, as reported by ACLED. In Panel B, $Homicides$ is the number of homicides in grid i in month m of year y, as reported by the Mexican Government. Standard errors, clustered at the grid level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A16: IHS transformed fatalities and homicides

	(1) $VPI^{w}_{i,y,m}$	(2) $VPI^{w}_{i,y,m}$	(3) $VPI^{w}_{i,y,m}$	(4) $VPI^{w}_{i,y,m}$
Panel A: ACLED				
$Fatalities_{i,y,m}$	0.3436***	0.0379***	0.0379***	0.0378***
	(0.0603)	(0.0106)	(0.0106)	(0.0106)
Observations	45,580	45,580	45,580	45,580
R-squared	0.6677	0.9737	0.9737	0.9738
Panel B: Crime				
$Homicides_{i,y,m}$	0.1810***	0.0307**	0.0306**	0.0304**
	(0.0152)	(0.0133)	(0.0133)	(0.0133)
Observations	45,580	45,580	45,580	45,580
R-squared	0.8512	0.9733	0.9733	0.9734
Grid FE	No	Yes	Yes	Yes
Year FE	No	No	Yes	Yes
Month FE	No	No	No	Yes

The outcome variable is VPI^{w}. In Panel A, $Fatalities$ is the IHS transformed number of violence- related fatalities in grid i in month m of year y, as reported by ACLED. In Panel B, $Homicides$ is the IHS transformed number of homicides in grid i in month m of year y, as reported by the Mexican Government. Standard errors, clustered at the grid level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A17: IHS transformed VPI as the outcome variable

(1)	(2)	(3)	(4)

	$VPI^{w}_{i,y,m}$	$VPI^{w}_{i,y,m}$	$VPI^{w}_{i,y,m}$	$VPI^{w}_{i,y,m}$
Panel A: ACLED				
$Fatalities_{i,y,m}$	0.0014***	0.0002***	0.0002***	0.0002***
	(0.0003)	(0.0000)	(0.0000)	(0.0000)
Observations	45,580	45,580	45,580	45,580
R-squared	0.6677	0.9737	0.9737	0.9738
Panel B: Crime				
$Homicides_{i,y,m}$	0.0008***	0.0001**	0.0001**	0.0001**
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Observations	45,580	45,580	45,580	45,580
R-squared	0.8512	0.9733	0.9733	0.9734
Grid FE	No	Yes	Yes	Yes
Year FE	No	No	Yes	Yes
Month FE	No	No	No	Yes

The outcome variable is $IHS - transformed\ VPI$. In Panel A, $Fatalities$ is the IHS transformed number of violence- related fatalities in grid i in month m of year y, as reported by ACLED. In Panel B, $Homicides$ is the IHS transformed number of homicides in grid i in month m of year y, as reported by the Mexican Government. Standard errors, clustered at the grid level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A18: Spatial spillovers of realized violence - All grids

	(1)	(2)	(3)

	$VPI^{w}_{i,y,m}$	$VPI^{w}_{i,y,m}$	$VPI^{w}_{i,y,m}$
Distance between grid i and j	100km	200km	300km
Panel A: ACLED			
$Fatalities_{j,y,m}$	1.2157** *	0.4947	0.4866
	(0.4660)	(0.3056)	(0.3378)
$Fatalities_{i,y,m}$	0.0226**	0.0240**	0.0233**
	(0.0110)	(0.0111)	(0.0100)
Observations	45,580	45,580	45,580
Panel B: Crime			
$Homicides_{j,y,m}$	0.7681	0.5905	0.5467
	(0.5483)	(0.4669)	(0.4425)
$Homicides_{i,y,m}$	0.0123**	0.0111**	0.0097**
	(0.0060)	(0.0050)	(0.0040)
Observations	45,580	45,580	45,580
Grid FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes

The outcome variable is VPI^{w}. In Panel A, $Fatalities$ is the number of violence- related fatalities, as reported by ACLED. In Panel B, $Homicides$ is the number of homicides, as reported by the Mexican Government. Conley (1999) clustered standard errors, accounting for spatial autocorrelation at the designated distance cutoff and temporal autocorrelation of 1 unit, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A19: Spatial spillovers of realized violence - Low population grids

	(1)	(2)	(3)

