Regional Studies, Regional Science

ISSN: 2168-1376 (Online) Journal homepage: www.tandfonline.com/journals/rsrs20

Measuring organised crime in Italy: dual composite indicators integrating text mining and official statistics

Antonio Fabio Forgione, Carlo Migliardo & Paolo Mustica

To cite this article: Antonio Fabio Forgione, Carlo Migliardo & Paolo Mustica (2026) Measuring organised crime in Italy: dual composite indicators integrating text mining and official statistics, Regional Studies, Regional Science, 13:1, 2652760, DOI: 10.1080/21681376.2026.2652760

To link to this article: https://doi.org/10.1080/21681376.2026.2652760

Published online: 29 Apr 2026.

REGIONAL STUDIES, REGIONAL SCIENCE
2026, VOL. 13, NO. 1, 2652760
https://doi.org/10.1080/21681376.2026.2652760

RSA Regional Studies Association
Routledge Taylor & Francis Group

OPEN ACCESS Check for updates

Measuring organised crime in Italy: dual composite indicators integrating text mining and official statistics

Antonio Fabio Forgione, Carlo Migliardo and Paolo Mustica

Department of Economics, University of Messina, Messina, Italy

ABSTRACT
This study introduces a dual, data-driven framework for mapping organised crime in Italy by constructing and cross-validating two complementary, per-capita indices aggregated through a non-compensatory composite methodology. The municipal index mines more than a decade of ANSA (Agenzia Nazionale Stampa Associata) news (2012–2023) using Natural Language Processing to extract and geolocate mafia-related events at the municipality level, and complements these signals with computer-vision/optical character recognition (OCR) extraction of clan presence from DIA (Direzione Investigativa Antimafia) investigative maps. The provincial index relies on official records of mafia-related offenses, the dissolution of local administrations due to mafia infiltration and the number of clans, all normalised by population. Across provinces, the composite indices align strongly (a correlation ≈ 0.75). Moreover, for the two offenses unambiguously linked to mafia activity (Type I offenses), news-based counts correlate closely with official records (0.92 for mafia association and 0.88 for mafia-type murders/attempted murders), supporting the use of media narratives as a proxy for underreported activity.

ARTICLE HISTORY
Received 18 July 2025
Accepted 21 March 2026

KEYWORDS
Organised crime; composite indicators; hyperlocal crime patterns; text mining; municipal crime data

JEL
K42; C43; C55; C81

1. Introduction

Organised crime, particularly mafia-style activities, poses unique challenges for researchers and policymakers because of its complex and often covert nature. Official crime statistics, typically aggregated at the provincial level or higher, may struggle to fully capture the relevance of mafia activities, which can be significantly heterogeneous across municipalities within the same province. Indeed, in many countries, data on mafia-related offenses are not available at the municipal level, limiting their ability to detect hyperlocal patterns of criminal activity. In Italy, a significant breakdown of this barrier is represented by the contribution of Dugato et al. (2020), who pioneered the measurement of mafia presence at the municipal level using distinct administrative datasets. While their work constitutes a fundamental benchmark, reliance on administrative records inherently entails challenges for continuous, fine-grained monitoring. Moreover, official statistics often aggregate mafia-related offenses (e.g., specific types of intimidation) within generic categories, making it difficult to isolate the share of crimes explicitly attributable to criminal organisations. Our study builds upon this path by proposing a complementary approach based on unstructured data. The challenges are exacerbated by the widespread issue of underreporting and the strategic resilience of mafia networks. The fear of reprisals and the clandestine nature of many operations result in a substantial discrepancy between actual criminal activity and what is documented by authorities. Furthermore, as demonstrated by studies on network disruption (Cavallaro et al., 2020), law enforcement actions often capture only the disrupted segment of these adaptable organisations. Consequently, administrative records may reflect investigative outcomes rather than the full underlying presence and activity of criminal clans. This disparity between official statistics and the on-the-ground reality necessitates the exploration of complementary data sources, such as news media, which can capture the publicly visible and signalling aspects of mafia activity.

CONTACT Carlo Migliardo cmigliardo@unime.it
Supplemental data for this article can be accessed online at https://doi.org/10.1080/21681376.2026.2652760

To address these pitfalls, this study introduces a dual data-driven framework for mapping organised crime in Italy by constructing and cross-validating two complementary indices designed to overcome the specific shortcomings of standard data analysis. The first index operates at the municipal level, leveraging advanced text mining techniques to analyse over a decade of news reports from ANSA (Agenzia Nazionale Stampa Associata), Italy's leading news agency. ANSA′s extensive and reliable coverage offers a unique opportunity to extract detailed data on mafia-related crimes across Italy's smallest administrative divisions, a level of granularity not publicly available in official statistics. By employing Natural Language Processing (NLP) methods, we transformed unstructured text into structured data and extracted references to locations, crime types and contextual details. This allowed us to identify criminal acts explicitly linked to mafia-style organised crime, as reported by journalists, providing a clearer and more focused picture than official statistics alone. Specifically, the reliance on news-based data offers the qualitative advantage of partially overcoming the under-reporting that plagues official statistics, particularly for mafia-related crimes where the victims silence often leads law enforcement to miss or group intimidatory acts (e.g., arson or attacks) together with generic crimes. Since mafia violence often has a signalling purpose, intended by clans to be public and visible to enforce territorial control, news agencies act as a sounding board for these signals, often providing the necessary context (e.g., linking an attack to a turf war or racketeering). Thus, NLP allows us to uncover hyperlocal outbreaks of organised crime that often go undetected in coarser datasets, enabling the identification of specific municipalities as hotspots for targeted intervention. To enhance the granularity and reliability of the municipal-level index, we incorporated computer vision (CV) techniques applied to investigative reports published semi-annually by Italy's Anti-Mafia Investigation Directorate (DIA – Direzione Investigativa Antimafia). These reports include maps indicating the surnames of mafia families present within municipalities, although the data are embedded as unstructured visual elements in the reports. By employing CV for image segmentation and Tesseract optical character recognition (OCR) for text extraction, we automatically extracted the number of clans present in each municipality from the maps.

The second index, calculated at the provincial level, complements the municipal analysis by drawing on official records of mafia-related offenses, such as damage by fire, extortion, money laundering, mafia-style homicide, mafia association, attacks and attempted mafia-style homicide, as well as the total number of local administrations dissolved due to mafia infiltration and the number of clans. Each sub-indicator was adjusted for the provincial population. This indicator, in line with previous studies adopting composite indicators (e.g., Bernardo et al., 2021; Calderoni, 2011; Dugato et al., 2020; Forgione & Migliardo, 2025), synthesises registered offenses directly tied to organised crime, alongside data on dissolved local governments, into a normalised benchmark that accounts for population differences. This provincial index provides a standardised metric for comparing the intensity of organised crime across local areas and highlights the mafias critical presence.

To aggregate the raw indicators into both the comprehensive municipal and provincial indices of mafia presence, as detailed in Section 3.3, we use an advanced technique that is robust against common statistical issues arising in averaging multiple sub-indicators.

In summary, the use of two indices is motivated by their complementary strengths: the municipal-level index offers unparalleled granularity for pinpointing localised criminal activity, while the provincial-level index provides a normalised, comparative perspective. These two elements contribute to a comprehensive understanding of organised crime, addressing the necessity for precise hotspot detection and the requirement for broader strategic insights. Cross-validation shows a strong alignment between the two composite measures (provincial correlation ≈ 0.75). Furthermore, for the two offenses inherently associated with mafia activity, the ANSA based counts exhibit a strong correlation with official records, with a coefficient of 0.92 for mafia-type criminal association and 0.88 for mafia-type murders or attempted murders; see Section 4.1, supporting the use of media narratives as a proxy for underreported activity. This not only underscores the efficacy of text mining as a research tool but also highlights its potential to enhance our comprehension of the extent and spread of organised crime.

