Organized Criminal Violence Event Data (OCVED)

Methodological Appendix

Javier Osorio

School of Government and Public Policy
University of Arizona

Version 2.0

April, 2022

Contents

1. Introduction

This note presents the technical elements and methodological procedures used to develop Organized Criminal Violence Event Data (OCVED) 2.0 in Osorio and Beltrán (2020). At its core, this research project integrates innovations in Machine Learning (ML) protocols and Natural Language Processing (NLP) tools in Spanish to generate a geo-referenced database at the daily municipal level on the violent presence of Organized Criminal Groups between 2000-2018.

To generate this database, the methodological strategy takes advantage of a large collection of articles from local and national newspapers and press releases from a variety of government agencies. To ensure the validity of the input corpora used in this study, the methodological approach relies on Machine Learning algorithms to classify relevant news stories. After selecting the relevant corpora, the research protocol uses Evetus ID Osorio and Reyes (2017), a rule-based event coding software designed to identify event data from text written in Spanish. In this particular application, Eventus ID is used exclusively to extract the names of organized criminal groups in Mexico and to pinpoint their location as referenced in the original text.

The resulting database presents geo-referenced data at unprecedented levels of granularity by comprising daily-municipal information about 10 main criminal organizations that can be further disaggregated into more than 200 criminal cells. The application also includes an interactive web interface presenting dynamic heat-maps of criminal territories. This information is the empirical foundation to identify distinct temporal and spatial trends in the development and contestation dynamics between criminal organizations during this time period.

Scholars and practitioners in the security, law enforcement, and development sectors require fine-grained and timely information to track and understand the dynamics of violence in highly complex and rapidly changing scenarios of intense conflict (Schrodt and Van Brackel 2013; Chojnacki et al. 2012). Unfortunately, such need is often hampered

by the slow pace, limited scope, and high cost of manually generated data (Schrodt et al. 2010). This problem is even more acute in developing countries thorn by violence where the availability of resources for research an analysis is scarce (Seybolt, Aronson and Fischhoff 2013; Davenport and Ball 2002; Zhukov, Davenport and Kostyuk 2019).

This study is part of a broader research agenda focused on two main areas of research. One branch focuses on developing methodological tools for event coding using natural language processing and machine learning based on a collaboration with political scientists and computer scientists based at the University of Texas - Dallas. This team has its foundations in the Open Event Data Alliance (OEDA) (`https://github.com/openeventdata`) and features prominent researchers such as Patric Bradt, Vito D'Orazio, and Latifur Khan, and a variety of graduate students. This branch has produced valuable methodological contributions to event coding that is available at `https://github.com/eventdata`.

The second branch of this research agenda focuses on applying different technological and methodological developments to analyze the behavior of armed actors in specific contexts. OCVED is the application of this research methodology to track the violent presence of criminal organizations in Mexico. Following a similar methodology, the Violent Presence of Armed Actors in Colombia (ViPAA) (Osorio et al. 2019) tracks the territorial presence of a variety of armed groups including government forces, paramilitary groups, insurgent organizations, and criminal organizations. ViPAA is available at `https://www.colombiaarmedactors.org/`.

OCVED contributes to other research efforts focused on tracking the presence of organized criminal groups in Mexico using computational social sciences approaches. In particular, OCVED provides a variety of contributions that help to advance the databases developed by Rios (2012) and Signoret et al. (2021). First, OCVED covers a broader time horizon than other projects by tracking the presence of armed actors from 2000 to 2018. Second, OCVED provides fine-grained information at the municipality daily

level, thus offering highly granular data across time and space. Third, OCVED tracks a larger number of criminal actors, either grouped into large criminal organizations, or by disaggregating them into a multitude of subgroups and spin-offs. Fourth, OCVED is transparent in reporting unidentified criminal organizations reported in the news. In this way, the project acknowledges the limitation, ambiguities, and difficulties of relying on news articles that do not always report with precision the specific organized criminal groups active in a given locale. In this way, OCVED offers an unprecedented level of detail in the efforts of tracking the violent presence of Organized Criminal Groups in Mexico.

2. Funding

The Organized Criminal Violence Event Data project has been possible thanks to the generous support of a variety of funding sources that contributed to different stages of the project. These sources include:

- The University of Arizona, Research, Discovery & Innovation, Technology and Research Initiative Fund (TRIF), 2018-2022.
- Cornell University, Mario Einaudi Center for International Studies Postdoctoral Fellowship, 2013-2014.
- Yale University, Pre-Doctoral Fellowship, Program on Order, Conflict and Violence, 2012-2013.
- Harry Frank Guggenheim Foundation, Dissertation Fellowship, Harry Frank Guggenheim Foundation Dissertation Fellowships, 2012-2013.
- The Kellogg Institute for International Studies - University of Notre Dame, Dissertation Year Fellowship, 2012-2013.

- National Science Foundation, Doctoral Dissertation Research Improvement Grant, award 1123572, 2011.
- Social Science Research Council – Open Society Foundations,Drugs, Security and Democracy Fellowship, 2011.
- United States Institute of Peace, Jennings Randolph Peace Scholarship Dissertation Program, 2011.
- The Kellogg Institute for International Studies - University of Notre Dame, Graduate Research Grant, 2011.

Dr. Javier Osorio is deeply thankful for all the support and trust received by these generous donors. Also, Dr. Osorio is thankful for all the research assistants that made this project possible at its different stages.

3. Violent Presence of Organized Criminal Groups

In line with the analysis of armed actors that the team of researchers conducted in Colombia (Osorio et al. 2019), this study uses the term *Armed Actor* to refer to state and non-state armed actors that exercise the organized use of violence in a specific territory to achieve political or economic goals. As indicated in Section 10, this study focuses on nine main organizations. These main actors are further disaggregated in a variety of specific branches, subgroups, or spin-offs.

The output data derived from this coding efforts provides indication of the *Violent Presence* of armed actors at the municipality-day level. Given the nature of the information source, this data reflects the location of armed actors involved in violent incidents which can be lethal (e.g. assassinations, armed clashes) or non-lethal (e.g. threats, displacement, beatings, kidnappings etc.). However, it does not reflect cases in which armed actors are present in a territory but exercise no violence (Arjona 2011). Such cases

would correspond to areas of dominant or monopolistic control of armed actors in which the use of violence might not be necessary (Kalyvas 2006; Reuter 2009). However, the information source might not report such incidents, thus inhibiting the software form detecting non-violent presence of armed groups. In consequence, inferences drawn from this data should consider a narrow approach focused on the violent presence of armed actors.

4. Database Content

Table 1 presents the list of variables included in the OCVED v2.0 database and provides a brief description of their content and variable type.

Table 1: OCVED v2.0 content

Variable name	Definition	Variable type
`date`	Date concatenated	Numeric
`date_elapsed`	Date elapsed	Date (mm/dd/yyyy)
`year`	Year	Numeric
`month`	Month	Numeric
`day`	Day	Numeric
`state`	State	Numeric
`mun`	Municipality	Numeric
`counter`	Number of records per municipality-day	Numeric
`actor_main`	Main Organized Criminal Group	Character
`actor_sub`	Sub group or spin-off	Character
`longitude`	Longitude	Numeric
`latitude`	Latitude	Numeric

The file `OCVED_v2.0.xlsx` is a spreadsheet containing geo-referenced data on the violent presence of Organized Criminal Groups in Mexico. The database includes:

- Geographic coverage: 1,475 municipalities (the entire country).
- Temporal coverage: daily data from 1/1/2000 to 12/31/2018.
- Main Orgarnized Criminal Groups coded:

  - Beltran Leyva
  - Cartel de Jalisco Nueva Generacion
  - Cartel de Juarez
  - Cartel de Sinaloa
  - Cartel de Tijuana
  - Cartel del Golfo
  - La Barbie
  - La Familia Michoacana
  - Los Zetas
  - Huachicoleros (oil thieves)
  - Unidentified criminal group
  - Other criminal groups

- Total number of observations coded: 64,895.