	$VPI^{w}_{i,y,m}$	$VPI^{w}_{i,y,m}$	$VPI^{w}_{i,y,m}$
Distance between grid i and j	100km	200km	300km
Panel A: ACLED			
$Fatalities_{j,y,m}$	0.9495***	0.5798**	0.3676
	(0.3421)	(0.2675)	(0.2898)
$Fatalities_{i,y,m}$	0.0238	0.0244	0.0246
	(0.0247)	(0.0239)	(0.0233)
Observations	22,790	22,790	22,790
Panel B: Crime			
$Homicides_{j,y,m}$	1.0256***	0.7746***	0.6346***
	(0.2400)	(0.1885)	(0.1774)
$Homicides_{i,y,m}$	0.0009	-0.0002	0.0000
	(0.0127)	(0.0125)	(0.0129)
Observations	22,790	22,790	22,790
Grid FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes

The outcome variable is VPI^{w}. In Panel A, $Fatalities$ is the number of violence- related fatalities, as reported by ACLED. In Panel B, $Homicides$ is the number of homicides, as reported by the Mexican Government. Conley (1999) clustered standard errors, accounting for spatial autocorrelation at the designated distance cutoff and temporal autocorrelation of 1 unit, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A20: Spatial spillovers of realized violence - High population grids

	(1) $VPI^{w}_{i,y,m}$	(2) $VPI^{w}_{i,y,m}$	(3) $VPI^{w}_{i,y,m}$
Distance between grid i and j	100km	200km	300km
Panel A: ACLED			
$Fatalities_{j,y,m}$	1.2129***	0.4905	0.4834
	(0.4655)	(0.3054)	(0.3385)
$Fatalities_{i,y,m}$	0.0226**	0.0241**	0.0233**
	(0.0110)	(0.0110)	(0.0100)
Observations	22,790	22,790	22,790
Panel B: Crime			
$Homicides_{j,y,m}$	0.7668	0.5897	0.5475
	(0.5479)	(0.4686)	(0.4464)
$Homicides_{i,y,m}$	0.0123**	0.0111**	0.0097**
	(0.0059)	(0.0050)	(0.0040)
Observations	22,790	22,790	22,790
Grid FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes

The outcome variable is VPI^{w}. In Panel A, $Fatalities$ is the number of violence- related fatalities, as reported by ACLED. In Panel B, $Homicides$ is the number of homicides, as reported by the Mexican Government. Conley (1999) clustered standard errors, accounting for spatial autocorrelation at the designated distance cutoff and temporal autocorrelation of 1 unit, in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A21: ACLED Fatalities in each event category as the predictor

	(1) $VPI^{w}_{i,y,m}$	(2) $VPI^{w}_{i,y,m}$	(3) $VPI^{w}_{i,y,m}$	(4) $VPI^{w}_{i,y,m}$	(5) $VPI^{w}_{i,y,m}$
$Fatalities_{i,y,m}$	0.0535**	-0.0918	-0.2115*	0.0001	0.0282***
	(0.0254)	(0.1371)	(0.1232)	(0.0445)	(0.0083)
Event category	Battles	Remote violence	Protests	Riots	Violence against Civilians
Observations	45,580	45,580	45,580	45,580	45,580
R -squared	0.9734	0.9731	0.9731	0.9731	0.9740
Grid FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes

The outcome variable is VPI^{w}. $Fatalities_{i,y,m}$ is the number of violence- related fatalities in grid i in month m of year y, for each event category, as reported by ACLED. Standard errors, clustered at the grid level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A22: Estimates based on UCDP fatalities

	(1)	(2)	(3)

	$VPI^{w}_{i,y,m}$	$VPI^{w}_{i,y,m}$	$VPI^{w}_{i,y,m}$
$Fatalities_{i,y,m}$	0.0000	0.0000	0.0001
	(0.0000)	(0.0000)	(0.0003)
Event category	Any	Non-state	One-sided
Observations	45,580	45,580	45,580
R-squared	0.9731	0.9731	0.9731
Grid FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes

The outcome variable is VPI^{w}. $Fatalities$ is the number of violence- related fatalities in grid i in month m of year y, as reported by UCDP. Standard errors, clustered at the grid level, in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Validation of Violence Perception Measures Against World Values Survey Data

As an additional validation exercise, we examine the relationship between survey-based measures of violence-related attitudes from the World Values Survey (WVS) Wave 7 and national homicide rates. This cross-national analysis demonstrates that attitudinal measures of violence perception correlate meaningfully with actual violence levels, providing further support for the construct validity of perception-based violence measures.