This study presents several novel aspects. The municipal mafia index is a composite measure derived from panel data that is mainly based on textual analysis.[1] Second, the provincial mafia index, which is also structured as a panel, represents the first indicator based on objective data aggregated through an innovative technique that is resilient to several limitations inherent in standard methodologies prevalent in the literature.

From a regional development perspective, organised crime significantly impacts local developmental trajectories by undermining institutional integrity. This deterioration is evidenced by a shift away from low levels of corruption, impartial public services and strict adherence to the rule of law, and is closely associated with varying socio-economic outcomes (Charron et al., 2014; Rodríguez-Pose, 2013). Within this context, organised crime functions as a perverse institution that distorts market mechanisms.

Institutional quality plays a crucial role in determining the effectiveness and returns of place-based policies and public investments, such as the EU Cohesion Policy. This influences project selection, implementation capacity and the ability to transform resources into growth (Crescenzi et al., 2016; Mendez & Bachtler, 2024; Rodríguez-Pose & Garcilazo, 2018). Empirical studies have validated these mechanisms. For example, Pinotti (2015) showed that the mafia leads to a structural gross domestic product (GDP) loss of approximately 16% in affected regions, mainly by replacing productive private capital with inefficient public investment. Furthermore, the presence of a mafia significantly deters foreign direct investment (Daniele & Marani, 2011) and undermines the administrative capacity to efficiently utilise EU funds, resulting in execution delays that negate the potential benefits of public spending (Arbolino & Boffardi, 2024).

In this context, the granularity of measurement assumes strategic importance. Standard governance indicators frequently lack the spatial resolution necessary to capture the localised nature of criminal influence. Within this framework, this study provides detailed indices that offer an objective perspective on the local institutional environment. Indeed, these metrics complement perception-based indicators that define the institutional traps impeding the escape from the poverty trap in underdeveloped areas of Italy (Charron et al., 2014).

The remainder of this paper is organised as follows: Section 2 offers a comprehensive literature review, examining traditional measures of organised crime (Section 2.1), methodologies for crime mapping through text mining (Section 2.2) and strategies for integrating textual data into composite indicators (Section 2.3). Section 3 outlines the data sources and methodological framework utilised, including the extraction of crime events from ANSA news archives via NLP (Section 3.1), the application of computer vision to analyse DIA reports (Section 3.2) and the employment of the Robust Multi-directional Benefit of the Doubt (RMDirBoD) technique to aggregate heterogeneous data into composite indices (Section 3.3). Section 4 presents the empirical findings and discussion, starting with a comparison of textual data against official records to evaluate reliability (Section 4.1), followed by the introduction of the Municipal Organised Crime index (Section 4.2) and concluding with the Provincial Organised Crime index (Section 4.3). Section 5 conducts robustness checks to validate the stability and reliability of the proposed composite indicators across alternative methodological specifications. Finally, Section 6 summarises the key findings, discusses their policy implications and outlines directions for future research.

2. Literature review

2.1. Traditional measures of organised crime

Several studies have devised indicators of organised crime relevance, relying on individual data sources. Some authors have proxied mafia presence by focusing on specific measures indirectly related to criminal organisations (i.e., murder, theft, robbery and fraud) (e.g., Alesina et al., 2019; Marselli & Vannini, 1997), while others have used direct measures of mafia-style involvement, such as the incidence of individuals implicated in mafia-type organisations per 100,000 people (Solivetti, 2016) and municipalities dissolved due to mafia infiltration (Coniglio et al., 2010).

Others have aggregated multiple crime-related variables to construct composite indicators reflecting the presence of criminal organisations. For instance, Van Dijk (2007) devised a composite organised crime index to assess organised crime levels across 113 countries. This index consolidates standardised scores derived from five indicators: the perceived prevalence of organised crime, rates of unsolved homicides, grand corruption, money laundering and the size of the illegal economy. Calderoni (2011) introduced a mafia index to quantify the presence of mafia organisations across various Italian provinces. The author initially identified the most pertinent indicators that signify mafia presence and subsequently selected relevant metrics, which included data on mafia-type associations, mafia-related homicides, the dissolution of city councils due to organised crime infiltration, and assets confiscated from organised crime entities.

Daniele and Marani (2011) investigated the influence of crime on foreign direct investment inflows across Italian provinces, developing an organised crime index derived from data on extortion, bomb attacks, arson and criminal associations. Similarly, Ganau and Rodríguez-Pose (2018) examined whether organised crime affects firms performance directly and indirectly by diminishing the positive externalities generated by the geographic concentration of market-related firms. In particular, the presence of organised crime was measured using a composite indicator that includes three crime variables (mafia-type associations, mafia-related murders and extortion). Dugato et al. (2020) introduced an additional indicator designed to assess the prevalence of mafia organisations within Italian municipalities.[2] It comprises five components: reported mafia associations, reported mafia murders/attempted mafia murders, active mafia groups mentioned in official reports, city councils dissolved due to mafia infiltration, and assets confiscated from the organised crime. Bernardo et al. (2021) developed a composite organised crime index for 103 Italian provinces by employing a stochastic-dominance efficiency procedure. This approach involved the endogenous derivation of weights through linear programming under dominance constraints for ten elementary indicators, including mafia-type murders, criminal and mafia associations, councils dissolved due to mafia infiltration, assets confiscated, extortion, kidnappings, arson attacks, drug trafficking, money laundering and corruption. Fioroni et al. (2025) explore the dynamics between organised crime, corruption and economic growth. They build a composite measure using factor analysis, combining selected mafia-related offenses with information on confiscated assets. Beyond their criminological significance, these measures directly address a fundamental debate on regional development regarding the influence of institutions and governance on regional disparities. The literature on regional studies documents substantial within-country variation in government quality and its strong association with the development level (Charron et al., 2014; Rodríguez-Pose, 2013). Empirical evidence consistently demonstrates that higher institutional quality enhances returns on public investment and mediates the effectiveness of EU Cohesion Policy expenditures (Crescenzi et al., 2016; Mendez & Bachtler, 2024; Rodríguez-Pose & Garcilazo, 2018). Institutional quality can also affect regional integration and market access, for example, by shaping interregional trade patterns (Barbero et al., 2021). We position our municipal and provincial indices as complementary measures of the local institutional environment, capturing one specific but economically significant dimension: the territorial embeddedness and visibility of organised crime.

As this overview shows, existing composite indicators of organised crime have relied exclusively on traditional data sources, such as official crime statistics, reports and surveys. However, integrating unstructured sources, such as textual data, can be particularly useful in cases where traditional structured data are unavailable or inaccessible to researchers. The next section reviews studies that have employed text mining techniques to map crimes.

2.2. Mapping crimes through text mining

Research employing text mining methodologies for crime mapping is a relatively recent development, predominantly situated within the domain of computer science literature. Arulanandam et al. (2014) proposed a systematic methodology for extracting crime-related information from online newspaper articles. They compiled a sample of articles on theft from four English-language newspapers across three countries: Australia, India and New Zealand. By integrating named entity recognition (NER) algorithms with conditional random field models, the authors successfully identified the location of theft in each article. Their approach demonstrated an accuracy rate ranging from 73 percent to 90 percent, contingent upon the newspaper analysed. Jayaweera et al. (2015) introduced a web-based crime analysis system aimed at identifying the locations of various crime types over time. To achieve this, the study utilised three English-language newspapers published in Sri Lanka as sources of crime incident details. The articles were analysed using entity extraction and classification techniques. The findings indicate that information regarding the date, location, and type of crime can be extracted with accuracies ranging from 74 percent to 82 percent. Hassan and Rahman (2017) implemented a system for the analysis of crime-related news from online newspapers, employing various data mining techniques. The authors integrated NER methods with Support Vector Machine (SVM) algorithms to extract crime locations from articles published in two English-language newspapers in Bangladesh. The results indicated that the system achieved an accuracy of approximately 72 percent. Bonisoli et al. (2021) proposed a methodology for the categorisation of crime in Italian press articles.