The first nine actors correspond to large criminal organizations comprising a multitude of internal branches, factions, or spin-offs. The three residual groups (Huachicoleros, Unidentified Criminal Group, and Other criminal groups) cluster a variety of small and fragmented groups that do not operate under a unified command.

Figure 1 presents the distribution of frequencies for each of the main types of actors included in the database. As displayed in the graph, about half of the events detected refer to unidentified criminal groups. In these cases, news articles tend to make generic references to "organized crime" or "drug cartels" without explicitly mentioning a specific criminal organization. The organized criminal groups that are most mentioned in the news articles during the period of analysis are the Sinaloa Cartel, La Familia Michoacana, and Los Zetas.

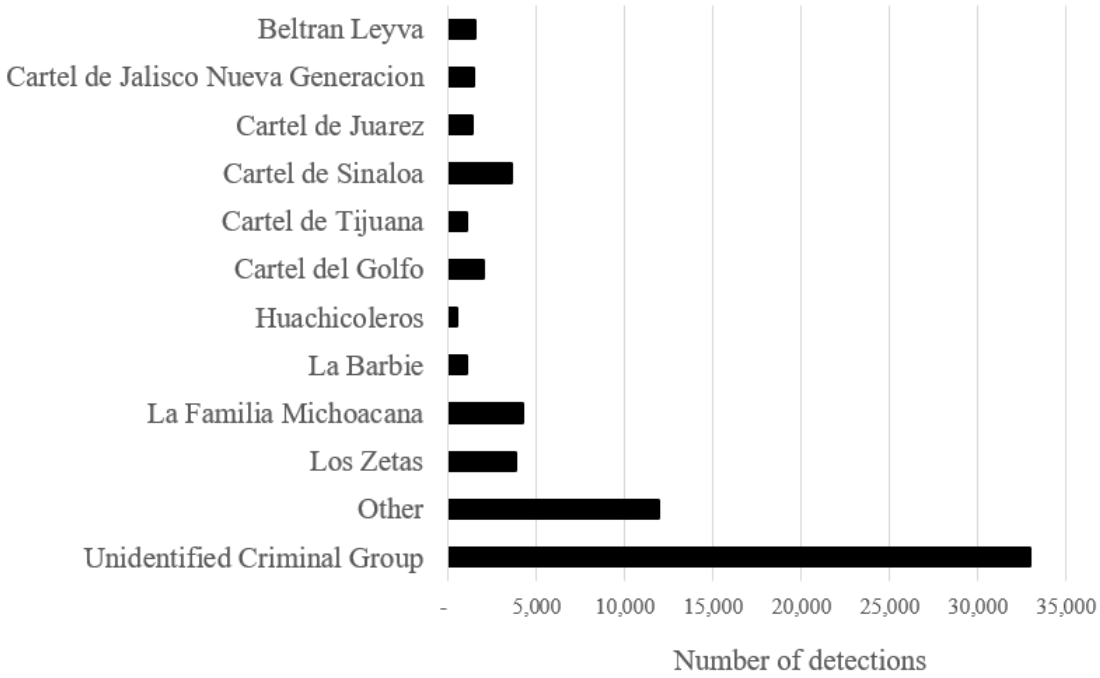


Figure 1: Frequency of actors coded in OCVED 2.0

The database distribution also includes the file `actors_and_locations.xlsx`, which contains actor and location codes in three tabs. The first tab, `actors`, presents the disaggregated list of main Organized Criminal Groups as well as their corresponding subgroups. The second and third tabs, `states` and `municipalities`, respectively, include the official state and municipality names and codes as provided by INEGI, the Mexican Census Bureau.

5. How to Cite

Users can cite the paper and the database as:

> Osorio, Javier, and Beltrán, Aejandro. (2020). "Enhancing the Detection of Criminal Organizations in Mexico Using ML and NLP", *2020 International Joint Conference on Neural Networks (IJCNN)*, Galsgow, UK, July, pp. 1-7, doi: 10.1109/IJCNN48605.2020.9207039, `https://ieeexplore.ieee.org/document/9207039`

Citation in BibTeX:

```
@article{Osorio2020,
author = {Osorio, Javier, and Beltrán, Aejandro},
title = {Enhancing the Detection of Criminal Organizations in Mexico Using
ML and NL},
url = {https://ieeexplore.ieee.org/document/9207039},
Journal = {2020 International Joint Conference on Neural Networks,  IJCNN 2020},
pages = {1--7},
year = {2019},
month = {July},
doi = {10.1109/IJCNN48605.2020.9207039},
place = {Glasgow, UK}
}
```

6. Data Access

The database can be requested at `https://www.ocved.mx/`. By submitting a data request, users will receive an automated email with the link to the database and its related documentation. The data and its corresponding documentation is also directly available in GitHub at: `https://github.com/javierosorio/OCVED_2.0`.

7. Interactive Web Map Application

OCVED has a website, `https://www.ocved.mx/`, that includes an interactive web application where users can visualize the temporal and spatial trends of Organized Criminal Violence using a heatmap to display the concentration of reports in a given location.

The interactive web map includes several features as indicated in Figure 3. The

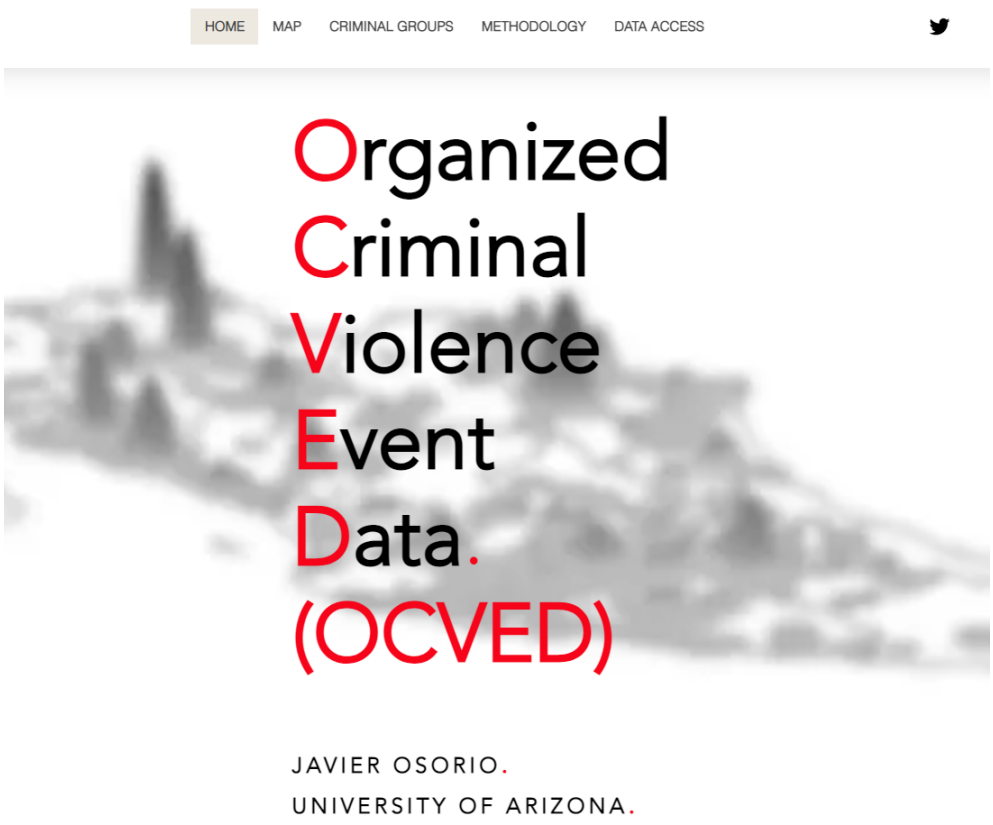


Figure 2: Website `www.ocved.mx`

map application allows the users to zoom in or out the territory to change the level of visualization (see function 1). At each scale, the map automatically adjusts the heatmap to provide the most accurate visualization of criminal group density in the selected area. The interactive map also enables a dynamic data visualization over time (see function 2). When activated, this function aggregates two years of data and renders a heatmap for each two-year period between 2000 and 2018. The map application also allows users to visualize a specific organization by selecting one or more main actors (see function 3). Finally, the interactive map allows users to select and visualize a specific subgroup associated with a main Organized Criminal Group (see function 4).