Data and Methods

World Values Survey Data

We utilize WVS Wave 7 (2017-2022), which includes two questions relevant to violence perceptions:

- **Q131**: Attitudes toward political violence, measured on a scale where higher values indicate greater tolerance for political violence
- **Q137**: Perceived security in one's neighborhood, measured on a 1-4 scale where 1 = "Very secure" and 4 = "Very insecure"

Negative response codes indicating missing data (don't know, refused, etc.) were recoded as missing prior to analysis.

Homicide Data

National homicide rates were obtained from the World Bank's World Development Indicators (indicator: VC.IHR.PSRC.P5), measured as intentional homicides per 100,000 population. Data covering 2010-2025 were retrieved, and each WVS country-year observation was matched to the nearest available homicide rate using a rolling join procedure.

Aggregation

For each country-year in WVS Wave 7, we calculated: 1. **Mean scores** for Q131 and Q137, representing average violence-related attitudes 2. **Standard deviations** for Q131 and Q137, representing within-country heterogeneity in attitudes

Sample

The merged dataset includes 62 country-year observations with complete data on all variables (mean and SD for both survey questions, and homicide rates).

Results

Correlation Analysis

Table A23 presents Pearson correlation coefficients between WVS attitude measures and national homicide rates.

Table A23: Correlation Between WVS Violence Attitude Measures and National Homicide Rates

Variable	Correlation (r)	95% CI	t-statistic	p-value	Significance
Mean Q131	0.344	[0.104, 0.547]	2.84	0.006	***
SD Q131	0.593	[0.403, 0.734]	5.70	< 0.001	***
Mean Q137	-0.385	[-0.579, -0.150]	-3.23	0.002	***
SD Q137	0.493	[0.277, 0.661]	4.39	< 0.001	***

Notes: N = 62 country-year observations. ** p < 0.01. Pearson product-moment correlations with World Bank homicide rates (per 100,000 population).*

Figures A9-A12 visualize these relationships with fitted linear regression lines.