They compiled a dataset comprising crime-related articles from two local newspapers in the Italian province of Modena, Modena Today and Gazzetta di Modena. Subsequently, they employed various algorithms to ascertain the crime category of each article. Their findings indicated that algorithms based on SVM exhibited superior performance in supervised text categorisation, whereas spectral and agglomerative clustering were most effective for unsupervised categorisation. Das et al. (2021) developed an advanced classifier for the analysis of crime reports, which not only forecasts crime categories but also self-updates with incoming crime data. Their study utilised English-language newspapers from India, the UAE and the USA, alongside narrative reports from the police and witnesses. They applied a bi-objective particle swarm optimisation method to this textual dataset to create a robust incremental classifier capable of dynamically categorising and predicting crime reports. The results of their study showed that this method generally outperformed the other classifiers. In their study, Boukabous and Azizi (2022) introduced a multimodal approach to crime prediction by integrating audio and textual data. They utilised a Convolutional Neural Network (CNN) and Bidirectional Encoder Representations from Transformers (BERT) models to analyse the audio and textual components extracted from the XD Violence video dataset. The results from these analyses were then synthesised into a unified decision vector, which effectively classified each observation as either associated with a crime or not. This integrated model demonstrated enhanced performance compared to the individual models when used separately.[3]

Text mining techniques offer substantial potential for the detection of organised crime through innovative crime indices. This potential is underscored by their demonstrated usefulness in social sciences, as evidenced by recent studies discussed in the following section.

2.3 Integrating textual data into indicators

The potential of unstructured texts has been widely exploited in economics and sociology for developing and analysing indicators. For instance, Park and Kremer (2017) introduced a methodology for categorising environmental sustainability indicators using text mining. They compiled 55 indicators and employed a topic model to systematically categorise them based on the conceptual similarities in their textual descriptions. Upon comparing this categorisation with the traditional top-down approach, the authors determined that their method could serve as a complementary tool, as it allows for the objective integration of conceptual similarities between indicators into their categorisation. A comparable methodology was suggested by Jacinto et al. (2020), who integrated text mining applied to scientific literature with expert opinion analysis to establish a database of social resilience indicators for floods, organised by dimensions. Textual data were employed to develop innovative indicators, as in Freunek and Niggli (2023), who established a framework for constructing the DynaPTI. This innovative patent technology indicator was developed to assess technological quality and activity using text mining. It employs textual information from patents as input to a BERT model to identify patents with the highest degree of technological similarity. In a similar vein, Salvatore et al. (2024) devised a methodology that integrates both conventional and textual data to create a composite indicator. This indicator functions as an innovative metric for assessing a company's dedication to corporate social responsibility and the effectiveness of its communication practices.

Furthermore, the use of unstructured text is commonly associated with its application, often as sentiment indicators, to refine macroeconomic and financial predictions. The study conducted by Feuerriegel and Gordon (2018) employed text mining techniques on regulatory disclosures to assess the long-term forecasts of stock indices. The findings indicate that text-based models have the potential to surpass baseline predictions derived from autoregressive models. Aprigliano et al. (2023) introduced two high-frequency text-based indicators, the Text-based Economic Sentiment Index (TESI) and Text-based Economic Policy Uncertainty (TEPU), to forecast Italian economic activity. These indices were constructed through a computational analysis of over 1.5 million articles from four prominent Italian newspapers, employing a novel domain-specific dictionary enriched with valence-shifting terms to better capture contextual polarity in economic narratives. The authors integrated these indicators into a Bayesian model averaging framework and a weekly GDP tracking model, demonstrating that their inclusion significantly reduced forecast uncertainty, particularly during recessions and improved the accuracy of key macroeconomic aggregates, such as GDP growth, household consumption and fixed investments, compared to traditional baseline models. The study further validates the relevance of newspaper-derived data through survey evidence, showing that Italian firms and households prioritise economic news from mainstream newspapers when making financial decisions, thereby reinforcing the informational value of textual indicators for real-time economic monitoring.

In conclusion, while traditional metrics for assessing organised crime mainly depend on structured data, unconventional data sources, such as textual data, offer potential for crime mapping applications. Notably, the limited availability of structured data positions textual sources as valuable alternatives, facilitating the development of hybrid composite indicators that integrate structured and text data.

3. Data and methods

As delineated in the preceding section, our dual comprehensive measure of organised crime is disaggregated at both the municipal and provincial levels and adjusted for the population. The municipal-level mafia index integrates data derived from text mining methodologies applied to ANSA news archives to identify crime-related narratives, computer vision analysis of investigative maps from Italy's DIA, and publicly accessible data on municipalities dissolved due to mafia infiltration. In contrast, the provincial mafia index synthesises structured data on mafia-related offenses, as provided by the Italian Ministry of the Interior, and presence (municipality dissolved and number of clans). This section introduces the data and methods used to construct composite indicators. In particular, Section 3.1 reports the NLP-driven extraction of crime events from ANSA news. Section 3.2 describes the computer vision workflow for DIA map analysis and Section 3.3 introduces the robust multi-directional Benefit of the Doubt (RMDirBoD) technique to aggregate these heterogeneous inputs into composite indices.

3.1. ANSA news

We gathered press news on organised crime from the Nexis Uni text archive (LexisNexis, 2024), which contains textual data from over 35 thousand sources, including news and legal publications from different countries. In particular, we considered Italian news on organised crime in Italy, published between 2012 and 2023. We used ANSA (Agenzia Nazionale Stampa Associata), the most important news agency in Italy, as a source of information. We considered only this source of information because ANSA has widespread diffusion throughout Italy, thanks to the presence of at least one office in each Italian region. This ensures homogeneous journalistic coverage of the country. From a network analysis perspective, comprehensive capillary coverage is essential. Drawing on Ficara et al. (2021), we emphasise the importance of minimising node removal, which refers to the missing data from peripheral municipalities in our study. This is due to its greater potential to distort the network structure compared to edge removal, such as occasional missed reports resulting from NLP filtering. While edge removal exerts a negligible impact on the network structure, the absence of nodes can lead to significant misinterpretations. Consequently, selecting a source with extensive national coverage, such as the ANSA, enables us to address this issue proactively. However, potential omissions in marginal areas remain an inherent limitation of press-based indicators.

Moreover, ANSA provides information that is often regarded as highly reliable, thus enabling accurate data extraction. Finally, considering a single source of information allows us to eliminate the risk of including the same news reported by others, as well as the bias of reporting news generally linked to a certain area by local newspapers or just news of interest to the entire country by national newspapers. From a practical perspective, most of the Italian news contained in Nexis Uni comes from ANSA, implying that a mapping of Italian news cannot ignore this agency.

To collect news on organised crimes, we defined a search query in the Nexis Uni search engine. Specifically, we considered news with at least one keyword associated with crimes relevant to the literature on organised crime (arson incidents, criminal attacks, extortion, mafia-style murders/attempted murders, membership in mafia-type criminal association, money laundering and usury) (Buscaglia & Van Dijk, 2003; Calderoni, 2011; Daniele & Marani, 2011). Corruption was excluded to maintain the analytical clarity. Although corruption is strictly correlated with organised crime (Fioroni et al., 2025), treating them as distinct phenomena allows for the investigation of their interaction, as the former is often an outcome of the latter. Adopting the framework proposed by Battisti et al. (2020), we distinguish between direct organised crime measures (Type I: mafia-style murders/attempted murders and membership in mafia-type criminal association) and indirect organised crime measures (Type II: arson incidents, criminal attacks, extortion, money laundering and usury). Because Type II offenses are not uniquely attributable to criminal syndicates in administrative data, using them without further information would introduce misclassifications. Therefore, we retained Type II events only when the ANSA/Nexis metadata and the article narrative explicitly framed them as organised-crime-related, using the Nexis Uni subject filter Organised Crime with relevance ≥25%. This reduces the misclassification bias that would arise if generic recorded crime counts were mechanically interpreted as mafia-related activity. This is particularly relevant for usury, which enters the municipal index only in its mafia-linked occurrences identified in the news; conversely, since official statistics lack such semantic filtering, usury was excluded entirely from the provincial reported-offense index to preserve reliability.

