RESEARCH PAPER

Estimating the number of attacks to civilians in Ukraine: a quantitative analysis from media sources

Alessio Farcomeni* and Antonello Maruotti§
*Tor Vergata University of Rome, Rome, Italy; §Libera Università Maria Ss Assunta, Rome, Italy

ARTICLE HISTORY
Compiled December 01, 2023

*Email: alessio.farcomeni@uniroma2.it

ABSTRACT

We describe a compiled database of daily reports of attacks involving civilians that are associated with the Russian aggression to Ukraine. We argue how obtaining reliable estimates would be beneficial for conflict resolution and post-war. Using appropriate statistical methods we correct for undercount and report unbiased estimates, with their associated confidence intervals. We discuss the territorial and time heterogeneity of civilian incidents. We argue that media sources, combined with appropriate methodologies, can be a timely and appropriate method for estimating the extent of harm in war times.

KEYWORDS
Civilian incidents; citizen science; data journalism; war

1. Introduction

The conflict in Ukraine is a complex and multifaceted event, giving rise to a multitude of narratives and perspectives from the various stakeholders involved. These narratives often differ significantly depending on the political, historical, and cultural contexts of the individuals or groups presenting them (Watanabe 2017; Lazarenko 2019; Götz and Staun 2022). The multiplicity of narratives surrounding the conflict underscores the deep divisions and competing interests at play, making it challenging to discern a single, definitive account.

Media have always played a significant role in shaping and amplifying the different narratives (Zeitzoff 2017; Makhortykh and Sydorova 2017; Zhabotynska and Velivchenko 2019). Both sides have utilized various media outlets, including state-owned or controlled channels, to disseminate their perspectives and influence public opinion (Roman, Wanta, and Buniak 2017; Lichtenstein and Koerth 2022; Zhabotynska and Ryzhova 2022; Wagnsson and Lundström 2022; Tolz and Hutchings 2023). The spread of misinformation, selective reporting, and biased interpretations have further fueled the multiplicity of narratives, making it increasingly difficult for individuals to form an unbiased understanding of the conflict. From a pure qualitative perspective, navigating the multiplicity of narratives surrounding the conflict in Ukraine requires critical thinking, careful analysis of different sources, and consideration of various perspectives. Recognizing the complexities, biases, and competing interests involved is essential in developing a comprehensive understanding of the conflict and working towards a peaceful resolution.

Quantitative analyses can be of paramount importance in order to gain a comprehensive and well-rounded understanding of the situation. By examining news articles, reports, and interviews from various media sources, both local and international, researchers and analysts can gain insights into the diverse viewpoints and perceptions of the conflict (Fernández et al. 2023; Lupu and Wallace 2022). Analyzing media data allows for the identification of potential misinformation and enables researchers to distinguish between reliable information and intentionally misleading narratives, helping to counter disinformation and enhance the accuracy of public discourse (Fengler et al. 2020). Moreover, media data analysis also enables the identification of trends and patterns in the reporting of the conflict over time. By examining the frequency of news coverage, researchers can identify shifts in the evolution of the conflict (Ptaszek, Yuskiv, and Khomych 2023). This time series analysis contributes to a deeper understanding of the evolving dynamics of the conflict and its implications. Moreover, such a type of analysis can help bridge gaps in information and provide valuable insights into underreported aspects of the conflict. It allows researchers to uncover events that may not be prominently featured in the media. By including a broader range of sources, such as independent journalists, citizen journalists, and social media content, we can gain a more nuanced understanding of the conflict and its impact on different communities.

We make a significant contribution to the existing literature by conducting a quantitative analysis of the number of attacks involving civilians in Ukraine. We use data from the media (mainly, Twitter recounts). A comprehensive list of sources reporting includes local and international outlets, government reports, NGOs, and citizen journalists.

Our main aim is estimating the *true* number of incidents that have potentially or actually caused harm to civilians. The necessary data for the task have been conveniently collected by Bellingcat (more details below), and is freely available on the website https://ukraine.bellingcat.com. Bellingcat has also done the necessary work to confirm the actual occurrence of each event reported. The source provides comprehensive reports of incidents in Ukraine involving civilian harm, targeting civilian infrastructure, or instances where rockets or missiles have clearly struck civilian areas. The database does not provide information on the number of dead or injured individuals for each event. The website is regularly updated, with new incidents being added almost daily, starting from February 24th, 2022. The data collection efforts are intended to continue until the conclusion of the conflict.