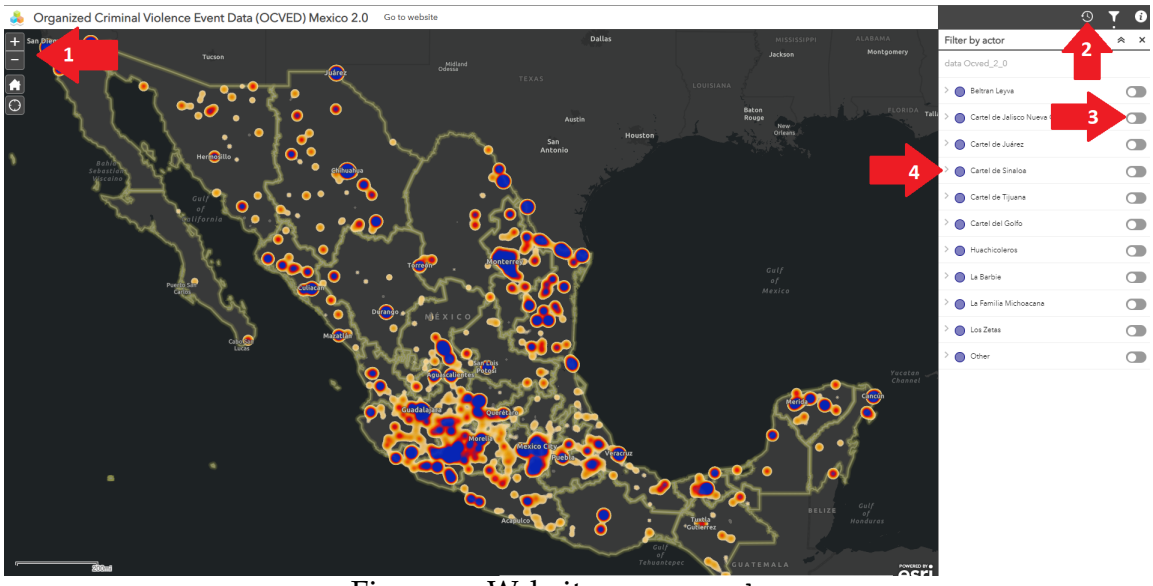


Figure 3: Website `www.ocved.mx`

8. Information Sources

Researchers have long warned about the biases and methodological problems of databases that rely on a single information source as they may suffer coverage bias (Davenport and Ball 2002, e.g.). To minimize concerns about coverage bias, OCVED relies on 105 information sources that issued news reports written in Spanish between 2000 and 2018.

Table 2 reports the main types of information sources, which includes four federal government agencies, 32 local government agencies (one per each state), 11 national newspapers, and 58 local newspapers (at least one per each state). This combination of official and public sources at the national and local level minimizes the risk of coverage and description bias in the database.

Table 2: Number of Information Sources

Source Type	Number of sources
Federal government agencies	4
Local government agencies	32
National newspapers and magazines	11
Local newspapers	58
Total	105

8.1 Government sources

Government information sources at the federal level include:

- Federal level: The Federal Security Ministry, *Secretaría de Seguridad Pública* (SSP)
- Federal level: The Army, *Secretaría de la Defensa Nacional* (SEDENA)
- Federal level: The Navy, *Secretaría de Marina Armada de México* (SEMAR)
- Federal level: The Office of the Attorney General, *Procuraduría General de la República* (PGR)
- State level: Official sources at the local level came from the offices of State Attorney Generals, *Procuradurías de Justicia Estatales* (PJEs) for each of the 32 states.

8.2 Newspaper sources

Scholars have long warned about the limitations and biases inherent to newspaper databases that rely on a single information source (Davenport and Ball 2002, e.g.). To reduce the problems of coverage bias, OCVED relies on a multitude of Mexican newspapers at the national and state-level written in Spanish. The list of national newspapers includes *Servicio Universal de Noticias* (also known as *El Universal*), *El Economista*, *El Financiero*, *Excélsior*, *Notimex–Nacional* (the state news agency), *Reforma*, *La Jornada*, *El Sol de México*, *Milenio Diario*, *Revista Proceso*, and *La Crónica de Hoy*. It is known that different national level newspapers in Mexico have different coverage and ideological orientations. For example, *Reforma* has better coverage of the north of the country and is usually considered to be a conservative newspaper. In contrast, *La Jornada* has better coverage of the south of Mexico and often takes a left-wing view in its reports. Having several national newspapers reduces the coverage and ideological limitations of each individual source.

National newspapers may pay limited attention to isolated events of organized criminal violence at the subnational level or they may also have limited space in their online

or printed outlets that may restrict the number of local level stories they publish. In consequence, news that are important at the local level often do not find their way up to national newspapers. Space limitations and editorial decisions often prevent a large number of local news stories from appearing in national newspapers. In order to minimize problems of media under-reporting between the national and local levels, this research also collected data from 58 local newspapers. Table 3 reports the complete list of sources considered in OCVED, including the national and local newspapers.

Table 3: List of Information Sources

Source type	**Name of information source**
Federal government agencies: 4	
	Army: Secretaría de la Defensa Nacional (SEDENA)
	Federal Police: Secretaría de Seguridad Pública (SSP)
	Navy: Secretaría de Marina Armada de México (SEMAR)
	Attorney General: Procuraduría General de la República (PGR)
Local government agencies: 32	
	State Attorney Generals for all states: Procuradurías Estatales
National level newspapers and magazines: 11	
	Servicio Universal de Noticias
	El Economista
	El Financiero
	Excélsior
	Notimex - Nacional
	Reforma
	La Jornada
	El Sol de México
	Milenio Diario
	Revista Proceso
	La Crónica de Hoy
Local level newspapers: 58	

Continued on next page

Table 3 – *Continued from previous page*

Source type	**Name of information source**
Aguascalientes	El Sol - Regional Newspapers
Baja California	El Mexicano
	El Sol - Regional Newspapers
	La Voz de la Frontera
Baja California Sur	El Sudcaliforniano
Campeche	Diario de Yucatán - Campeche
Chiapas	Notimex - Estados
Chihuahua	Diario de Juárez
	El Diario de Chihuahua
	El Diario de Delicias
	El Diario de Nuevo Casas Grandes
	El Diario de Parral
	El Sol - Regional Newspapers
Coahuila	La Opinión
	Milenio de Torreón
	El Sol - Regional Newspapers
Colima	Notimex - Estados
Distrito Federal	El Sol - Regional Newspapers
Durango	El Sol - Regional Newspapers
Estado de México	Milenio Estado de México
Guanajuato	Periódico A.M. Celaya
	Periódico A.M. Guanajuato
	Periódico A.M. Irapuato
	Periódico A.M. La Piedad
	Periódico A.M. León
	Periódico A.M. San Francisco del Rincón
	Milenio - León
	El Sol - Regional Newspapers
Guerrero	El Sol - Regional Newspapers
Hidalgo	Milenio Pachuca

Continued on next page

Table 3 – *Continued from previous page*

Source type	**Name of information source**
	El Sol - Regional Newspapers
Jalisco	Mural - Newspaper
	Milenio Guadalajara
Michoacán	El Sol - Regional Newspapers
Morelos	Ecos de Morelos - La Unión de Morelos
	El Sol - Regional Newspapers
Nayarit	Notimex - Estados
Nuevo León	Milenio Diario de Monterrey
	El Norte - Newspaper
Oaxaca	Notimex - Estados
Puebla	Milenio Puebla
	El Sol - Regional Newspapers
Querétaro	Diario de Querétaro
	El Occidental
	El Sol - Regional Newspapers
Quintana Roo	Notimex - Estados
San Luis Potosí	El Sol de San Luis
	El Sol - Regional Newspapers
Sinaloa	El Sol - Regional Newspapers
Sonora	Notimex - Estados
Tabasco	Milenio Villahermosa
Tamaulipas	Milenio Diario de Tampico
	El Sol - Regional Newspapers
Tlaxcala	El Sol - Regional Newspapers
Veracruz	Milenio Xalapa
	El Sol - Regional Newspapers
Yucatán	Diario de Yucatán
Zacatecas	El Sol - Regional Newspapers

Since the content of the news articles in EMIS is protected by copyright to the original newspapers and their authors, it is not possible to make them publicly available.