In our analysis of the news data, we extracted the date, title and content of the articles based on the labels provided by the Nexis Uni database. However, accurately identifying the location of each news event requires a more sophisticated approach. This complexity stems from the fact that the location reported by Nexis Uni often does not match the actual crime site. Typically, the location label refers to a broader provincial area, omitting the specific municipality where the incident took place, or it may incorrectly identify the capital of Italy or the relevant Italian region. Additionally, it is crucial to recognise that the label specifies only one location, thereby excluding other areas affected by a news event. Consequently, the true location(s) of a crime must be deduced from the content of the article. To achieve this objective, we employ a two-stage strategy to enhance the quality of crime localisation in Italian municipalities, as recommended in the literature (Arulanandam et al., 2014; Hassan & Rahman, 2017; Jayaweera et al., 2015). In the initial stage, we implemented an NER model to automatically identify locations within the text. Subsequently, in the second stage, we classified each sentence containing keywords related to locations and crimes as indicative of the crime location.

Upon closer examination of the methodology, the initial phase involved the utilisation of spaCy, a Python library capable of executing various tasks in Natural Language Processing, including NER.[4] A notable advantage of spaCy is its provision of pretrained models for conducting NER in multiple languages, including Italian. We applied the model to the collected news articles on organised crime, thereby extracting the locations mentioned in the text. Focusing on Italian municipalities, we filtered the extracted locations using the list provided by the Italian National Institute of Statistics (ISTAT).

In the second stage, we segmented the news articles into individual sentences, focusing on those containing keywords pertinent to municipalities and criminal activities. Although we employed the Python library Natural Language Toolkit to ensure that sentence segmentation adhered to the syntactic rules of the Italian language, we also developed regular expressions to filter sentences based on specified keywords. During the second stage, we conducted a binary classification process by manually labelling each sentence as indicative of the crime location. Specifically, a sentence was classified as a crime location if it referenced municipalities where the crime occurred and/or where an arrest or seizure occurred. To further enhance data quality, we manually refined the extracted NER tags within these sentences by removing tags for locations not associated with crime sites and adding tags for crime locations initially overlooked by the NER model. To quantitatively evaluate the reliability of our methodology, we employed a representative sample of 368 sentences for dual validation. Initially, we assessed the performance of the original NER extraction by calculating precision and recall. Specifically, we compared the tags originally extracted by the NER model with the manually refined ones, resulting in a precision of 0.871 and a recall of 0.953, indicating satisfactory performance of the NER model. Subsequently, an independent operator labelled the same sentences as crime locations to evaluate the robustness of the classification process. The inter-rater agreement, measured using Cohens Kappa, was 0.855, indicating strong agreement between the annotators.

In summary, while the automated pipeline adopted in the first stage provided the core of our dataset, the human-in-the-loop strategy implemented in the second stage enabled the extraction of organised crime information at the municipal level with the highest possible accuracy by manually validating all matches to exclude false positives and recovering relevant events missed by automated filtering. Figure 1 presents a flowchart that encapsulates the two-stage strategy delineated so far.

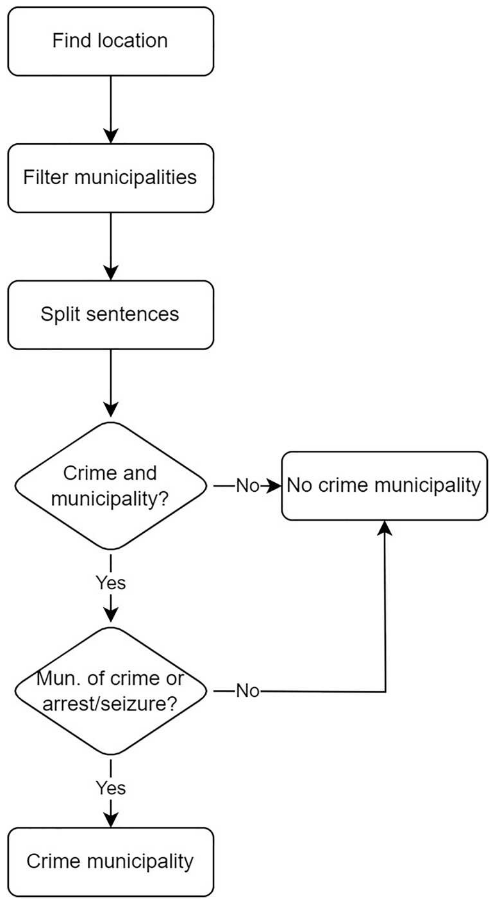

Figure 1. Flowchart to detect the location of the crime in the news.

The final dataset includes 2528 observations, each of which is associated with a specific municipality.[5] We then created dummy variables that captured whether the news referred to the i-th mafia-type crime.

$$\text{Crime}_{\,i} = \begin{cases} 1 & \text{if the news includes keywords on the crime } i \\ 0 & \text{otherwise} \end{cases}$$

with i = {*arson incidents, criminal attacks, extortion, mafia-style murders/attempted murders, membership in mafia-type criminal association, money laundering and usury*}. Table 1 details the distribution of the analysed crimes within our dataset.

It is important to note that crime dummies are not mutually exclusive: a single news item frequently reports multiple crimes (e.g., an arrest for both extortion and usury). Consequently, the sum of crime mentions detected (4487) exceeds the total number of news items (2528), reflecting the fact that mafia-related news often describe complex criminal scenarios rather than isolated offenses.

Table 1. Frequency of crime mentions in the analysed news.

Crime	Frequency	% of total observations ($N = 2528$)
Extortion	1478	58.5%
Membership in mafia-type criminal association	940	37.2%
Mafia-type murders/attempted murders	682	27.0%
Money laundering	530	21.0%
Usury	487	19.3%
Attacks	321	12.7%
Arson	49	1.9%

Note: The crime dummies are not mutually exclusive. A single observation (news item) may be tagged with multiple crimes. Therefore, the sum of crime frequencies (4487) exceeds the number of observations (2528).

To construct the panel dataset, we subsequently aggregated the micro data at the municipality-year level. Specifically, for each municipality and year, we calculated the per-capita intensity of each crime type by summing the corresponding dummy variables across all attributed news articles and dividing by the resident population.

3.2. DIA reports

Another source of unstructured information used in our analysis is reports periodically published by the DIA, the Italian multi-force investigation body under the Ministry of the Interior, whose main task is to fight mafia-related organised crime in Italy. These reports contain detailed information on the results achieved by the DIA during its activity. The DIA reports include maps delineating the municipalities where mafia groups are active. Given that our inputs were images (maps) containing textual information for extraction, we employed CV models to automatically process these data. First, we employed the Python library OpenCV to segment the maps into white boxes that contained the names of the municipalities where the mafia groups operated. Given that these white boxes were static images, we utilised Googles Tesseract optical character recognition engine to extract the textual information contained within them. In conclusion, typographical errors in the extracted texts were corrected using the Fuzzy-Wuzzy Python library, which specialises in string matching. Figure 2 presents an example of a pipeline that facilitates the extraction of municipalities from maps where mafia groups are known to operate.

Similarly, we applied this approach to extract the names of clans, which typically appear in uppercase, as shown in Figure 2 from the white boxes. The total number of clans for each municipality was determined by summing the number of names identified across the various white boxes associated with that municipality.

To validate the variables derived from the DIA maps, we constructed a ground truth dataset by manually annotating 198 randomly sampled white boxes (5% of the total) with the correct clan counts and municipality lists. Subsequently, we calculated the mean absolute error (MAE) for the clan counts and the precision and recall for the municipality lists, resulting in an MAE of 0.040, a precision of 0.761 and a recall of 0.884. The low MAE value indicates a high degree of accuracy in the number of clans extracted from each white box, implying an average error of only 0.04 units per map.