It is important here to highlight the website's own disclaimer, which states: ”while we are attempting to collect as many incidents as possible, we cannot possibly guarantee to collect them all. [...] Therefore, this map is not an exhaustive list of civilian harm in Ukraine but rather a representation of all incidents which we have been able to collect and of which we have been able to determine the exact locations.” This challenge is further acknowledged by The Office of the UN High Commissioner for Human Rights (OHCHR), which reports a total of 23,375 civilian casualties in Ukraine as of May 1st, 2023, with 8,709 fatalities and 14,666 injuries. However, it is important to note that the OHCHR also acknowledges that the actual figures may be ”considerably higher.” We firmly assert that the data provided by the Bellingcat platform holds significant value and can be effectively adjusted for potential undercount through appropriate statistical methods, which we describe and use below. We provide therefore reliable estimates, together with an assessment of uncertainty (by means of confidence intervals), which we believe can be useful to objectively and timely inform the political and public debate about the ongoing war in Ukraine.

A strategy similar to what we propose can actually be used more in general: undercounting is a common issue in conflict monitoring, as it is inherently challenging to gather comprehensive information in such circumstances. The use of media content for correcting for undercount has also been discussed in other contexts, e.g., Farcomeni (2022).

In the following we use population size estimation methods, a.k.a. capture-recapture methods (Böhning, Bunge, and van der Heijden 2018; McCrea and Morgan 2014), to obtain reliable estimates of the true extent of civilian harm in the Ukraine war, how it is evolving, and its spatial distribution. In conflict situations, attacks may go unreported or be inadequately documented due to factors such as limited access, lack of resources, or fear of reprisals. Since media outlets often report on different incidents, capturing a wide range of information, analyzing the overlaps between various media sources can provide insights into the total number of attacks that have occurred. Capture-recapture methods can help address these gaps by estimating the *hidden* attacks that are not captured by any individual source; and are inferred through overlaps between sources (Murphy 2009; Parada et al. 2023). In our data for each incident we have information such as the date, location, nature of the attack, and type of location (e.g., school, private building, etc.). Each reported attack can be considered a ”capture” event from the data analysis perspective. An overlap occurs when two or more sources report on the same attack. By analyzing the extent of these overlaps, it becomes possible to estimate the number of unreported or *hidden* attacks. In detail, we provide a straightforward procedure to estimate the magnitude of the underreporting by considering the bias-corrected version of Chao estimator (Chao 1987; Chao and Colwell 2017). Chao's lower bound estimator is designed to be robust against biases and reporting variations; it does not rely on specific assumptions about the distribution or structure of the data, making it suitable for situations where reporting sources may differ in terms of coverage, accuracy, or reporting probabilities. This robustness ensures that the estimator provides a conservative estimate that accounts for potential discrepancies and biases in the observed data, which likely arise in conflict situations. Chao's lower bound estimator, thus, offers a useful approach for communicating the potential range of the number of attacks, helping to inform decision-makers, policymakers, and humanitarian organizations about the minimum scale of the conflict and assist in resource allocation and planning.

The rest of the paper is as follows: in the next section we provide a literature review and connect our contribution to the extant literature on peace research. In Section 3 we briefly describe the Bellingcat data and outline the proposed methodology for correcting undercount. In Section 4 we show our estimates for the number of incidents involving civilians. In Section 5 we give some concluding remarks about our findings.

2. Underreporting in Conflict Data: Insights from Recent Research and Motivation

Our research can be easily cast into the recent literature on the importance of addressing underreporting in conflict data. Understanding and analyzing conflict data is pivotal for comprehensive assessment of conflicts worldwide. It plays a crucial role to plan support for civilian recovery, to provide reliable legal accountability (i.e. quantifying the war crimes properly in front of courts and tribunals), to influence policy decisions for peace initiatives and, obviously, to fight misinformation and avoid a biased public opinion. Reliable estimates of civilian attacks, war crimes, casualties might be useful for conflict resolution, as they might facilitate peace negotiations. The availability of underestimated figures might delay the onset of ceasefires, causing additional harm to populations involved.

In general, underreporting of critical events can significantly impact the accuracy and reliability of conflict datasets. Several recent studies, aiming at quantifying the underreporting rate, have delved into this issue, shedding light on the complexities surrounding underreporting and its implications for conflict analysis. In the following, we summarize recent works sharing the same philosophy as ours, i.e. increasing the attention on the bias that underreporting may lead to if not properly addressed.

Cook et al. (2017) addressed the inherent underreporting present in binary data collected from various sources. Their work emphasized the need for robust methodologies to correct biases arising from underreported conflict data focusing on measurement error and misclassification. By proposing models to account for underreporting, a strategy to remedy potential bias is proposed by further remarking that researchers possessing more than one source of data-generating information can achieve this desired result.

Weidmann (2016) conducted a comprehensive analysis of reporting bias in conflict event datasets, introducing a diagnostic measures. The research highlighted the necessity of acknowledging and understanding the impact and uncertainty of reporting bias when utilizing conflict event data. By recognizing reporting biases, researchers can refine methodologies to improve the accuracy of conflict analyses.