8.3 News Gathering

To gather information from Mexican news outlets in a systematic manner, OCVED relied on the services of Emerging Markets Research, Data and News (EMIS) (formerly known as Infolatina), a large collection of newspapers containing indexed news articles. As a first stage in the information gathering process, OCVED used a sophisticated query in the EMIS search engine to identify potentially relevant news articles. The query has two main components. The first section includes a list of terms used as search criteria. Many of them are stemmed words followed by an asterisk sign (*) at the beginning or end of the stem to enable any combination of words that match the stem. For example, the stemmed term "`narco*`" would match any combination of words such as `"narco"`, `"narcos"`, `"narcotráfico"`, `"narcótico"`, `"narcotraficante"`, `"narco-bodega"`, etc. In this way, the query provides an effective way of having and encompassing search with a relatively small number of terms.

The second component of the query includes a long list of terms used to exclude articles that may not be relevant for this study but may tangentially include some of the search terms indicated in the first part of the query. This is an important part of the query as it aims to exclude news reports that refer to sports[1], or may make reference to deaths in the context of violent conflicts in other parts of the world (e.g. "Iraq"), natural disasters, diseases, etc. The EMIS search engine also includes a feature to filter out duplicate news stories. Although it is not 100% accurate, this de-duplication approach helps to minimize artificial inflation of relevant news articles. The text below presents the query used in EMIS to identify potential news articles relevant to OCVED:

[1]News articles on sports often use bellicose language such as "attack" or "assault" that may be conflated with similar terms in the context of organized criminal violence events.

("crimen organizado" OR "delincuencia organizada" OR AFI OR amapola OR arma* OR armas OR asesin* OR atac* OR ataque OR capo OR cartel OR cocaina OR cuerp* OR crim* OR delincuen* OR decapit* OR dispar* OR drog* OR ejecucion* OR ejecuta* OR ejercito OR enfrenta* OR herid* OR mariguana OR marina OR mata OR mata* OR mato OR militar* OR muert* OR narco* OR PFP OR pistol* OR polici* OR restos OR rifl* OR secuestr* OR sicari* OR SSP OR tortur* OR zeta*) NOT *(Qaeda OR Oppenheimer OR "apoyo al campo" OR "Banco de Mexico" OR Laden OR "bolsa Mexicana" OR "canasta basica" OR "Carlos Marin" OR Aristegui OR "Gomez Leyva" OR invitado OR Dresser OR "desastre natural" OR "desastres naturales" OR Bartolome OR Fondevilla OR forestal OR "JAQUE MATE" OR Chabat OR "Fuentes Aguirre" OR Zuckermann OR Curzio OR "liga mexicana" OR lector OR Meyer OR Loret OR "Luis Rubio" OR Merino OR "medio ambiente" OR "Obra publica" OR "Obras publicas" OR "Plaza Publica" OR "regreso a clases" OR Sarmiento OR "sintesis de medios" OR "sintesis de prensa" OR "sintesis informativa" OR "tala clandestina" OR taxi OR "templo mayor" OR tolvanera OR "toma clandestina" OR accidente OR actor OR actriz OR acuerdo OR adelata OR Afganistan OR aguacero OR AH1N1 OR alimento OR AMLO OR anuncia OR aprueba OR atropella OR bacteria OR ballena OR beatificacion OR belleza OR bono OR bronquitis OR caiman OR calor OR cancer OR campeonato OR cantante OR carnaval OR ciclista OR cine OR clima OR colera OR comentario OR comisionado OR cortocircuito OR cultura OR debate OR delfin OR deporte OR desbordamiento OR dialogo OR ebrio OR ecologia OR educacion OR Egipto OR empresa OR escritor OR espectaculos OR estrellas OR Europa OR eutanasia OR evenenado OR felicita OR festejo OR fiscal OR frio OR futbol OR Gaddafi OR ganancia OR granizo OR gripa OR gripe OR inaugura OR indigna OR infarto OR influenza OR infraccion OR infraccion OR Ingalterra OR intoxicado OR inundacion OR invito OR Irak OR islam OR juego OR jugar OR justific* OR laboral OR Latinoamerica OR Libia OR llam* OR afirm* OR nieg* OR exig* OR reclam* OR critic* OR promet* OR llueven OR lluvia OR Marruecos OR medicament* OR medicin* OR metrobus OR Mitofsky OR modernizacion OR multa OR multa* OR mundial OR música OR nepotismo OR nuclear OR Obama OR olimpiada OR opinion OR Osama OR Paulette OR pelicula OR peticion OR pirata OR pirotecnicos OR PROFEPA OR promo* OR propon* OR rabia OR reclam* OR reconoc* OR religi* OR reproch* OR ring OR rugido OR sarampion OR sindica* OR sismo OR supervisa OR talamontes OR Tabaco OR teatro OR temperatura OR terremoto OR tormenta OR torneo OR trascendio OR tuberculosis OR urge OR urgen OR utilidad* OR vacacio* OR varicela OR veneno OR videojuego OR viru* OR EZLN OR EPR OR Zapatist* OR "nota diplomatica" OR fraude OR "subcomandante marcos" OR ambulantes OR pescado OR tortuga OR IFE OR "dia del trabajo" OR "desvio de recursos" OR campeonato OR dictador OR suicid* OR "casas de bolsa" OR Iraq OR hipotec* OR Appo OR Appistas OR "Boletin de prensa" OR misa OR (reforma & (energetica OR laboral OR politica OR tributaria)))*

9. Document Classification

A considerable challenge in text analysis relates to effectively identifying relevant news articles that are pertinent to the domain of study. Given the enormous amounts of information available on the web, it is very easy to get inundated with irrelevant information that, if kept, could artificially inflate the volume of information to be processed in could lead to including false positives in the output data. Both of these risks are likely to distort the conclusions derived from the data. In addition, failing to identify the relevant types of documents could generate problems of false positives in the event coding process. The query discussed in section 8.3 helps to reduce the amount of not valid news articles, but this approach still is not sufficient to ensure high quality information input.

To select relevant documents that inform the event coding process, this research used two different approaches. First, in the initial stage of this project, the methodology relied on manual classification of relevant news articles using a team of human coders applying the rules indicated in a codebook. The output of this process generated a collection of relevant news articles containing information about organized criminal violence in Mexico. In the second stage, the methodology relied on the insights gained from the manual classification task to inform a Machine Learning algorithm to classify relevant news articles. The following sections discuss both the manual and automated approaches used to classify news articles as domain relevant for this study.

9.1 Manual Document Classification

The manual classification consisted of a group of research assistants who received specific training and supervision to classify news stories and press releases as relevant or not based on the inclusion and exclusion criteria outlined below. The research protocol used manual classification to identify relevant stories published between 2000 and 2010.

Definitions and research assumptions

This section presents the definitions and assumptions that guide this research project and describes the criteria for selecting press releases in the stage of extracting reports of violence from websites.

Following Reuter (2009), this study defines *organized crime* as a set of criminal groups that are organized in a stable and hierarchical manner that rely on the use of violence or its credible threat to establish their presence in a given territory or in a certain illicit market. This territorial or sector presence generates economic benefits to the organization and its members.