The precision score indicates that approximately 24% of the extracted municipality names were false positives (or noise), while the recall value suggests that about 12% of the municipalities contained in the white boxes were not detected by the algorithm (false negatives). These results appear satisfactory, as

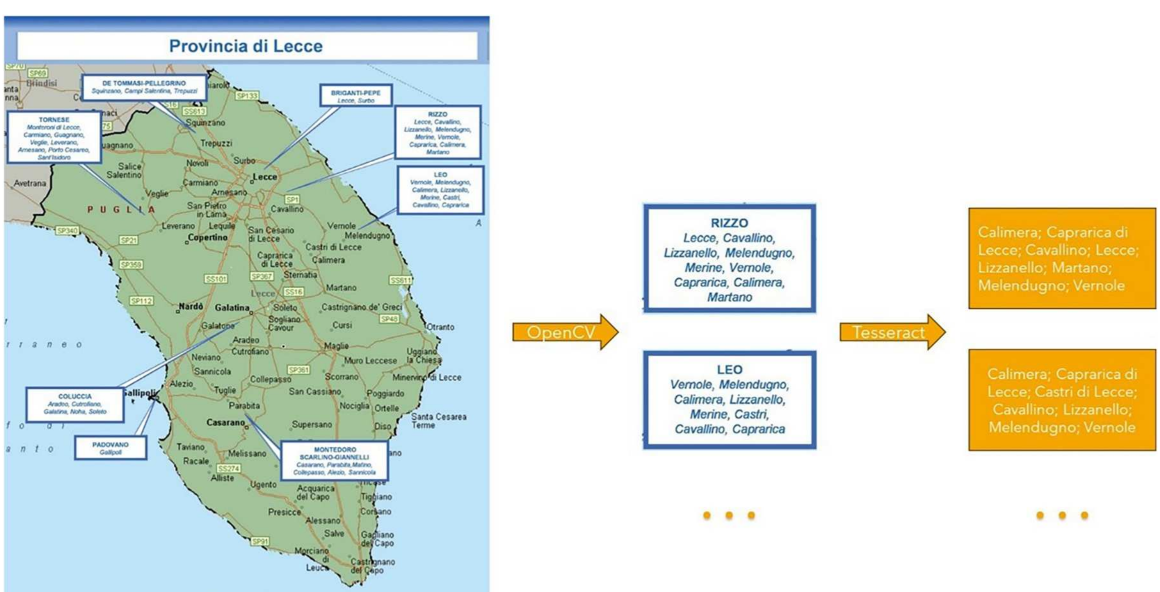

Figure 2. Example of the pipeline for extracting municipalities from maps. Source: Direzione Investigativa Antimafia (2012–2023).

false positives generated by the algorithm are quite uniformly distributed across Italian municipalities (97% of these erroneous detections appearing with a frequency of three or fewer), suggesting that municipalities are not subject to systematic bias. On the other hand, false negatives, which are of particular concern as they represent missing nodes that may bias the results according to Ficara et al. (2021), are limited to around 10%.

3.3. Robust multi-directional benefit of doubt technique

To aggregate heterogeneous sub-indicators into a single measure of organised crime at both the municipal and provincial levels, we employ the Robust Multi-directional Benefit of the Doubt (RMDirBoD) framework (Vidoli et al., 2024). This choice represents a methodological evolution designed to overcome the limitations of standard linear indices and deterministic non-parametric models, specifically addressing data challenges such as fragmentation and high variability (Union & Centre, 2008). This approach offers three hierarchical improvements.

First, regarding the weighting scheme, the method avoids the rigidity of fixed-weight aggregation. Consistent with the BoD philosophy, it endogenously assigns data-driven weights to maximise the score of each unit. This ensures that the Composite Indicator respects the specific data structure of each territory, avoiding the arbitrariness of equal or expert-based weighting.

Second, regarding the benchmarking process, the framework improves upon conventional BoD models. While standard approaches often impose a uniform direction of improvement-potentially introducing bias-the *multidirectional* structure calculates a unit-specific trajectory (Fusco, 2023). This flexibility is crucial for ensuring that territories with distinct criminal profiles (e.g., coercive violence vs. illicit markets) are evaluated against realistic, tailored targets rather than a generic standard.

Third, regarding aggregation properties, the model mitigates the *full compensability* and outlier sensitivity common to both linear indices and deterministic frontiers. In standard approaches, excellent performance in a single dimension can fully offset the severe deficits in others. The RMDirBoD addresses this by estimating the efficiency of repeated subsamples of size $M < N$ drawn with replacement. This resampling strategy reduces the dominant influence of extreme values, ensuring that the final indicator reflects a balanced assessment rather than one driven by isolated anomalies.

Formally, let y_{iqt} denote the normalised value of indicator $q = 1, \ldots, Q$ for unit $i = 1, \ldots, N$ in year t. Specifically, normalisation was performed separately for each year to account for temporal shifts in the data distribution. We estimate the composite indicator on a balanced panel by computing year-specific frontiers to account for temporal dynamics. The robust score is obtained through a Monte Carlo approximation based on B replications. In each replication, a reference set of size $M < N$ is drawn with replacement, and the units efficiency is benchmarked against this partial frontier.

The final RMDirBoD composite indicator corresponds to the expected value of the scores across replications and can be expressed as:

$$CI_{it}^{\mathrm{RMDirBoD}} = 1 - \frac{\beta_{it}^{*} \sum_{q=1}^{Q} g_{PI,it,q}}{\sum_{q=1}^{Q} (y_{iqt} + \beta_{it}^{*} g_{PI,it,q})} \tag{1}$$

where $g_{PI,it,q}$ denotes the unit-specific potential improvement for indicator q, and β_{it}^{*} is the associated scaling factor derived from the robust resampling scheme (for analytical details, see Vidoli et al., 2024). The resulting score ranges between 0 and 1, with higher values indicating a greater relative intensity of organised crime.

We estimate the RMDirBoD composite indicator maintaining a consistent set of sub-indicators and fixed tuning parameters across years. Robustness was achieved through repeated subsampling of the reference set, with the subsample size set to $M = 1500$. This value was selected using the standard diagnostic method based on the percentage of outperforming units, which became negligible and stabilised at this value (see Figure A.1 in the Appendix in the online supplemental data). We set the number of replications to $B = 1000$ to ensure a stable Monte Carlo approximation of the robust scores. As additional sensitivity checks, we also computed the Multi-directional Benefit of the Doubt approach (Fusco, 2023), that is a non-robust version of the index, without the robust subsampling step, and obtained very similar results, with very high correlations with the RMDirBoD-based indicator. Similarly, re-estimating RMDirBoD using both lower and higher replication counts yields very high correlations with the baseline RMDirBoD specification, indicating that the results are not driven by the Monte Carlo design. The provincial index is computed analogously year-by-year from official data, applying the same normalisation and RMDirBoD aggregation, with $M = 32$ and $B = 300$ given the smaller cross-sectional sample size.

The RMDirBoD model presents an advanced, data-driven methodology for constructing a CI for organised crime, effectively balancing flexibility, robustness and empirical precision. By integrating the multidirectional perspective of the MDirBoD with robust optimisation techniques, this approach offers a dependable tool for assessing crime prevalence and informing policy interventions at both the municipal and provincial levels in Italy.

4. Results and discussion

4.1. Comparing textual data with official data

Before proceeding with the estimation of the composite indicator, we evaluated the reliability of our textual indices concerning the crimes under investigation by comparing them with the official statistics from the Italian Ministry of the Interior. In particular, we plotted the frequency of recorded crimes for a specific offense against the number of news articles on crimes extracted from the textual dataset. Due to the lack of official municipal data, the data were compiled at the provincial level. Figure 3 illustrates the scatter plots for two crimes that are definitively linked to mafia-type offenses: membership in mafia-type criminal association and mafia-type murders/attempted murders.