In a related vein, Demarest and Langer (2022) focused on the inclusion and exclusion of events in datasets derived from newspapers. They outlined the pitfalls and offered guidelines for researchers leveraging newspapers as sources in conflict studies. Understanding these pitfalls can assist in navigating potential biases arising from the selection of data sources, as the quality of conflict event data can be affected by a wide range of errors.

Raleigh, Kishi, & Linke (2023) explored, from a pure qualitative perspective, the complexities associated with political instability patterns, emphasizing how such patterns might be obscured due to scope conditions, coding choices, and data sources used in constructing conflict datasets. Their findings underscored the impact of dataset construction on the comprehension of political instability trends within conflict zones.

Another aspect investigated by Dorff, Henry, & Ley (2023) focused on the potential deterrence of detailed reporting due to violence against journalists in conflict zones. This study revealed how intimidation tactics might hinder accurate reporting on conflict events, highlighting the multifaceted challenges in collecting reliable data.

Parkinson (2023) delved into the political economy of media-sourced conflict data, illuminating the factors influencing the reporting of conflict events by the media. By addressing the challenges and biases inherent in conflict data reported by media outlets, the study highlighted the complexities of using media-sourced data for conflict analysis.

Finally, Miller et al. (2022) proposed an agenda for addressing biases present in conflict data. Their work aimed to improve the accuracy and reliability of such datasets by offering insights into mitigating biases and enhancing methodologies for collecting and analyzing conflict-related information.

Collectively, these studies discuss the multifaceted challenges associated with underreporting in conflict data. They emphasize the critical need for refined methodologies, greater awareness of biases, and improved data collection practices to enhance the accuracy and reliability of conflict datasets, thus enabling more informed analyses and policy decisions in conflict resolution and peace building efforts.

We would like to further emphasize the importance of accurate data for post-conflict reconstruction, humanitarian aid, and healing processes for affected communities. An unbiased evaluation allows policy makers to compare programme achievements to planned objectives and activities. As discussed in "International Prosecutors" edited by Luc Reydams, Jan Wouters, and Cedric Ryngaert, the role of evidence collection, methodologies for analyzing conflict-related data, and the significance of accurate information are crucial in international legal proceedings. While this reference does not directly address the statistical methods employed in conflict analysis, it could provide broader insights into the importance of robust data and methodologies in supporting legal actions related to conflicts, which indirectly underscores the significance of accurate data analysis, including capture-recapture methods, in such contexts.

Moreover, our research delves into the complexities of information dissemination during conflicts. In the context of the Ukraine-Russia war, capture-recapture methods offer insights into the difficulties faced by peacekeeping missions in gathering and disseminating accurate data due to the nature of the conflict; challenges such as misinformation, biased reporting, or limitations in accessing comprehensive information are often encountered in conflict zones. These challenges reflect the complexities and difficulties faced in accurately capturing and analyzing data related to conflicts, challenges which methodologies, such as capture-recapture data analysis, can straightforwardly address.

Boulding (2018), Collier (2003), and Galtung (1969) are specifically devoted to conflict studies and peace research, addressing various aspects of conflicts and their resolution. The estimation of the true number of attacks in Ukraine aligns with the core themes discussed in these landmark papers. Just to briefly summarize, estimating the true number of attacks in Ukraine contributes by offering empirical insights into the nature and scale of conflict, aiding in understanding the dynamics of violence and the impact on society. Similarly, it provides critical quantitative data that sheds light on the severity and frequency of violence in the region. At last, by revealing potential undercount or inaccuracies in reported attacks, it unveils the need for more accurate data in peace research, which can inform the development of effective peace-building strategies.

Such insights are essential in understanding the impact of conflict on development and formulating effective policies to break the cycle of violence. This aligns with Boulding's focus on comprehending conflict behaviors and their implications for defense strategies, with Collier’s analysis of the intricate relationship between conflict, civil wars, and development and with Galtung's work that delves into the concepts of violence and peace, advocating for peace-oriented research and strategies.

3. Materials and Methods

Bellingcat is an esteemed and independent collective comprising researchers, investigators, and citizen journalists. Since the onset of the aggression against Ukraine, the collective has been diligently collecting information concerning incidents involving civilians. To ensure the reliability, authenticity, and accuracy of the data, rigorous checks are conducted to verify the sources, identify potential manipulation, and ascertain the precise location of each incident. The data, encompassing a list of one or more sources, incident location, classification of the affected area, weapon systems employed in the attacks, and a concise description, has been generously made available at the URL `https://ukraine.bellingcat.com/`.