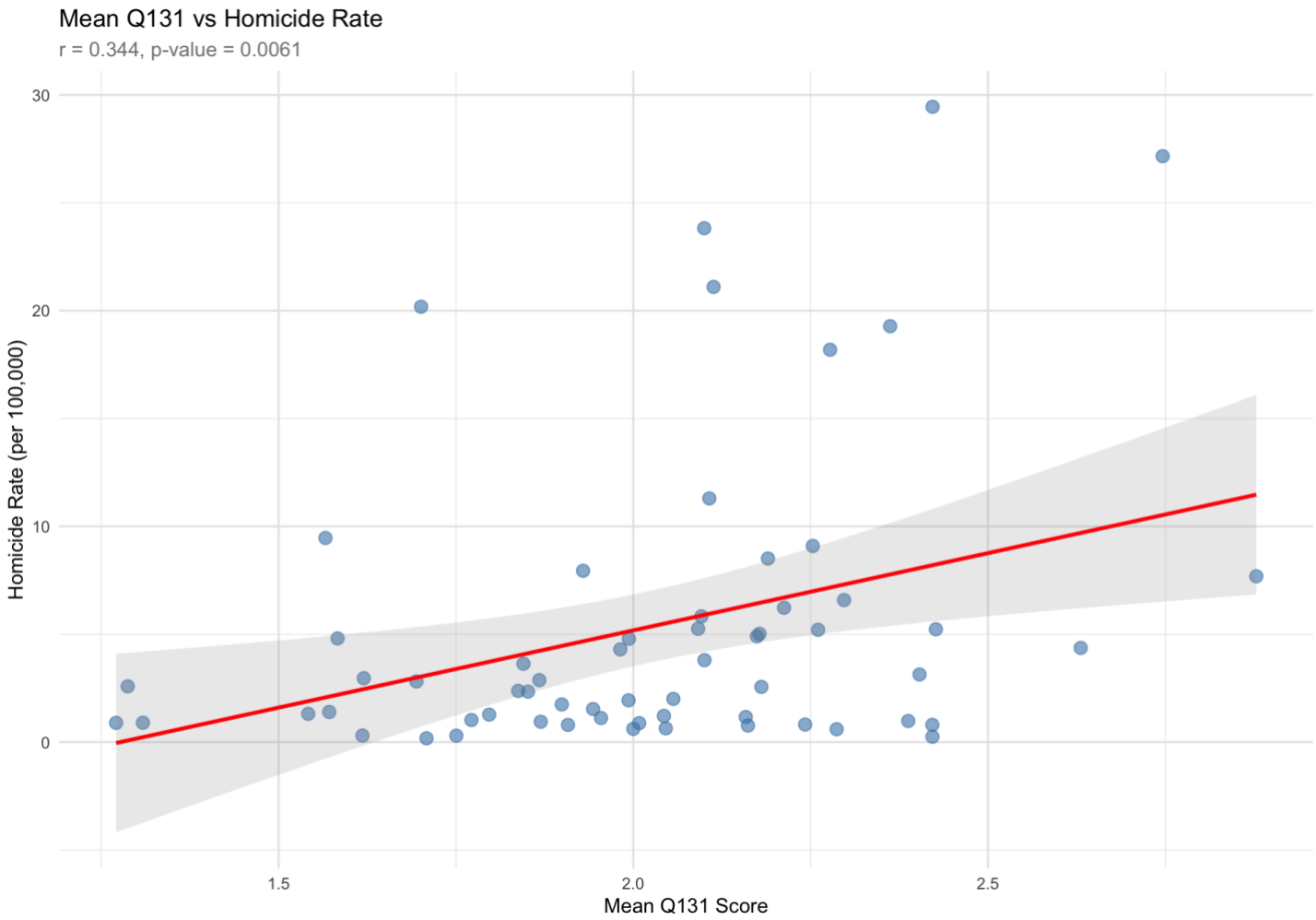


Figure A9: Mean Q131 (tolerance for political violence) vs. national homicide rate. Points represent country-year observations (N = 62). Blue line shows fitted linear regression with 95% confidence interval. r = 0.344, p = 0.006.