Despite the inherent limitations of comparing our data at a more aggregated level, we identified a distinct linear correlation between the two crimes under investigation. Specifically, the linear correlation coefficients for membership in mafia-type criminal association and mafia-type murders or attempted murders were 92% and 88%, respectively. This finding suggests that the frequency of news reports on crime serves as a reliable proxy for the number of recorded crimes. This strong correlation can be attributed to at least two factors. First, the official data we analysed pertain to crimes that, by their nature, are necessarily perpetrated by members of criminal organisations and occur at a relatively low frequency compared to other crimes that can be committed by individuals outside such organisations (e.g., money laundering). Second, our textual data, which merely count the number of news articles related to criminal syndicates for a specific

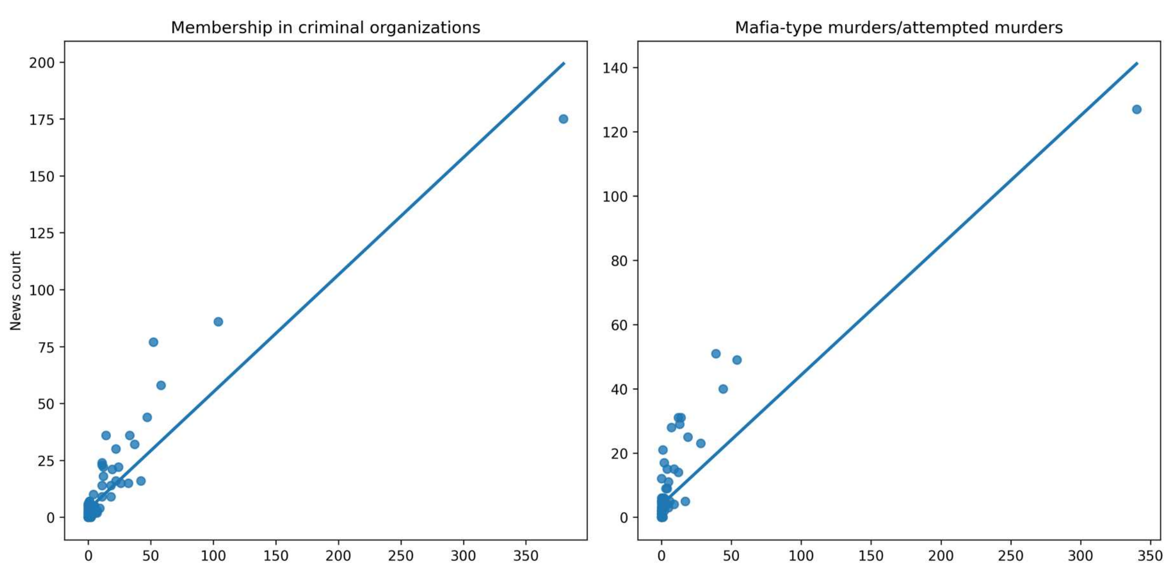

Figure 3. Recorded crimes vs. news articles: membership in mafia-type criminal association compared to mafia-type murders/attempted murders.

Note: All scatter plots include the trend line.

offense rather than the number of offenses, exhibit a pattern similar to that of official data, which refers to a relatively small number of crimes committed exclusively by members of criminal organisations.

A similar pattern is observable in offenses that are not exclusively associated with organised crime syndicates, such as arson, criminal attacks, extortion, money laundering and usury. Figure 4 presents scatter plots comparing textual data with official statistics.

Although the linear correlation coefficients for these crimes were lower than those for mafia-related criminal associations and homicides, they remained within an acceptable range. Excluding arson incidents, which exhibited a correlation of 39%, the remaining correlation coefficients were moderate to strong, ranging from 56% for attacks to 65%, 76% and 77% for extortion, money laundering and usury, respectively. It is important to note that the offenses in question are not exclusively perpetrated by individuals affiliated with organised crime groups. Unlike crimes such as mafia-type criminal associations or mafia-related homicides, which inherently require a demonstrable connection to organised crime for classification, the other offenses analysed here do not necessitate such a link for official recognition. For example, although money laundering is frequently associated with criminal syndicates, it may also be conducted by individuals acting independently of such networks. Another emblematic example is arson, which, on the one hand, is a common crime that can be committed by anyone;[6] on the other hand, arson committed as an intimidatory act tends not to be reported to authorities due to fear of retaliation. In this context, news-based data offer a qualitative advantage, albeit being more selective. While the media may not capture every minor incident (resulting in a lower volume of observations compared to official records), they effectively capture events with a strong signalling purpose. Mafia violence is often designed by clans to be public and visible to enforce territorial control. Even in the absence of formal reports by the victims, news agencies act as a sounding board for these visible signals, often providing the necessary context (e.g., racketeering) that official records might miss or misclassify as common delinquency.

In conclusion, official statistics for non-mafia-related offenses suffer from a dual limitation: they are inflated by generic crimes (noise) while simultaneously missing the specific mafia attribution due to victim silence. Therefore, considering that major news agencies typically follow verification routines and that

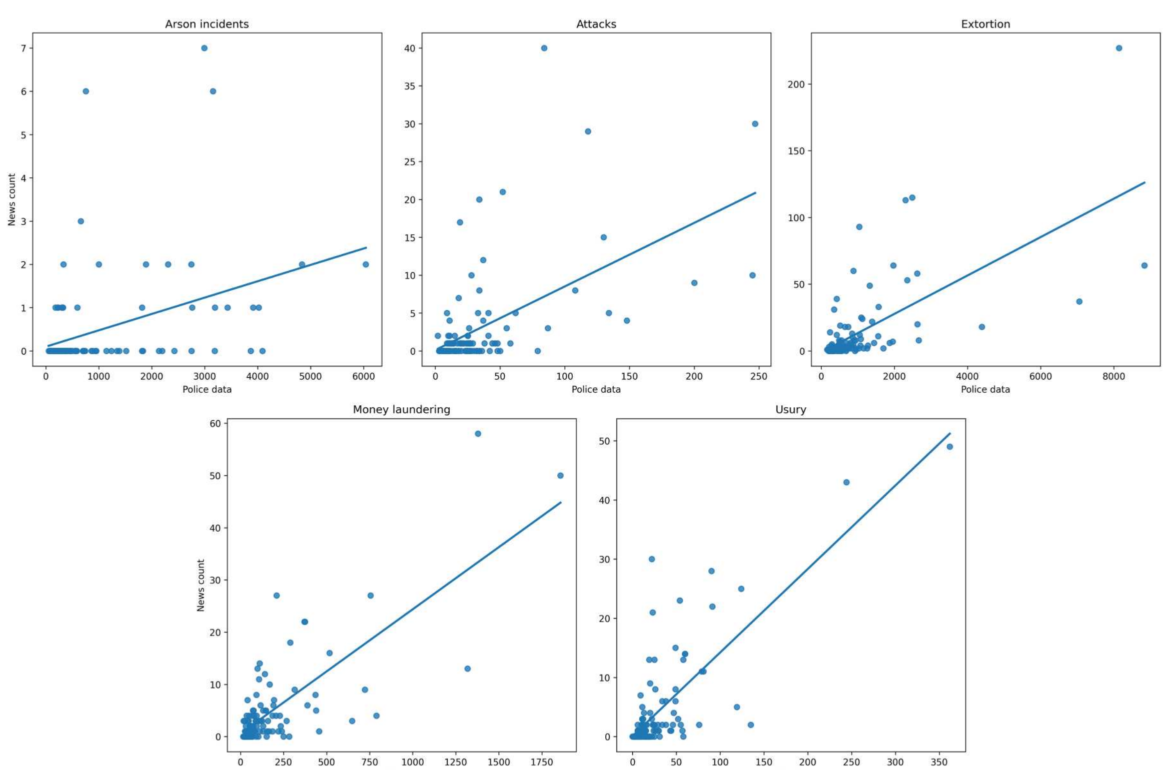

Figure 4. Recorded crimes vs. news articles: arson incidents, attacks, extortion, money laundering and usury.
Note: All scatter plots include the trend line.

media coverage focuses on the visible signals of mafia power, news-based data may be more appropriate for investigating the relevance of organised crime.