In the preparation of our study, we established a data freeze on May 31, 2023. All the findings presented can be readily updated to reflect the latest data release, as we have provided the accompanying `R` code as supplementary material.

Between February 24th, 2022 and the freeze date, the website documented a total of 1,046 attacks, with approximately half of them (569) attributed to a single source. The incident with the highest number of sources (23) pertains to an attack that occurred at the Retroville shopping mall in Kyiv on March 20, 2022. The city of Kharkiv emerged as the most frequently affected area in the available data, with 161 attacks. This was followed by Mariupol (60) and Mykolaiv (57). Among the 941 incidents with detailed information on the type of area affected, residential areas accounted for 437 incidents, an alarming 123 attacks targeted school or childcare areas, and 58 attacks were directed towards healthcare facilities. Additionally, there were 131 and 86 reported attacks in commercial and industrial areas, respectively. A summary representation of the data at hand by (the most relevant) locations and types is given in Figure 1.

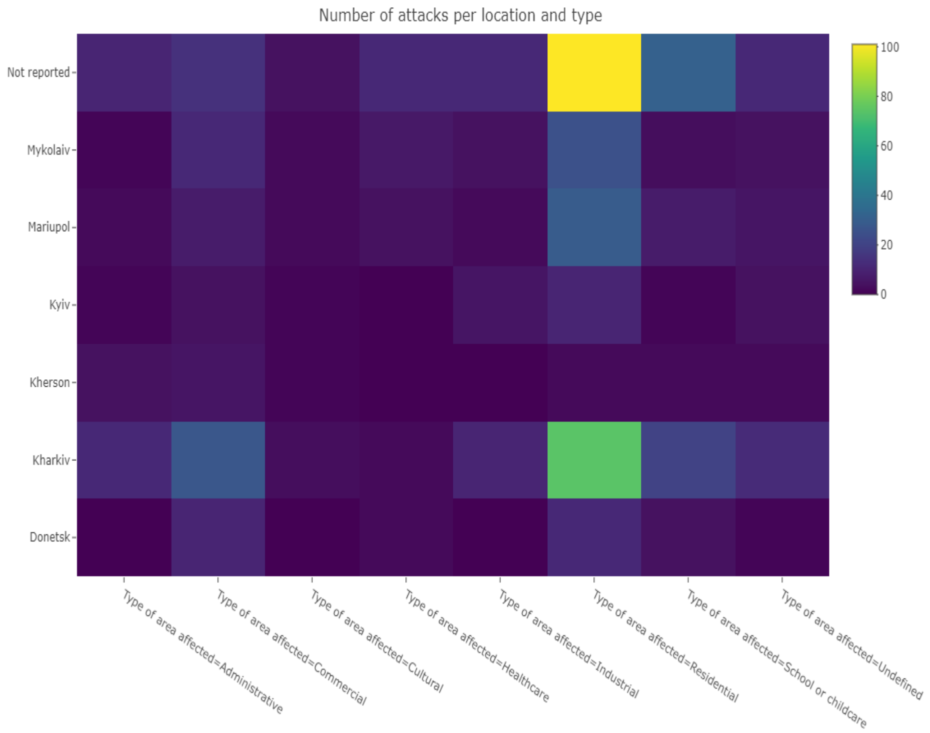


Figure 1.: Number of attacks per location and type

It is crucial to note that all the aforementioned numbers are based on observed data, primarily relying on social media reports and recounts of events that have been verified by Bellingcat. As a result, it is expected that the true figures are higher than what has been reported. Importantly, the extent of undercounting may vary across different types of areas. For instance, attacks in healthcare areas might be marginally more likely to be reported due to heightened indignation and outrage, compared to incidents occurring in administrative areas. Consequently, making reliable comparisons based on percentages becomes challenging. Our primary objective in this study is to address the undercounting issue and provide accurate estimates for the number of attacks involving civilians. The key to achieving this lies in leveraging the fact that Bellingcat meticulously reports all available sources for each event.

To estimate the number of attacks that go unreported by any detected source, we employ the bias-corrected version of Chao’s estimator (Chao 1987; Bohning, Kaskasamkul, and van der Heijden 2019; Chao and Colwell 2017; Farcomeni and Dotto 2021). This estimator guarantees that the estimates serve as a lower bound for the true unobserved number of events and performs well even with low counts. Technical details are provided in the Appendix. In summary, the main ingredients of the bias-corrected Chao’s estimator are briefly outlined. The collected data represent the number of media sources reporting an attack, suggesting the use of count data distributions (Winkelmann 2008). While the Poisson model may serve as a natural starting point, it imposes restrictive assumptions by assuming a unit variance-to-mean ratio. Thus, the Poisson distribution may not be suitable for analyzing the current data, which exhibit overdispersion or underdispersion primarily due to unobserved heterogeneity. To account for heterogeneity in estimating the population size, the Poisson parameter is often treated as an unobserved random variable with a latent distribution (Chao 1987). A non-parametric bias-corrected estimator of the population size is obtained as follows:

$$\widehat{N}=n+f_1(f_1-1)/[2(f_2+1)]$$

Here, f_1 represents the number of attacks recorded exactly by one source (singletons), f_2 denotes the number recorded by two sources (doubletons), and n is the observed number of attacks. The underlying idea of this non-parametric estimator is that f_1 and f_2 accurately reflect the number of missed individuals and the number of individuals captured but not yet recaptured, respectively.

By employing the bias-corrected Chao’s estimator, we address the challenge of estimating the unobserved number of attacks while taking into account potential biases introduced by incomplete reporting. This methodology provides a valuable approach to obtain reliable estimates of the population size in capture-recapture studies with count data.

4. Results

Using our approach we can estimate that the number of attacks involving civilians in the period between February 2022 and May 2023 is 1761 (95% confidence interval: 1601-1921). The extent of undercount in Bellingcat data is therefore at least (1601/1046-1)≈50%. The confidence interval is moderately narrow, indicating not very much uncertainty about the large number of incidents involving civilians estimated.

In Table 1 we report the observed and estimated number of attacks by type of area affected. Confidence intervals indicate moderate uncertainty. They sometimes include the observed counts, indicating that some undercount shall be actually expected for all type of areas with the exclusion of cultural, administrative, and healthcare areas. As a general comment, the accuracy of information on the number of attacks can vary depending on the type of targets involved, and as such some are more likely to be reported than others. To summarize, the following aspects could motivate such differences, making healthcare facilities, cultural and administrative buildings of major importance in monitoring the number of attacks:

- Visibility and documentation
- International scrutiny
- Targeted nature of attacks
- Media priorities

Attacks on healthcare facilities, cultural and administrative buildings often receive more attention from various stakeholders, including international organizations, human rights groups and, accordingly, media outlets. These types of targets are considered critical infrastructure or institutions, and their destruction or damage tends to have significant repercussions. As a result, there is often more visibility and documentation of attacks on such facilities, making the data more reliable. Similarly, they typically have well-established reporting mechanisms in place, increasing the probability attacks on them to be reported and documented accurately. Furthermore, there is a different perception in the public opinion driving media attention for different types of targets. Attacks on healthcare facilities, cultural and administrative buildings generally lead to international outrage and condemnation. Consequently, these incidents often trigger investigations and assessments by international bodies, such as the United Nations or independent commissions; again increasing the reliability of the reported data. In other words, there might be a media reporting bias, corresponding to different levels of priorities: media outlets may prioritize certain types of attacks over others. Attacks on critical infrastructure and institutions often receive more attention due to their perceived impact on society and international norms. This can result in a reporting bias where attacks on residential, commercial, and educational targets are relatively under-reported, leading to potential inaccuracies in the data. The targeted nature of attacks, thus, also play a role. Attacks on residential areas, commercial establishments, and schools may involve a broader range of targets and can be more scattered across different locations, making more challenging to track and verify attacks on individual homes, small businesses, or scattered schools compared to concentrated attacks on specific infrastructure or institutions.

Table 1.: Observed and estimated attacks on civilians between February 24th, 2022 and May 31st, 2023; by type of area affected

Area	Observed	Estimate	95% confidence interval	

			Lower bound	Upper bound
Residential	437	740	633	846
Commercial	131	237	169	304
School or childcare	123	211	155	267
Undefined	90	141	100	182
Industrial	86	123	91	155
Healthcare	58	99	58	142
Administrative	42	68	42	100
Cultural	26	31	26	41

In Table 2 we report the observed and estimated number of attacks by city. Only the most frequently involved areas have been included in the table. Despite being by far the most frequently reported city, the estimated number of attacks in Kharkiv are almost twice the reported ones. For all other frequently attacked cities we do not have strong evidence of undercount, due to large standard errors. Nevertheless, Chao bias-corrected lower-bound estimates are frequently about twice the observed counts.