In line with Kalyvas (2006), *violence* is defined as deliberate action to inflict physical harm on people and their property. Violence can be used strategically to prevent certain behavior or tactically to eliminate a target. *Organized criminal violence* refers to both the process and the result of violence perpetrated by criminal organizations or by the coercive apparatus of the state in its fight against criminal groups. This research focuses on large-scale violence related to organized crime. To do this, it analyzes the aggregate levels of violence caused by the struggle between the state and criminal organizations, as well as the violence between rival criminal organizations, violence within them, and violence against the general population.

This research assumes that organized criminal groups in Mexico are not only dedicated to the cultivation, production, reception, transportation and local and international sale of illicit drugs. Criminal groups in Mexico also operate broadly in other illicit markets such as kidnapping, money laundering, extortion, and human trafficking, oil theft.

Inclusion Criteria

The following article selection criteria apply for the extraction of press releases issued by government departments, as well as for press releases produced by national and local newspapers and magazines.

The news articles relevant to this study must include *factual* information related to organized criminal violence incidents in Mexico or about the law enforcement efforts directed at criminal organizations.

An event is defined by three key elements:

- Source: Refers to the actor who perpetrates the violent action
- Action: Refers to the violent act itself
- Target: Refers to the actor against whom the violent action is directed

In the example "a group of hit-men attacked federal police", the three elements that define the event are: `<source = hitmen>`, `<action = attacked>`, and `<target = federal police>`.

To **include** a press release as relevant, it must contain one of the following elements:

1. Violent acts such as confrontations, murders, kidnapping, extortion, torture or assaults allegedly perpetrated by members of criminal organizations or state security forces.
2. Acts of physical or material violence, wither they are lethal or non-lethal, that are explicitly related to organized crime or that presume the actions of these armed groups.
3. Threats of the use of physical or material violence. These include news articles that talk about how people react to violence or threats of violence.
4. Violent acts that do not make explicit mention of organized crime but that present characteristic features of the *modus operandi* of this type of organization, such as:
    - Use of high caliber weapons
    - Indications of torture or mutilation
    - Executions with shot on the head

- Groups or armed actors
- Vehicle convoys
- Narco-messages "*Narcomantas*"

5. Include news articles containing information about the actions of government agencies or officials in relation to activities directed against criminal groups. This may include:
    - Action of government agencies or officials in the fight against organized crime
    - Government agencies or officials as victims of organized crime
    - This also includes former officials
6. Seizures of weapons, properties, drugs, laboratories, money, vehicles allegedly related to organized crime.
7. Release of kidnapped persons.
8. Arrest of people allegedly related to organized crime.
9. News articles related to kidnapping, extortion, seizure of large sums of money or money laundering even if they do not explicitly mention drug trafficking.
10. Include articles on the actions of self-defense groups.
11. Include news articles on violence in prisons.
12. Include notes on forced displacement due to violence.

Exclusion Criteria

The document classification should **exclude** news articles that meet any of the following criteria:

1. Opinion articles or editorials from newspapers and magazines.

2. Statements, speeches or opinions of government authorities (national or foreign), opinion leaders, churches, civil organizations, journalists, etc.
3. Press releases from government secretariats that refer to guerrilla groups (e.g. EZLN, EPR, EPRI).
4. Press releases from government ministries that are not related to the fight against organized crime, although they are related to general activities of said ministry (e.g. exclude SEMAR press releases that report the rescue of fishermen).
5. Exclude news articles related to:
    - Crimes of the common jurisdiction (e.g. robbery, assault, rape)
    - Wrongful death (e.g. accidents)
    - Sexual crimes.
    - Passion crimes.
    - Femicides.
    - Human trafficking (e.g. prostitution rings or migrants).
    - Adult or child pornography.
    - Accidents.
    - Natural Disasters.
6. Exclude news articles or press releases that make general recounts of results in the fight against organized crime that summarize activities of more than one month (e.g. monthly reports, government reports, reports from NGOs or international organizations).
    - As an exception, include news articles that summarize events from the last few days. For example, Monday notes often summarize what happened over the weekend. In those cases, we do include the note.

7. Exclude articles that refer to violent events related to organized crime that have occurred in another country, even if they include Mexican actors.
8. Exclude notes that refer to acts of corruption by government officials (e.g. politicians, military, police).
    - The exception is if the story refers to collusion or corruption related to criminal organizations (e.g. the narco paying the police). In those cases, we do include the note.
9. Request of individuals or organizations to government authorities to investigate a crime.
    - Do not include news articles where a judge issues an arrest warrant (this is just a warrant but does not mean that a person has been arrested).
    - Do not include stories where a private person or agency (lawyer, politician, group, NGO) blames or accuses another person of committing a crime (this can be considered only as a opinion, but not fact).
10. Announcements or notifications from authorities about the initiation of investigations or ongoing investigations.
11. Social movements or protests claiming lack of security.

News Gathering

Each member of the research team had access to EMIS and ran the query discussed in section 8.3 on the selection of newspapers indicated in section 8.2. Research assistants conducted this search in a systematic manner by progressively covering month by month for each given year covered in the project. This helped making a steady and systematic progress in the news gathering protocol. In addition, the EMIS search engine allows

sorting the news articles in chronological order, which enabled research assistants to exlude duplicated or closely duplicated articles that appeared before.

As the research assistants identified a relevant news articles, they gathered their URLs in a spreadsheet. Based on this list of selected news stories, the next step in the methodology consisted on downloading the content of .html files of the selected news articles using Web Text Downloader. In contrast to indiscriminate web scrapers that download the entire content of websites, Web Text Downloader is a customized application specifically developed for this project that allows targeted scraping within password protected environments, such as the EMIS interface.

The output of this stage was a collection of news articles from national and local newspapers as well as press releases from government agencies between 2000 and 2010.

9.2 Document Classification using Machine Learning

The process of automated document classification includes two main steps. The first one focuses on scraping news articles in order to update the corpus and cleaning the text. The second stage consists of relying on a variety of ML algorithms to identify the best performing classifier.

9.2.1 Scraping and pre-processing

The first step for classifying relevant documents using Machine Learning algorithms is to gather an updated collection of news articles from 2010-2018. To do so, the team of researchers relied on EMIS University, a news aggregator containing a large collection of newspapers in multiple languages. To query EMIS, the team of researchers relied on the applications' internal search engine to assess 77 Mexican newspapers published in Spanish between 2010-2018. The output contained thousands of potentially relevant news articles. To scrape all the search results we use `Selenium` Selenium (2013), this helped us to save each article using a unique identifier. In this way, we gathered a

collection of documents including the news article's title, content, and corresponding unique URL. To process the news articles, we use `BeautifulSoup` Richardson (2007) to parse and clean the content of the articles and save the processed news stories in `.json` format. The result of this process generated a corpus comprising a total of 158,514 articles.

9.2.2 Machine Learning classifier

In order to identify the presence of Organized Criminal groups in Mexico from news papers, it is essential that the input information is valid. Otherwise, there is a high probability of generating false positives derived from invalid news stories. The collection of news articles retrieved from the EMIS search engine includes a non-ignorable number of non-relevant news articles. To address this challenge, the team of researchers relied on a variety of Machine Learning (ML) algorithms to classify the news articles. The ML tasks corresponds to a binary classification in which the algorithm uses training data annotated by human coders to classify a news article as relevant or not relevant.

The first step of the classification process is to generate the training data. To do so, the team relied on the support of three human coders who annotated the news articles as reject/accept (0.1) in a sample of 30,842 articles. Following the conceptual guidelines of the codebook discussed above, the annotators classified as accept those articles describing events of organized criminal violence. These type of incidents include descriptions of confrontations between criminals; armed clashes between criminals and government authorities; arrests; drug seizures; seizures of assets or weapons; or the capture of high-profile targets. The human coders marked as irrelevant articles not making direct reference to organized criminal violence events, articles presenting editorial opinions about criminal violence, or summaries from government authorities providing a cumulative security activity report. The human annotators also classified an additional random sample of 1,000 articles to assess their degree of agreement. The intercoder

reliability of their annotations reached 90.4% and a Fleiss' Kappa of 0.704.