4.2. Introducing the municipal organised crime index

This section introduces the composite indicator designed to map the presence of mafias at the municipal level: the Municipal Organised Crime index (MOC). Official data on mafia-related activities often lack municipal-level granularity and are not publicly available. Textual data from news articles provide a valuable alternative, addressing data scarcity and allowing socio-economic analysis to be conducted on the effect of organised crime based on granular micro data. Moreover, extracting textual information from public news sources facilitates the rapid collection of data, in contrast to official data, which are often not released immediately. Finally, textual data may enhance accuracy by enabling the systematic selection of news articles using metadata tags that explicitly associate offenses with organised crime groups. This filtering capability allows researchers to distinguish between crimes perpetrated by individuals unaffiliated with mafia organisations and those directly linked to them.

The MOC index builds upon and complements existing municipal-level methodologies, most notably the pioneering mafia presence index (MPI) as established by Dugato et al. (2020). While MPI utilises administrative and judicial records to capture the judicially relevant manifestations of mafia activity, such as arrests and confiscations, our MOC index employs news data to capture the publicly visible and signalling aspects of such activity, including attacks and turf wars intended as messages. Furthermore, our index broadens the scope of the phenomena considered by incorporating both power syndicate and enterprise syndicate crimes (Type II crime measures), provided these are explicitly associated with mafia groups in news narratives. It also utilises a panel structure and a robust aggregation technique (RMDirBoD) that is well suited for dynamic and fine-grained analysis.

The MOC index integrates primarily unstructured data on offenses that are frequently cited in the literature (e.g., Dugato et al., 2020). In line with Battisti et al. (2020), we have previously differentiated between direct organised crime measures (Type I, e.g., mafia-type murders) and indirect organised crime measures (Type II, e.g., usury). To better characterise the multidimensional nature of mafia activity, we also adopt the conceptual framework proposed by Mocetti and Rizzica (2024). Adhering to their classification, we categorise our sub-indicators into three distinct domains.[7] The first comprises objective indicators of organised crime, capturing the phenomenon directly through variables such as mafia-type murders/attempted murders, membership in mafia-type criminal association, municipalities dissolved due to mafia infiltration, and the number of mafia groups. The second domain pertains to power syndicate crimes, referring to the sphere of territorial control through coercion and violence, measured by arson incidents, attacks and extortion. The third domain covers enterprise syndicate crimes, capturing illicit market activities, specifically money laundering and usury. Table 2 lists the variables used in constructing the composite indicator grouped by domain, together with the sources consulted and the types of data processed.

Adopting the multidomain taxonomy proposed by Mocetti and Rizzica (2024) provides a broader interpretation of the MOC index beyond its technical composite nature. Objective indicators capture the direct manifestations of mafia presence, while power-syndicate crimes (i.e., warning-signal events) and

Table 2. The components of the MOC index categorised by domain.

Variable	Source	Type of data
Objective indicators (Type I)		
Mafia-type murders/attempted murders	ANSA news	Textual data
Membership in mafia-type criminal association	ANSA news	Textual data
Municipalities dissolved due to mafia	Ministry of the Interior	Structured data
Number of mafia groups	DIA reports	Image data
Power syndicate crimes (Type II)		
Arson incidents	ANSA news	Textual data
Attacks	ANSA news	Textual data
Extortion	ANSA news	Textual data
Enterprise syndicate crimes (Type II)		
Money laundering	ANSA news	Textual data
Usury	ANSA news	Textual data

Note: All data refer to the period between 2012 and 2023.

enterprise-syndicate crimes serve as proxies for the coercive and market-oriented channels through which mafias operate locally. This distinction is significant because, as demonstrated by Cavallaro et al. (2020), enforcement actions often reveal only the disrupted segment of resilient criminal networks, whereas media narratives can capture the publicly visible and signalling dimension of coercion that underpins territorial control. Therefore, the municipal-provincial dual strategy should be understood from complementary perspectives institutional detection versus public visibility rather than as a substitutable measure of the same construct.

Four of the offenses used in our MOC index were also used in the construction of Dugato et al.'s (2020) MPI. However, our MOC index introduced five additional Type II crime measures (arson incidents, attacks, extortion, money laundering and usury) that were omitted from the final version of the MPI, partly because of the challenges involved in linking such offenses exclusively to mafia groups. Conversely, our approach, detailed in Section 3.1, enabled us to obtain granular information on these crimes in relation to criminal organisations. In addition, the *RMDirBoD* technique is more effective for the accurate aggregation of raw data than simpler methods previously used for crime data aggregation.

To standardise the indicators, we used the min max normalisation technique. This method rescales each indicator to a common range of [0, 1], ensuring that all variables contribute equally to the analysis without being influenced by their original scales (Cinelli et al., 2021). For a given indicator x, the normalised value x' is calculated as follows:

$$x'_{it} = \frac{x_{it} - \min_t(x)}{\max_t(x) - \min_t(x)} \tag{2}$$

where x_{it} is the value for unit i in year t; $\min_t$ and $\max_t$ computed across units in year t.

Figure 5 shows the geographical distribution of the MOC index:

The map displays the MOC index using the standard deviation classification, as applied by (Dugato et al., 2020) for the MPI. This choice is motivated by the fact that the vast majority of municipalities (about 85% of

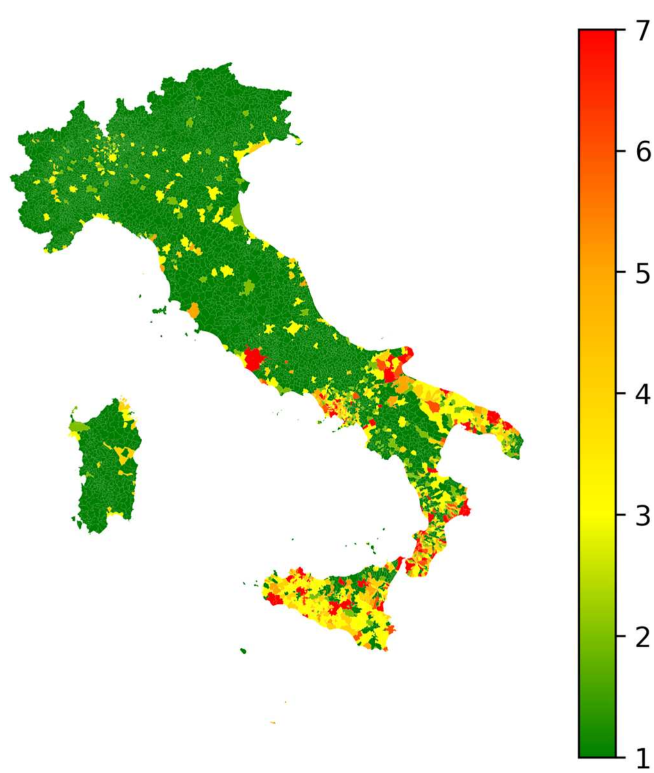

Figure 5. The geographical distribution of the MOC index.

Note: The MOC index map adopts the standard deviation classification as the MPI map presented in (Dugato et al., 2020). Specifically, categories 1–7 correspond to the following intervals: the minimum; between the minimum and the mean; between the mean and the mean plus one standard deviation; between one and two standard deviations above the mean; between two and three standard deviations above the mean; between three and four standard deviations above the mean; and values greater than four standard deviations above the mean.

the roughly 7900 municipalities) take a zero value in our index, implying a highly skewed distribution concentrated at zero. Consequently, the empirical distribution of the index is highly skewed and concentrated at zero. This implies that the remaining municipalities where organised crime is observed exhibit values distributed along the tail of the distribution, essentially representing the outliers. Displaying the MOC index without using classes fails to highlight the presence of rare values. In contrast, the standard deviation classification emphasises outliers, that is, municipalities affected by organised crime, which are the focus of our analysis.