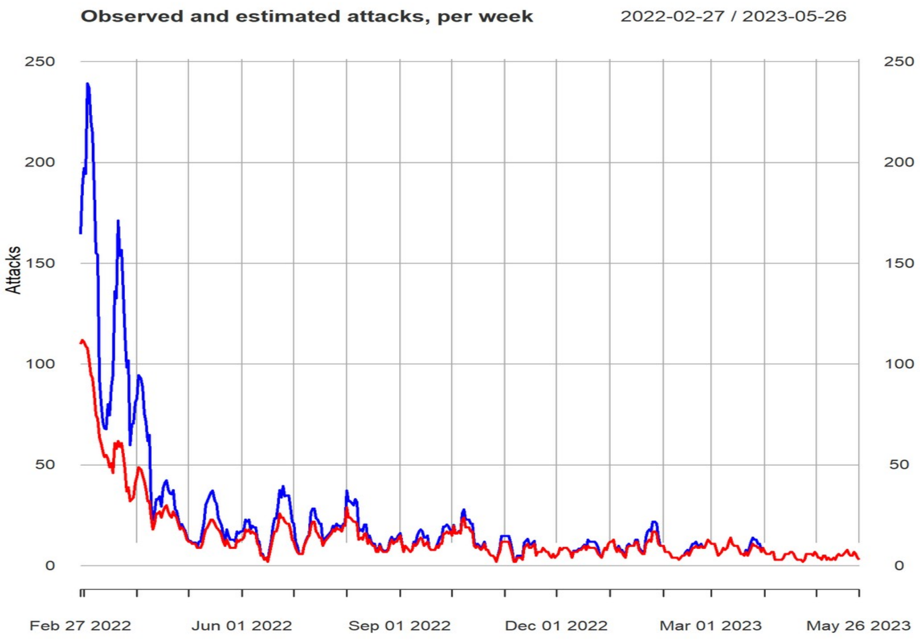
Two main factors could contribute to explain the relevant number of attacks in Kharkiv, namely, the geographical location and the strategic significance. Kharkiv is located in northeastern Ukraine, relatively close to the border with Russia and the separatist-controlled territories in Donetsk and Luhansk. Its proximity to the conflict zone makes it potentially vulnerable to cross-border shelling, infiltration, or other forms of attacks. Consequently, being an important economic and industrial hub, its strategic significance make it an attractive target for destabilization efforts or attempts to disrupt Ukraine's functioning and can be thought as a way to weaken morale, exert control, or intimidate population. These characteristics made Kharkiv a major target of the Russian attacks, and several factors can contribute to the challenges in accurately reporting and documenting attacks, from security concerns to disrupted infrastructures and self-censorship. Its geographical location increases the lack of security near the conflict zone, with higher risks for journalists, independent observers, and humanitarian organizations attempting to gather information on attacks. These risks may include threats from armed groups, cross-border shelling, or the presence of landmines and unexploded ordnance. As a result, access to certain areas or specific information might be restricted, hindering the ability to report on attacks comprehensively. Similarly, being a strategic target, the attacks led to damage to infrastructure, including communication networks, power supply, and transportation systems. Such disruptions can hamper the collection and transmission of information, making it more challenging for journalists and observers to report on attacks accurately. At last, we would remark that in regions at the core of the conflict zone, there may be a climate of fear and self-censorship, where individuals are hesitant to share information due to concerns about reprisals or personal safety, leading to undercounting or underreporting.

Table 2.: Observed and estimated attacks on civilians between February 24th, 2022 and May 31st, 2023; by area. Only the most frequent areas are reported in the table.

Area	Observed	Estimate	95% confidence interval	
			Lower bound	Upper bound
Kharkiv	161	316	227	404
Mariupol	60	159	60	260
Mykolaiv	57	115	57	187
Donetsk	32	74	32	147
Kyiv	29	34	29	45
Kherson	19	30	19	59

At last, we conduct a time-dependent analysis, focusing on weekly data, in order to appraise the evolution over time.

Figure 2.: Time series of estimated number of attacks on civilians per week between February 24th, 2022 and May 31st, 2023. Observed attacks in red, total estimated attacks in blue



From Figure 2 it can be clearly seen that after an initial “shock and awe” period with very large number of attacks per week, the phenomenon has become slightly stable. A cyclical seasonality can be observed, with regular and almost equidistant periods of high frequency followed by shorter periods of low frequency of events. The total number of attacks was strongly underreported in the early week of the conflict, whilst nowadays it is minor or almost null. Accordingly, we can conclude that the level of underreporting of the number of attacks in Ukraine indeed varies over time, and even if it is challenging to provide a definitive explanation, we can argue that: in the initial stages of the conflict, there was chaos and confusion, making it difficult to gather accurate and timely information; as the conflict progresses, it attracts increased international attention, leading to heightened scrutiny and reporting on attacks; the conflict dynamic evolved over time, with shifting territorial control, changes in tactics, and varying levels of intensity; and of course organizations involved in monitoring and documenting conflict-related incidents refined their methodologies, strengthen their networks, and establish better data collection and verification processes.