In order to effectively train a ML algorithm, the training data requires a balanced number of annotations in each of the classification categories, such that the categories are evenly divided as 50%-50%. This helps to provide the about same number of examples to the ML algorithm so it learns how to classify the "accept" news articles as well as the "reject" stories. Failing to provide a balanced training data would bias the ML algorithm by providing more examples in one category than in the other one. In consequence, the ML algorithm would likely preform much better at classifying the category for which it has a larger number of examples and display lower performance in identifying the category with fewer examples.

Initially, the training data scraped from EMIS was not balanced as it contained an acceptance rate of 23%, while the rest were not-relevant news articles. To balance the training data, we relied on a set of news articles from a manually collected by oso-rio2015contagion, which comprises a collection of relevant articles for the same time period. Adding this set of relevant news articles, increased the sample of the training data to a total of 60,837 news articles. The balance assessment indicates that 61% of them correspond to the category "accept". This classification distribution is not problematic as it provides slightly more examples for the ML algorithm to accurately classify relevant news articles.

Before feeding the training data to the ML algorithm, it is necessary to pre-process the text. To do so, the data preparation process consists of normalizing the text and eliminating diacritic characters commonly present in Spanish, remove digits, punctuation marks, and stop words. To reduce words to their lemma, we use the Spanish language lemmatizer from `SpaCy` (Honnibal and Montani 2017). This lemmatization procedure helps to make the ML training more efficient and effective by reducing words to their root mode. The next step consists of converting the data into a features matrix capped at 5,000 features using `TfidfVectorizer` from `sci-kit learn` Pedregosa et al. (2011). The

pipeline shuffles and splits the training data into 5 folds, evaluates each model using k-fold cross validation, and assigns 10% of data for testing.

To discriminate between the relevant and not-relevant news articles, we consider a variety of Machine Learning algorithms for this binary classification task. We use a wide range of ML models, including traditional approaches, ensemble methods, to deep learning models. There are three Convolutional Neural Network (CNN) models reported from a random grid search, with a shared vocabulary size of 85,178 and an embedding dimension of 50. We include transformer models using the `simple transformers` library from Rajapakse (2020) and the transformers library `Hugging Face` (Wolf et al. 2019).

Instead of using a single ML algorithm, the application follows the standard in computer sciences and considers multiple algorithms and compares their performance using the F-1 score. The F1 score is the standard evaluation metric for ML models given its weighted performance reporting that takes into account both precision and recall. In this way, our classification task puts a variety of models to compete so that we can identify the one providing the best performance. The binary classification considers eleven distinct ML models:

- AlBERT using the `albert-base-v1` model
- CNN 1 with 128 filters and a kernel of size 3
- CNN 2 with 64 filters and a kernel of size 7
- CNN 3 with 128 filters and a kernel size of 7
- Extreme Gradient Boosting (XGB)
- Logistic Regression (LR)
- Multilingual BERT uncased for Spanish with 5-fold cross validation.
- Multilingual DistilBERT uncased

- Multinomial Naive Bayes (NB)
- Random Forest Classifier (RF)
- Support Vector Machine (SVM)

Figure 4 reports the performance of different models based on F1 scores. Starting with the models reporting the lowest performance, the Extreme Gradient Boosting (XGB) model reports an F1 average of 0.902. The next model is ALBERT, which reaches an F1 average of 0.914 across folds. The Multinomial Naive Bayes (NB) model reports an average F1 score of 0.915 across folds. The different Convolutional Neural Network models provide similar results. CNN 3 averages 0.926 F1, while CNN 2 reports an average of 0.9268 F1, and CNN 1 produces marginally better results with an F1 of 0.9269. Multilingual BERT-uncased performs at an F1 of 0.928, which is the same results obtained by the Random Forest Classifier (RF). The DistilBERT model has slightly better results with an average F1 of 0.937. There is a substantive increase in classification performance with the Support Vector Machine (SVM) model, which produces an F1 average 0.947. Finally, the Logistic Regression (LR) model provides the best performance with an average F1 of 0.949. It is interesting how the simplicity of the probabilistic approach of the LR model outperforms highly sophisticated ML algorithms such as BERT or SVM.

Given that the LR classifier provides the best performance, this is selected as the algorithm to classify the entire collection of 158,514 news articles. Before applying the LR model, the team pre-processed the entire corpus with the normalization, cleaning, and lemmatization processes used for the training data. Then, the researchers applied the LR classifier to this universe of stories for binary classification. The result of this classification task is 43,681 news articles selected as relevant for organized criminal violence. To validate the classification output, a group of human coders briefly reviewed a small sample of the classified articles and confirmed adequate performance.

Details on the technical implementation of the different ML models used in this anal-

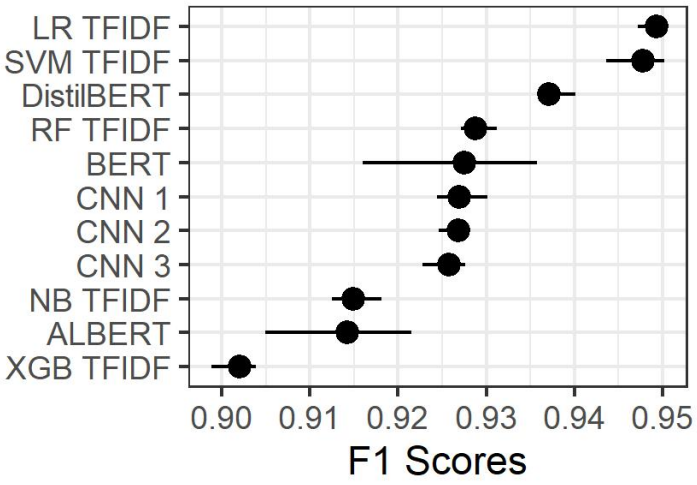


Figure 4: Machine learning model performance.

ysis is available at: `https://github.com/AlejandroBeltranA/OCVED-ML`

10. Computerized Text Extraction

To generate the database of violent presence of Organized Criminal Groups in Mexico, the project relies on Eventus ID, a software for supervised event coding from text written in Spanish (Osorio and Reyes 2017). Eventus ID belongs to the family of event coding protocols that rely on sparse parsing to process the text using a rule-based approach. Following the approach of other event coders such as KEDS, TABARI, and Petrarch (Schrodt, Davis and Weddle 1994; Schrodt 2009; Schrodt, Beieler and Idris 2014), Eventus ID uses a set of dictionaries or coding rules to identify event data, which is defined as a categorical description of someone (source actor), doing something (action), to someone else (target actor), in a give date (time), and in a specific place (location). In this way, the software processes textual information in order to identify who did what to whom, when and where. Once the coder identifies the event components in the corpus, the program extracts the textual information and transforms it into numeric format for data visualization or statistical analysis.

In this project, we are only focused on identifying the geographic location of Orga-

nized Criminal Groups as reported in the news stories. To conduct this task, the coding protocol only uses a dictionaries of actors and locations (states, municipalities, and location filters). Advanced users will be able to find further technical details about Eventus ID and execution instructions in Github: `https://github.com/javierosorio/Eventus_ID_2.0`.

Figure 5 describes the general process of geo-locating actors using Eventus ID. The algorithm implements the following steps:

1. The software takes as the main input the corpus of *Noche y Niebla* narratives of human rights violations.

2. Eventus ID uses the content of the actors dictionary as search criteria to look for a match in each line of the corpus.

3. If there is an actor match, Eventus ID stores the matching text and identifies the corresponding numeric code as indicated in the actors dictionary.