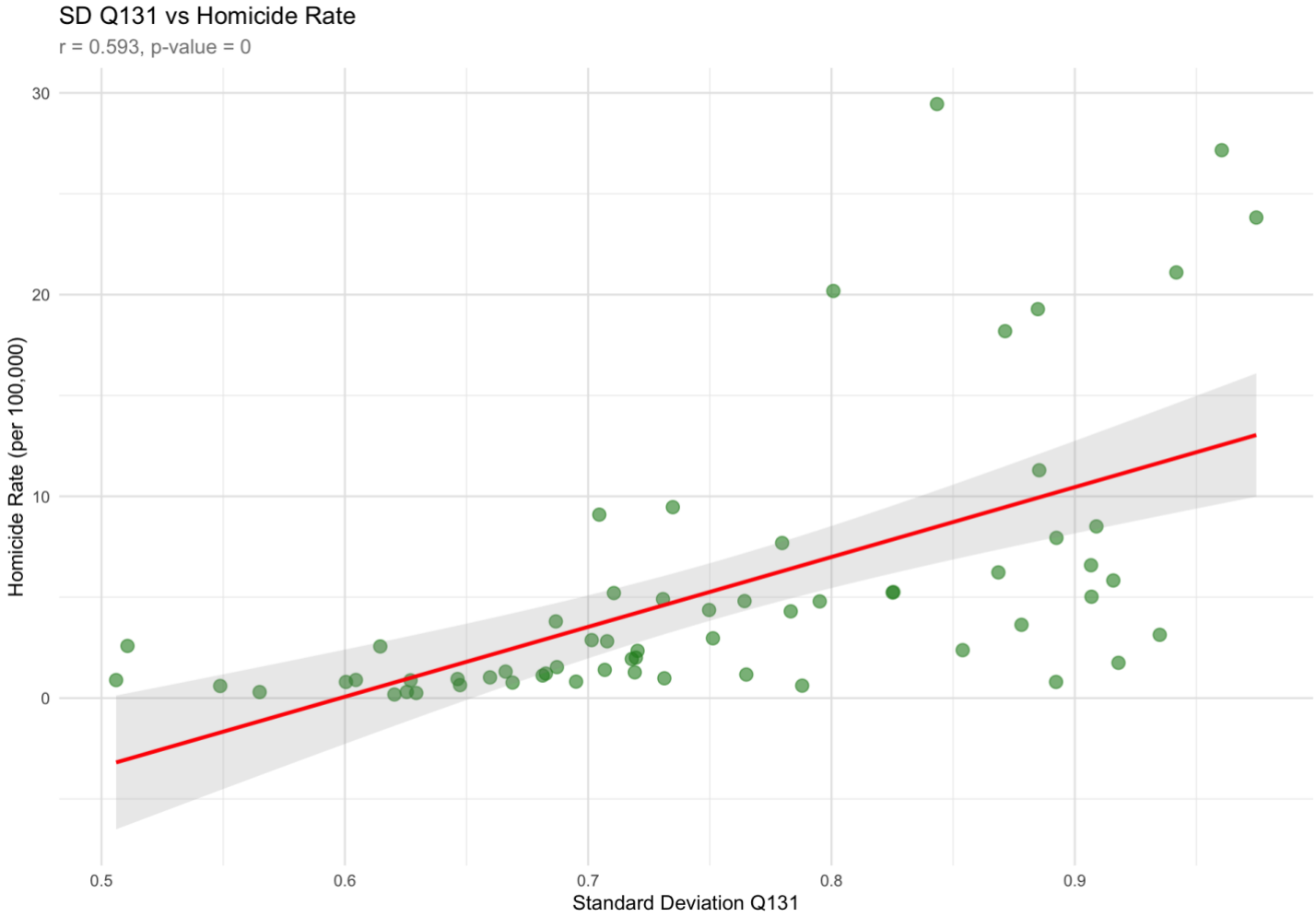


Figure A10: Standard deviation of Q131 vs. national homicide rate. Points represent country-year observations (N = 62). Blue line shows fitted linear regression with 95% confidence interval. r = 0.593, p < 0.001.