5. Discussion

In this work we have presented estimates, also stratified by week, type of area, and partly by city, for the number of attacks involving civilians the ongoing Ukranian war. As mentioned in the introduction, Chao bias-corrected estimates are simple to use and can be readily used to correct for undercount any time a list of sources is reported for each event of interest. Unfortunately we were not able to locate viable data about the number of victims of the conflict. The official UN estimates available are aggregated and no sources are reported. An indirect estimate for the number of victims could be obtained through statistics about the average number of deaths and/or injuries per event, which is also not reported by the UN or other sources that we could locate. Attacks on civilians seem to have hit quite hard the city of Kharkiv, possibly due to early involvement in the conflict; and has not spared healthcare and childcare areas, including schools.

It is widely acknowledged that many official counts are simply lower bounds for the true ones in the context of conflict data. We have argued about the importance of reliable corrected estimates for peace building and post-war. We conclude by remarking that the very occurrence of undercount can be often simply tested (Farcomeni, 2023).

References

Boulding, K. E. 2018. *Conflict and defense: A general theory*. Pickle Partners Publishing.

Böhning, D., P. Kaskasamkul, and P. G. M. van der Heijden. 2019. “A modification of Chao’s lower bound estimator in the case of one-inflation.” *Metrika* 82: 361–384.

Böhning, Dankmar. 2010. “Some general comparative points on Chao’s and Zelterman’s estimators of the population size.” *Scandinavian Journal of Statistics* 37 (2): 221–236.

Böhning, Dankmar, John Bunge, and Peter GM van der Heijden. 2018. *Capture-recapture methods for the social and medical sciences*. CRC Press Boca Raton.

Chao, A. 1987. “Estimating the population size for capture-recapture data with unequal catchability.” *Biometrics* 43: 783–791.

Chao, A., and R. K. Colwell. 2017. “Thirty years of progeny from Chao’s inequality: estimating and comparing richness with incidence data and incomplete sampling.” *SORT* 41: 3–54.

Collier, P. 2003. *Breaking the conflict trap: Civil war and development policy*. World Bank Publications.

Cook, S. J., Blas, B., Carroll, R. J., & Sinha, S. (2017). “Two wrongs make a right: Addressing underreporting in binary data from multiple sources”. *Political Analysis*, 25(2), 223-240.

Demarest, L., & Langer, A. (2022). “How events enter (or not) data sets: the pitfalls and guidelines of using newspapers in the study of conflict”. *Sociological Methods & Research*, 51(2), 632-666.

Dorff, C., Henry, C., & Ley, S. (2023). “Does violence against journalists deter detailed reporting? Evidence from Mexico”. *Journal of Conflict Resolution*, 67(6), 1218-1247.

Farcomeni, A. 2022. “How many refugees and migrants died trying to reach Europe? Joint population size and total estimation.” *Annals of Applied Statistics* 16: 2339–2351.

Farcomeni, A. 2023. “A likelihood ratio test for completed sampling in population size estimation studies.” *Biometrical Journal*, 65: 2200129.

Farcomeni, A., and F. Dotto. 2021. “A correction to make Chao estimator conservative when the number of sampling occasions is finite.” *Statistics and Probability Letters* 176: 109154.

Fengler, Susanne, Marcus Kreutler, Matilda Alku, Bojana Barlovac, Mariella Bastian, Svetlana S Bodrunova, Janis Brinkmann, et al. 2020. “The Ukraine conflict and the European media: A comparative study of newspapers in 13 European countries.” *Journalism* 21 (3): 399–422.

Fernández, Óscar, Marie Vandendriessche, Angel Saz-Carranza, Núria Agell, and Javier Franco. 2023. “The impact of Russia’s 2022 invasion of Ukraine on public perceptions of EU security and defence integration: a big data analysis.” *Journal of European Integration* 45 (3): 463–485.

Galtung, J. (1969). “Violence, peace, and peace research”. *Journal of peace research*, 6(3), 167-191.

Götz, Elias, and Jørgen Staun. 2022. “Why Russia attacked Ukraine: Strategic culture and radicalized narratives.” *Contemporary Security Policy* 43 (3): 482–497.

Lazarenko, Valeria. 2019. “Conflict in Ukraine: multiplicity of narratives about the war and displacement.” *European Politics and Society* 20 (5): 550–566.

Lichtenstein, Dennis, and Katharina Koerth. 2022. “Different shows, different stories: How German TV formats challenged the government’s framing of the Ukraine crisis.” *Media, War & Conflict* 15 (2): 125–145.

Lupu, Yonatan, and Geoffrey PR Wallace. 2022. “Targeting and Public Opinion: An Experimental Analysis in Ukraine.” *Journal of Conflict Resolution* 00220027221121139.

Makhortykh, Mykola, and Maryna Sydorova. 2017. “Social media and visual framing of the conflict in Eastern Ukraine.” *Media, war & conflict* 10 (3): 359–381.