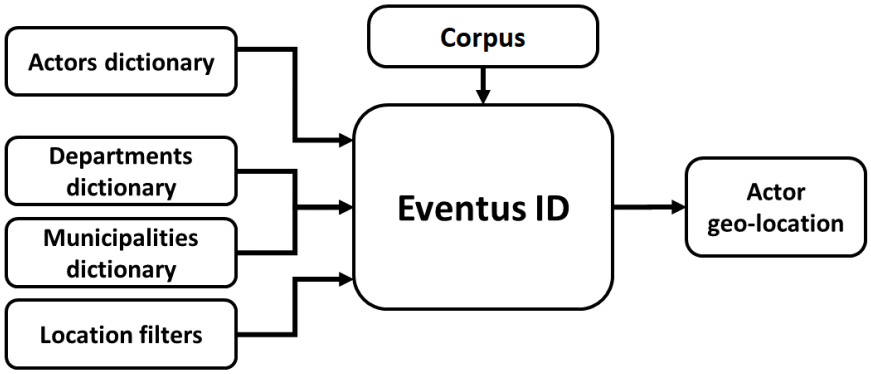
4. For each corpus line containing an actor match, the program uses the dictionaries of departments and municipalities (which correspond to states and counties, respectively) to look for a location name (also know as toponym) in the same line in which it identified an actor.

5. If there is a match with the locations dictionaries, the program then verifies the location filters to identify whether or not this location name is a false positive or an actual toponym.

6. If there is no match in the location filters, Eventus ID saves the coded actor and locations in the output database.

The actors dictionary contains an encompassing list of 8,576 names of armed actors (last update on 12/29/2021). The list of nouns contains a collection of names of

Figure 5: Actor geo-location using Eventus ID



state agencies related to security and law enforcement, as well as Organized Criminals Groups. The actors dictionary also includes acronyms, subgroup names, main leaders, and aliases. For redundancy purposes, the actors dictionary considers variations of different armed actor's names. This is important as it allows to cover different variations in unstructured text. The list of actors considered in this study concentrates on nine main Organized Criminal Groups and three residual categories. These main groups are disaggregated into their corresponding factions, sub groups, or spin-offs. In total, the analysis tracks the presence of 140 specific criminal groups. Table 4 presents the list of Organized Criminal Groups considered in this research grouped by the main organization. The list also presents a set of subgroups, branches, and spin-off criminal groups associated with the larger criminal organizations.

To identify the locations, Eventus ID relies on two dictionaries containing detailed lists of toponyms. One of the dictionaries corresponds to states and the another one for municipalities (equivalent to counties). In this way, the software is capable of geo-referencing actors at two levels of sub-national analysis. The states dictionary contains the names of all 32 states in Mexico. In a similar manner, the municipalities dictionary includes all the municipalities in the country and a variety of name variations. These dictionaries include the official state or municipal codes used by Mexican Census authority, Instituto Nacional de Estadística, Geografía e Informática (INEGI). Using these official location codes facilitates merging the output coding data with official statistics

Table 4: List of Organized Criminal Groups

BELTRAN LEYVA	Los Ardillos	Los Colmenos
Beltran Leyva	Los Rojos	Los Coyotes
Cartel Independiente de Acapulco	**LA FAMILIA MICHOACANA**	Los Danieles
La Nueva Administracion	La Familia Michoacana	Los Dannys
La Oficina	Caballeros Templarios	Los Dedos
Limpia Mazateca	El Charro	Los Flacos
Los Pelones	La Empresa	Los Garibay
Nuevo Cartel de la Sierra	La Nueva Familia Michoacana	Los Gaseros
CARTEL DE JALISCO NUEVA GENERACION	Los Jaguares	Los Gatos
	Los Perez	Los Indios
Cartel de Jalisco Nueva Generacion	Los Pumas	Los Japos
Cartel de Colima	Los Troyanos	Los Jarochos
Cartel del Milenio	Los Viagras	Los Jarquin
Gente Nueva	**LOS ZETAS**	Los Juchitan
La Resistencia	Los Zetas	Los Juniors
Los Mata Zetas	Banda del Chaparro	Los Kinkones
Los Valencia	Cartel del Noreste	Los Lagartus
CARTEL DE JUAREZ	Los Broncos	Los Limones
Cartel de Juarez	Los Cotorros	Los Maximos
La Linea	Los Guerreros	Los Mellizos
Los Aztecas	Los Lancheros	Los Mudos
CARTEL DE SINALOA	Los Numeros	Los Negros
Cartel de Sinaloa	**UNIDENTIFIED CRIMINAL**	Los Ninnos de Oro
Artistas Asesinos	Unidentified Criminal Group	Los Nortenos
Cartel del Pacifico	Criminal	Los Nuevos Pelones
El Chapo	Drugtrafficker	Los Ortiz
El Guero Palma	Hitman	Los Pajaros
Los Jaguares	Key operator	Los Palafox
Los Salazar	Leader	Los Panchos
Nacho Coronel	Cartel member	Los Pedraza
Zambada	**OTHER**	Los Perros
CARTEL DE TIJUANA	Cartel de Neza	Los Petriciolet
Cartel de Tijuana	Cartel de Oaxaca	Los Pipas
Arellano Felix	Cartel de Tlahuac	Los Pipos
Faccion de El Teo	Individual	Los Punta Norte
CARTEL DEL GOLFO	Los Antrax	Los Purina
Cartel del Golfo	Los Babicoras	Los Rancheros
Los Cuervos	Los Benitez	Los Ratoncitos
Los Escorpiones	Los Cachos	Los Rodolfos
Los Halcones	Los Campesinos	Los Santeros
Los Linces	Los Caraveo	Los Simpson
Los Metros	Los Cazo	Los Smith
HUACHICOLEROS	Los Chatos	Los Tablajeros
Huachicoleros	Los Chavez	Los Temixco
Cartel de Santa Rosa de Lima	Los Chibuyar	Los Thunder
LA BARBIE	Los Chinconcuac	Los Veneros
La Barbie	Los Chinos	Los Yonqueros
Cartel del Pacifico Sur	Los Cholos	Los Zampayo
Guerreros Unidos	Los Chutas	Los Zodiaco
La Barredora	Los Ciruelos	M60
La Mano con Ojos	Los Colin	Union Tepio

and other databases.

A common challenge in computerized event data relates to correctly identifying the place of occurrence of an event (Chalabi 2014; Lee, Liu and Ward 2019). To address the challenge of geographic disambiguation, Eventus ID uses a filter of locations to eliminate matches that might look as a location names, but do not actually refer to locations. For example, the locations filter includes the name of "`Cartel de Sinaloa`" to prevent the program erroneously identify the name of this criminal organization as the state of "`Sinaloa`."

After detecting an actor in a specific line, Eventus ID uses the state and municipality dictionaries to identify the location of an event as mentioned in such line of the corpus. After identifying a location, the software verifies if there is a match in the location filter. If that is the case, the software ignores this name as a possible location. If there is no match, the program saves the location name and corresponding numeric code in the output database.

For each municipality, the database includes the geographic longitude and latitude coordinates as reported by the Mexican Instituto Nacional de Estadística, Geografía e Informática. This allows projecting the coded data into a map. It is important to clarify that the projection *does not provide the exact location* of an event as based on a neighborhood, specific street address, latitude or longitude coordinates, or a geo-tag marked in the field. The projection in the map is just a generic approximation of the location based on the geographic coordinates of the capital of each municipality.

After coding actors and locations using Eventus ID, the methodology considers a post-coding process for cleaning the data in Stata. This protocol includes a deduplication routine to eliminate multiple matches of the same actor in the same municipality day. In this way, the analysis avoids artificial inflation and considers a conservative approach for including in the cleaned database one single actor per organization, for each municipality-day. The protocol also includes a process for location deduplication and

cleaning to reduce concerns of false positives. The final data set includes the type of actor, the specific organization recorded, the date of occurrence, as well as the department and municipality identified.