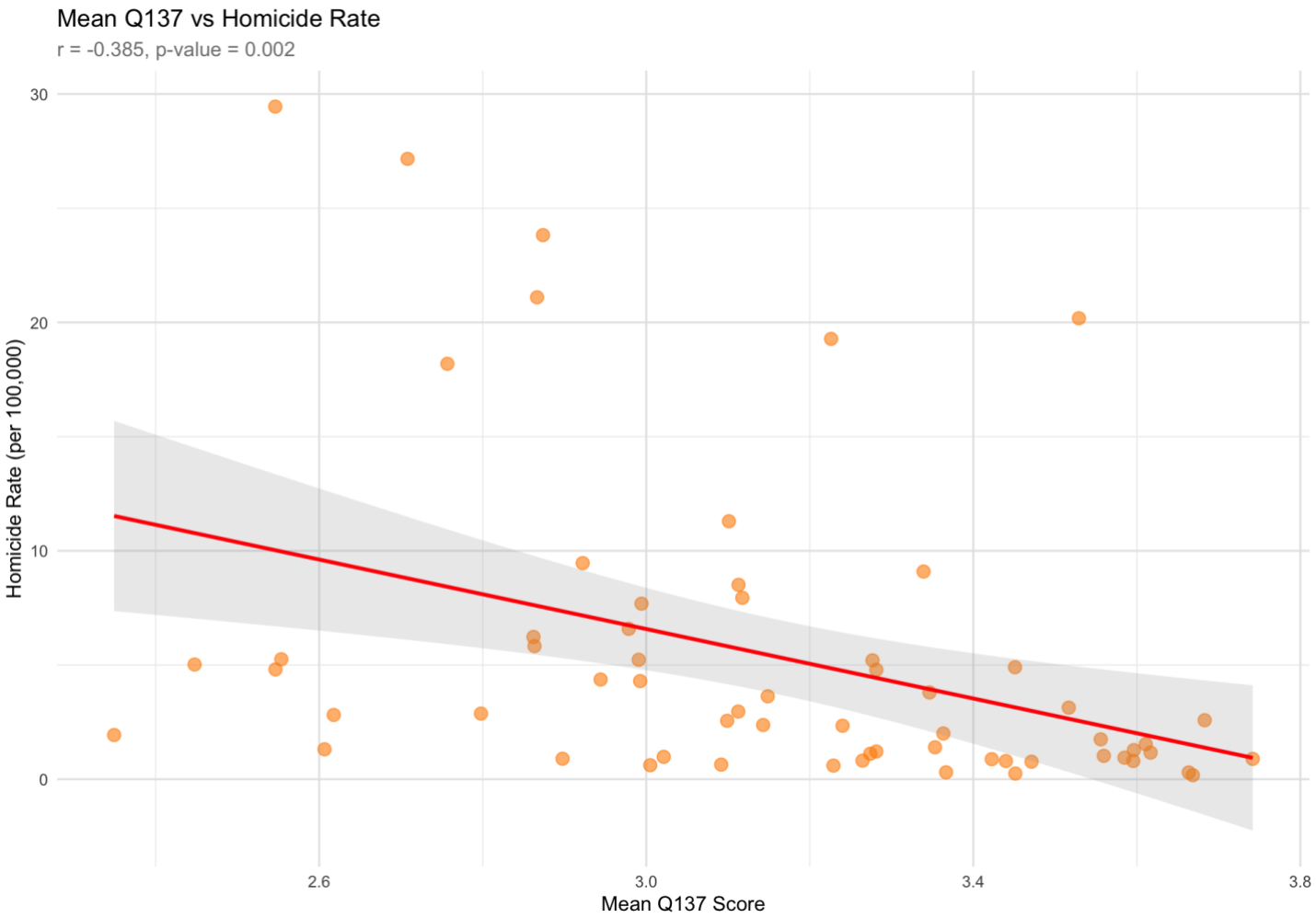


Figure A11: Mean Q137 (perceived neighborhood insecurity) vs. national homicide rate. Points represent country-year observations (N = 62). Blue line shows fitted linear regression with 95% confidence interval. r = -0.385, p = 0.002.