McCrea, Rachel S, and Byron JT Morgan. 2014. *Analysis of capture-recapture data*. CRC Press.

Miller, E., Kishi, R., Raleigh, C., & Dowd, C. (2022). “An agenda for addressing bias in conflict data”. *Scientific Data*, 9(1), 593.

Murphy, Joe. 2009. “Estimating the World Trade Center tower population on September 11, 2001: a capture–recapture approach.” *American Journal of Public Health* 99 (1): 65–67.

Parada, Vanessa, Larissa Fast, Carolyn Briody, Christina Wille, and Rudi Coninx. 2023. “Underestimating attacks: comparing two sources of publicly-available data about attacks on health care in 2017.” *Conflict and Health* 17 (1): 3.

Parkinson, S. E. (2023). “Unreported Realities: The Political Economy of Media-Sourced Data”. *American Political Science Review*, 1-6.

Ptaszek, Grzegorz, Bohdan Yuskiv, and Sergii Khomych. 2023. “War on frames: Text mining of conflict in Russian and Ukrainian news agency coverage on Telegram during the Russian invasion of Ukraine in 2022.” *Media, War & Conflict* 17506352231166327.

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), 1-17.

Reydams, L., Wouters, J., & Ryngaert, C. (Eds.). (2012). *International prosecutors*. Oxford University Press.

Roman, Nataliya, Wayne Wanta, and Iuliia Buniak. 2017. “Information wars: Eastern Ukraine military conflict coverage in the Russian, Ukrainian and US newscasts.” *International Communication Gazette* 79 (4): 357–378.

Tolz, Vera, and Stephen Hutchings. 2023. “Truth with a Z: disinformation, war in Ukraine, and Russia’s contradictory discourse of imperial identity.” *Post-Soviet Affairs* 1–19.

Wagnsson, Charlotte, and Magnus Lundström. 2022. “Ringing true? The persuasiveness of Russian strategic narratives.” *Media, War & Conflict* 17506352221101273.

Watanabe, Kohei. 2017. “The spread of the Kremlin’s narratives by a western news agency during the Ukraine crisis.” *The Journal of International Communication* 23 (1): 138–158.

Weidmann, N. B. (2016). “A closer look at reporting bias in conflict event data”. *American Journal of Political Science*, 60(1), 206-218.

Winkelmann, Rainer. 2008. *Econometric analysis of count data*. Springer Science & Business Media.

Zeitzoff, Thomas. 2017. “How social media is changing conflict.” *Journal of Conflict Resolution* 61 (9): 1970–1991.

Zhabotynska, Svitlana, and Olha Ryzhova. 2022. “Ukraine and the West in pro-Russia Chinese media: A methodology for the analysis of multimodal political narratives.” *Cognition, Communication, Discourse* (24): 115–139.

Zhabotynska, Svitlana, and Valentina Velivchenko. 2019. “New media and strategic narratives: the Dutch referendum on Ukraine–EU Association Agreement in Ukrainian and Russian Internet blogs.” *European security* 28 (3): 360–381.

Appendix

Formally, let X_i, $i = 1,\ldots,N$ denote the number of sources reporting attack i, and let $p_x = \Pr(X_i = x)$. Also let f_x denote the frequency of attacks reported exactly x times, $x = 0,1,\ldots,m$. As $X_i = 0$ is not observed, the corresponding f_0 is unknown and should be estimated. As X_i takes only non-negative integer values, the Poisson model with a single homogeneous parameter λ may represent a natural starting point. As mentioned in the main text, this model is restrictive because it assumes a unit variance-to-mean ratio. Hence, to account for heterogeneity due to over/under-dispersion, we consider the following marginal distribution which entails a mixture of Poisson distributions

$$p_x(\lambda)=\int_0^{\infty}\frac{\exp(-t)t^x}{x!}\lambda(t)\,dt$$

where the mixing distribution density $\lambda(t)$ is unknown. Under this model, the conventional estimator of Chao $n+f_1^2/[2(f_2)]$ is asymptotically unbiased for N, but experiences overestimation bias for small sample sizes as we have in the considered empirical setting. For small population sizes, a bias-correction should be used. The reason for the bias-adjustment is as follows. Ideally, we would like f_1^2/f_2 to be close to $E(f_1)^2/E(f_2)$. However, the estimator f_1^2/f_2 estimates $E(f_1^2/f_2)$, and $E(f_1)^2/E(f_2)$ and $E(f_1^2/f_2)$ are not necessarily close. As it turns out, an excellent bias-corrected estimator is provided by

$$\widehat{N} = n + f_1(f_1 - 1)/[2(f_2 + 1)]$$

details are given in Böhning (2010).