11. Technical Aspects of the Map

This section provides technical information related to the Geographic Information Systems (GIS) used in this project. We used ArcGIS Pro v2.2 to develop the map and specify its main configuration. After developing the map, we used ArcGIS Online to deploy the interactive map application in the web. Both ArcGIS Pro and ArcGIS Online are proprietary products of the Environmental Systems Research Institute (ESRI).[2] We used these programs through the institutional licenses of the University of Arizona. The coordinates system used in this project corresponds to WGS84 Web Mercator (Auxiliary Sphere).

The visualization presented in the map corresponds to heat maps of violent presence of armed actors. A heat map is a GIS analysis tool that provides a visual representation of high-density of occurrence or clustering of a phenomenon. In contrast to hot-spot analysis tools that rely on statistical measures of spatial auto-correlation, heat maps provide a more flexible (yet, less rigorous) way of clustering the observations and assigning a color gradient to indicate higher concentration.

The raw information used for creating the heat map came from the computerized data generated with Eventus ID. Each data point corresponds to the identification of a specific armed actor in a given municipality-day as mentioned in the *Noche y Niebla* narratives. As indicated before, the incidents do not have the specific xy coordinates of their occurrence. Instead, their position in the map is assigned to the corresponding latitude and longitude of the center of the municipality. As a result, points often overlap

[2] See `https://www.esri.com/`.

in the map projection. This is becomes evident when users zoom into the map.[3]

To build a heat map, ArcGIS uses a normal distribution (also known as Gaussian distribution) to assess the strength of the influence of each point over an specific area as determined by a radius.[4] The program renders a higher intensity to areas that concentrate a higher density of data points. To represent these areas, the program assigns a more intense color gradient than the one assigned to low intensity areas.

In contrast to a fixed raster visualization, the ArcGIS Online application dynamically adjusts the radius to render the heat map based on the projection scale that the user prefers. In this way, the heat map will provide a more aggregate or disaggregate clustering of the data points as the user zooms out or into the map.

12. Related Publications

- Osorio, Javier, Mohamed Mohamed, Viveca Pavon, and Susan Brewer-Osorio. (2019). "Mapping Violent Presence of Armed Actors in Colombia", *Advances of Cartography and GIScience of the International Cartographic Association*, 1(16):1-9, https://www.adv-cartogr-giscience-int-cartogr-assoc.net/1/16/2019/.

- Osorio, Javier and Alejandro Reyes. (2016). "Supervised Event Coding from Text Written in Spanish: Introducing Eventus ID." *Social Science Computer Review*, 35(3):406–416, https://journals.sagepub.com/doi/abs/10.1177/0894439315625475.

[3]To identify how many points overlap in a specific location, users can use the mouse to make a right click on any cluster. This will open a pop-up window indicating the number of points overlapping in this location.

[4]For technical details about the methodology used to generate the heat maps, see: https://developers.arcgis.com/javascript/latest/api-reference/esri-renderers-HeatmapRenderer.html.

13. Version History

- **Version 2.0**:
  - Release date: April, 2022
  - Revised list of actors.
  - Improvement in geographic disambiguation.
- **Version 1.0**:
  - Release date: July, 2020
  - First public version (Beta)

References

Arjona, Ana M. 2011. "Presencia vs. violencia: problemas de medición de la presencia de actores armados en Colombia.".
URL: *http://focoeconomico.org/2011/12/20/presencia-vs-violencia-problemas-de-medicion-de-la-presencia-de-actores-armados-en-colombia/*

Chalabi, Mona. 2014. "Mapping Kidnappings in Nigeria." *Five Thirthy Eight* .
URL: *https://fivethirtyeight.com/features/mapping-kidnappings-in-nigeria/*

Chojnacki, Sven, Christian Ickler, Michael Spies and John Wiesel. 2012. "Event Data on Armed Conflict and Security: New Perspectives, Old Challenges, and Some Solutions." *International Interactions* 38(4):382–401.

Davenport, Christian and Patrick Ball. 2002. "Views to a kill: exploring the implications of source selection in the case of Guatemalan State Terror, 1977-1995." *Journal of Conflict Resolution* 46(3):427–450.

Honnibal, Matthew and Ines Montani. 2017. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." To appear.

Kalyvas, Stathis. 2006. *The Logic of Violence in Civil Wars*. onnecticut: Cambridge University Press.

Lee, Sophie, Howard Liu and Michael Ward. 2019. "Lost in Space: Geolocation in Event Data." *Political Science Research and Methods* 7(4):871–888.
URL: *https://arxiv.org/abs/1611.04837*

Osorio, Javier and Alejandro Beltrán. 2020. Enhancing the Detection of Criminal Organizations in Mexico using ML and NLP. In *IEEE World Congress on Computational Intelligence (WCCI)*. Gasgow, Scottland: .
URL: *https://wcci2020.org/ijcnn-2020-program/*

Osorio, Javier and Alejandro Reyes. 2017. "Supervised event coding from text written in Spanish: Introducing Eventus ID." *Social Science Computer Review* 35(3):406–416.
URL: *https://doi.org/10.1177/0894439315625475*

Osorio, Javier, Mohamed Mohamed, Viveca Pavon and Brewer-Osorio Susan. 2019. "Mapping Violent Presence of Armed Actors." *Advances in Cartography in GIScience of the International Cartographic Association* pp. 1–16.
URL: *https://www.adv-cartogr-giscience-int-cartogr-assoc.net/1/16/2019/*

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg et al. 2011. "Scikit-learn: Machine learning in Python." *Journal of machine learning research* 12(Oct):2825–2830.

Rajapakse, Thilina. 2020. "Simple Transformers.".
URL: *https://simpletransformers.ai/*

Reuter, Peter H. 2009. "Systemic Violence in Drug Markets." *Crime, Law and Social Change* 52(3):275–284.

Richardson, Leonard. 2007. "Beautiful soup documentation." *April* .
URL: *https://www.crummy.com/software/BeautifulSoup/bs4/doc/*

Rios, Viridiana. 2012. "How Government Structure Encourages Criminal Violence. The causes of Mexico's Drug War.". Place: Cambridge, MA.

Schrodt, Philip A. 2009. "TABARI. Textual Analysis by Augmented Replacement Instructions.".
URL: *http://eventdata.parusanalytics.com/software.dir/tabari.html*

Schrodt, Philip A., Brandon Stewart, Jennifer Lautenschlager, Andrew Shilliday, David Van Brackel and Will Lowe. 2010. "Automated Production of High-Volume, Near-Real-Time Political Event Data." unpublished.

Schrodt, Philip A. and David Van Brackel. 2013. Automated Coding of Political Event Data. In *Handbook of Computational Approaches to Counterterrorism*, ed. Devika Subramanian. New York: Springer pp. 23–50.

Schrodt, Philip A., John Beieler and Muhammed Idris. 2014. Three's a Charm?: Open Event Data Coding with EL:DIABLO, PETRARCH, and the Open Event Data Alliance. In *International Studies Association*. Toronto: .
URL: *http://parusanalytics.com/eventdata/papers.dir/Schrodt-Beieler-Idris-ISA14.pdf*

Schrodt, Philip A., Shannon G. Davis and Judith L. Weddle. 1994. "Political Science: KEDS - A program for the machine coding of event data." *Social Science Computer Review* 12(4):561–587.

Selenium. 2013. "Selenium webdriver." *Selenium HQ* .
URL: *https://selenium.dev/documentation/en/webdriver/*

Seybolt, Taylor B., Jay D. Aronson and Baruch Fischhoff, eds. 2013. *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict*. Oxford: Oxford University Press.

Signoret, Patrick, Marco Alcocer, Cecilia Farfan-Mendez and Fernanda Sobrino. 2021. "Mapping Criminal Organizations.".
URL: *https://doi.org/10.7910/DVN/NoKGCZ*

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz and Jamie Brew. 2019. "HuggingFace's Transformers: State-of-the-art Natural Language Processing.".

Zhukov, Yuri M., Christian Davenport and Nadiya Kostyuk. 2019. "Introducing xSub: A new portal for cross-national data on subnational violence." *Journal of Peace Research* 56(4):604–614.