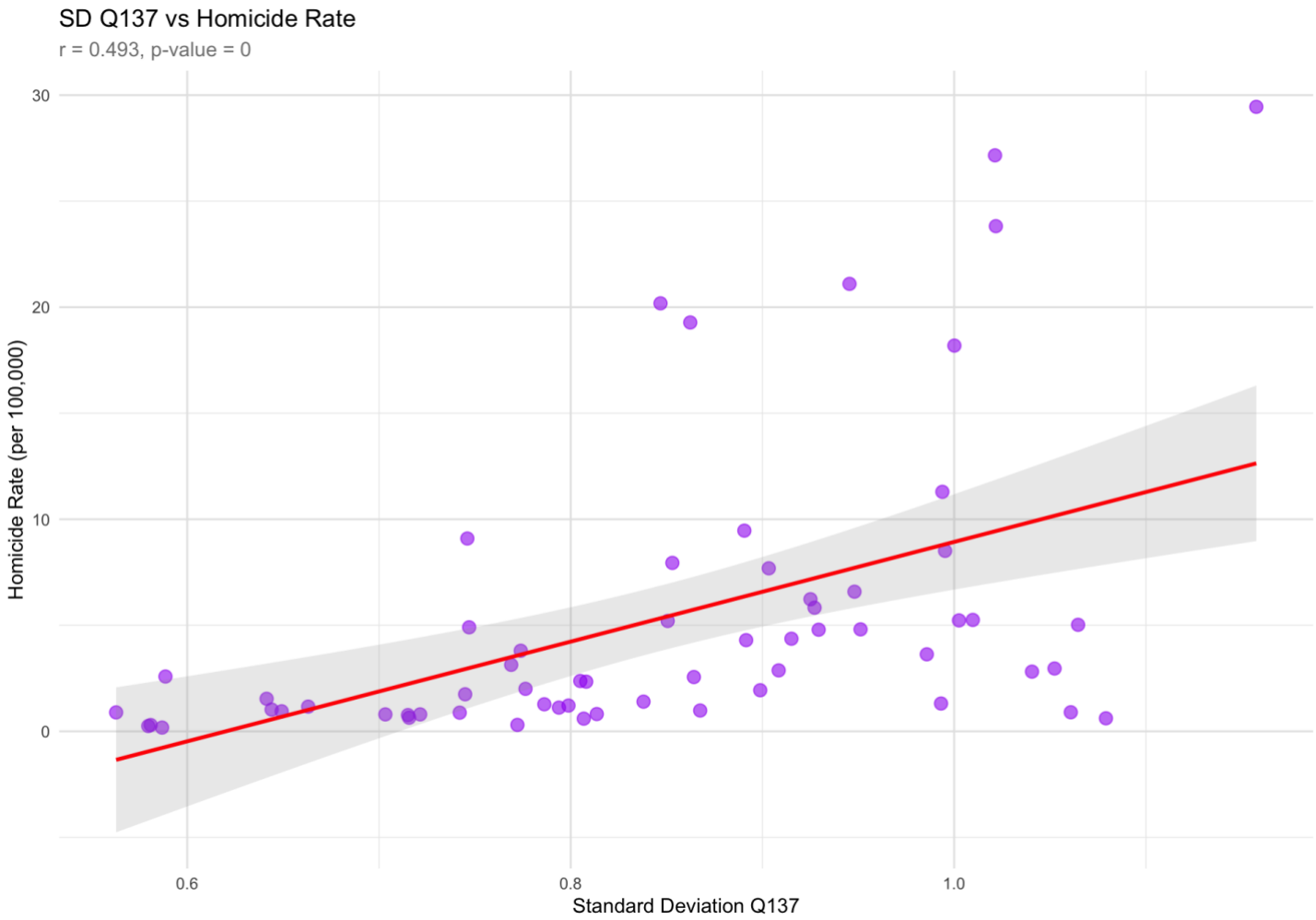


Figure A12: Standard deviation of Q137 vs. national homicide rate. Points represent country-year observations (N = 62). Blue line shows fitted linear regression with 95% confidence interval. r = 0.493, p < 0.001.

All four correlations are statistically significant at the $p < 0.01$ level, demonstrating robust associations between survey-based violence attitudes and actual violence levels across countries.

Temporal Alignment of VPI and ACLED

To examine the temporal relationship between VPI and ACLED fatalities, we constructed a normalized comparison that accounts for the different scales of the two indicators. For each measure, we calculated 30-day centered moving averages to smooth daily fluctuations while preserving meaningful temporal dynamics. We then expressed both series as fractions of their respective 2023 annual means, such that values above 100% indicate periods exceeding the 2023 baseline and values below 100% indicate periods falling short of it. This normalization enables direct visual comparison of relative trends across indicators measured in different units.

Figure A13 presents this comparison across three panels. The top panel displays the full study period (January 2020–May 2024), while the middle and bottom panels provide detailed views of two periods of particular interest: April–August 2020 and December 2023–May 2024.

This analysis reveals that VPI aligns with ACLED during specific episodes of violence escalation. In June–July 2020, both indicators rise sharply, corresponding to documented violence escalation in several Mexican states, including protests in Guadalajara following the death of Giovanni López in police custody and escalating cartel violence in Guanajuato that prompted a presidential visit (see main text, Figure 5). Similarly, from February–

May 2024, both indicators increase during Mexico's presidential election campaign period, consistent with the documented surge in political violence during this electoral cycle.

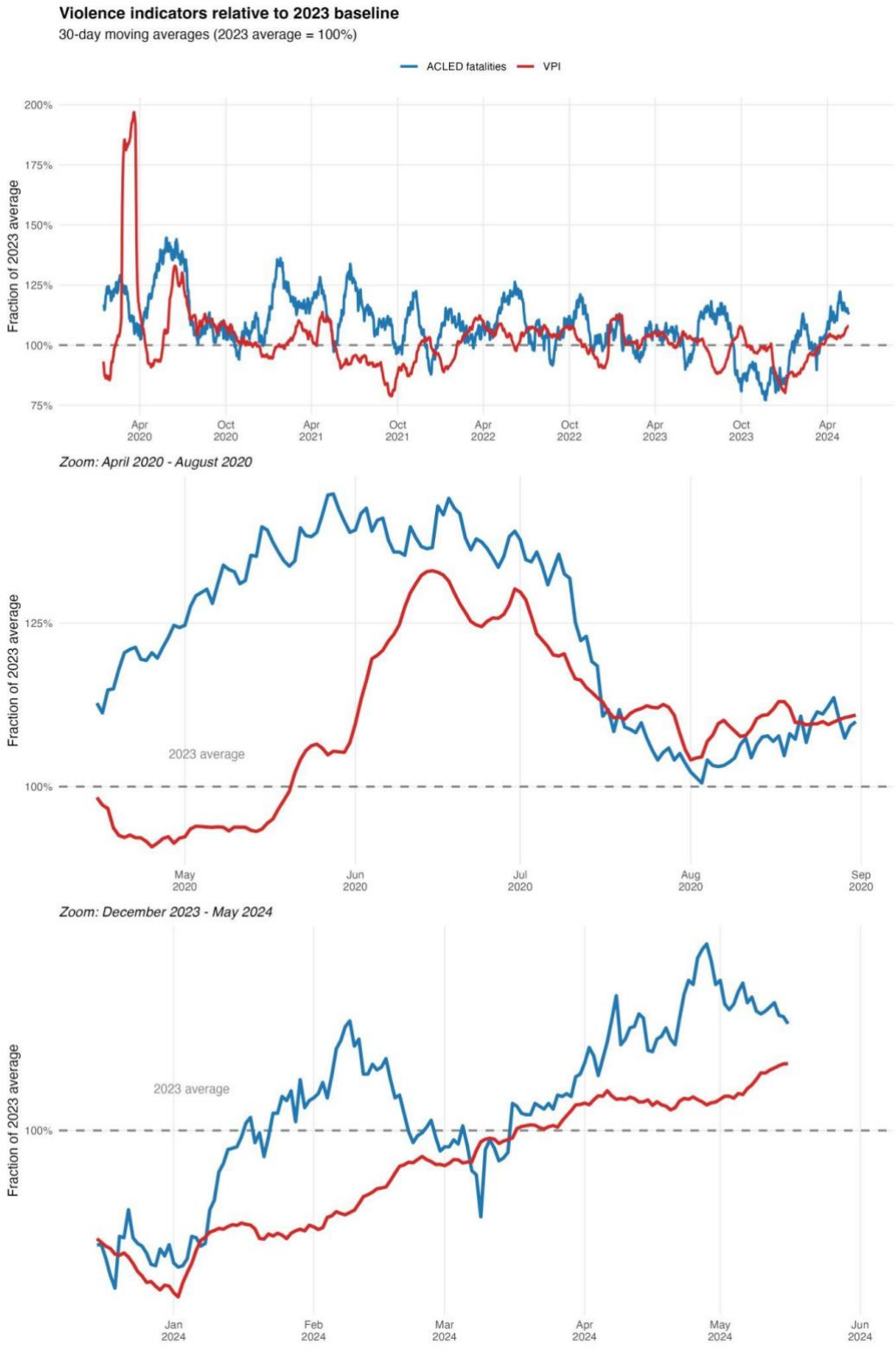


Figure A13: Temporal comparison of Violence Perception Index (VPI) and ACLED fatalities, January 2020–May 2024. Both indicators are expressed as 30-day centered moving averages normalized to their respective 2023 annual means (dashed line at 100%). Top panel: full study period. Middle panel: detailed view of April–August 2020, showing alignment during the June–July violence escalation. Bottom panel: detailed view of December 2023–May 2024, showing alignment during the presidential election campaign. The early 2020 VPI spike (top panel) reflects nationwide discourse on gender-based violence not captured by ACLED fatality counts.

The analysis also reveals an important divergence in early 2020, where VPI spikes dramatically (approaching 200% of the 2023 baseline) while ACLED fatalities remain relatively stable. This period corresponds to the nationwide protests and intense public discourse surrounding gender-based violence, culminating in the March 2020 International Women's Day demonstrations. This pattern illustrates the VPI's capacity to capture violence-related discourse and public concern that does not manifest in event-based fatality counts.

These findings clarify the appropriate use case for VPI. Rather than serving as a tool for long-term trend analysis, VPI is best suited for detecting localized perturbations, both temporal and geographic. Its value lies in identifying spikes in violence-related discourse that may signal emerging hotspots, shifting community concerns, or events generating public fear, whether or not these correspond to increases in documented fatalities. In this sense, VPI provides complementary intelligence for early warning systems and real-time monitoring, detecting signals that event-based datasets may miss or capture only with delay.