As expected, organised crime is more prevalent in southern Italy, where the country's major criminal organisations originate (Camorra in Campania, Cosa Nostra in Sicily, ′Ndrangheta in Calabria and Sacra Corona Unita in Apulia). More interestingly, the intensity of organised crimes presence in southern municipalities appears remarkably consistent with the Dugato et al.'s (2020) MPI.[8] In other words, the distribution of the MOC index is not random, as it shows that southern Italy is more affected by mafia presence, with some southern municipalities exhibiting higher mafia presence than others, consistent with findings from previous studies. This suggests that, despite the challenges of integrating textual data, we successfully captured a reliable representation of the diffusion of organised crime across Italian municipalities. When properly processed, textual data can serve as a valuable source of information for analysing phenomena when official data do not exist or are not publicly available.

To the best of our knowledge, existing composite indicators of organised crime usually map the structural distribution of the phenomenon, often aggregating data over multi-year periods to provide a consolidated picture of mafia presence. However, considering that criminal infiltration is recognised as a dynamic institutional pathology (Rodríguez-Pose, 2013), analysing the annual evolution of our composite indicator constitutes a significant novel contribution, providing new insights into the long-term patterns and shifts associated with criminal presence. Specifically, this could help address research questions regarding whether the spread of organised crime follows similar trends across geographical areas, despite their underlying structural differences. The time evolution of the MOC index for southern, central and northern Italy is shown in Figure 6.

As expected, structural differences are evident in terms of organised crime presence between the south and the rest of the country. However, looking at the temporal dynamics, an interesting difference emerges between central and northern Italy. While the north exhibits a flat and stable trend, central displays a cyclical pattern that, albeit with significantly lower intensity, partially mirrors the fluctuations observed in the south, characterised by recurrent peaks followed by subsequent declines. This finding is particularly

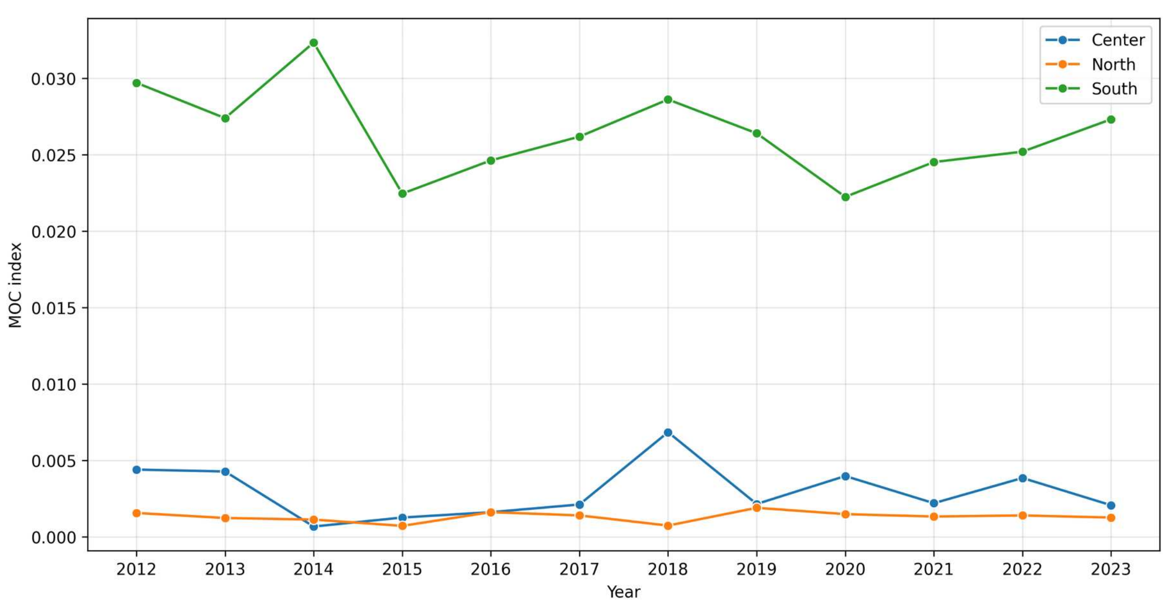

Figure 6. Time evolution of the MOC index by geographical area.

Note: The plot shows the average MOC index value across geographical macro regions (south, central and north, following the National Institute of Statistics classification).

noteworthy when considered in the context of Cavallaro et al. (2020), who proposed that mafia clans possess an inherent resilience that enables them to reorganise despite disruptions centred on key figures, such as arrests. In the south and, to a lesser extent, in the centre, where mafia groups have consolidated their presence, phases of intense criminal visibility are followed by contractions, likely resulting from the periodic disruption of criminal networks by law enforcement agencies, after which criminal organisations eventually regroup. For completeness, we also provided a grid of annual maps in the Appendix in the online supplemental data (Figure C.2), illustrating the geographical distribution of the MOC index for each year.

4.3. Provincial organised crime index

We also computed the organised crime index at the provincial level, utilising official data supplied by the Italian Ministry of the Interior – Department of Public Security (Ministero dell'Interno – Dipartimento della Pubblica Sicurezza, 2024) for the period spanning 2012–2023. In alignment with the aforementioned literature that utilises official data to compute comprehensive indicators (e.g., Calderoni, 2011; Daniele & Marani, 2011), we incorporated offenses that facilitate the identification of mafia syndicates. Moreover, the index encompasses the number of mafia groups identified from DIA maps. Consistent with the framework proposed by Mocetti and Rizzica (2024), we categorise these indicators into the following domains: Objective indicators (mafia-type murders/attempted murders, membership in mafia-type criminal association, municipalities dissolved due to mafia infiltration, and the number of mafia groups); Power syndicate crimes (arson incidents, attacks and extortion); Enterprise syndicate crimes (money laundering).

The aggregation technique employed was the RMDirBoD methodology. The map of the Provincial Organised Crime index is shown in Figure 7.

By computing this composite indicator across a limited set of territorial units (107 provinces), the aggregation process led to a more uniform distribution of values. Consequently, the categorisation of provinces into distinct groups, as required for the MOC index, was rendered unnecessary.

The Provincial Organised Crime index indicates elevated values in the southern provinces, aligning with the well-documented presence and influence of mafia organisations in these areas. This index provides a

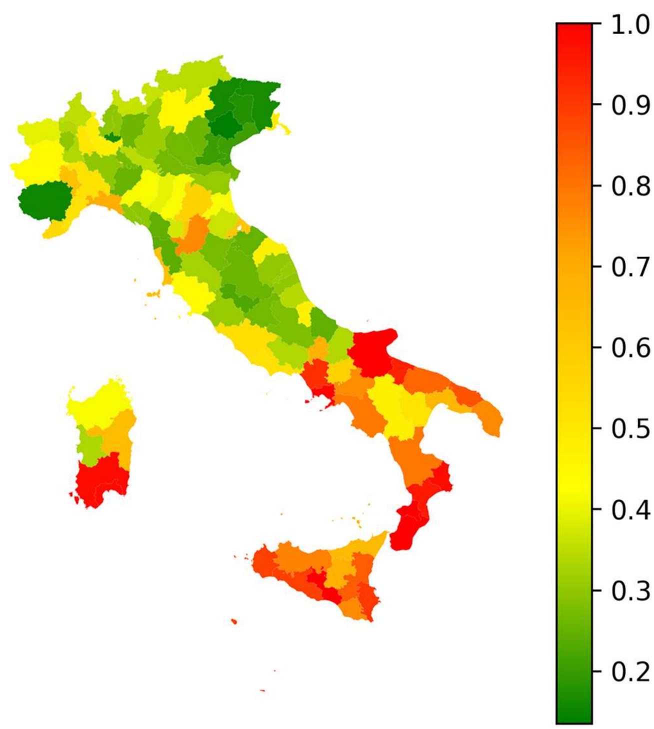

Figure 7. The provincial organised crime index.

macro level perspective that complements the municipal-level findings. For comparative analysis, we aggregated the municipal text-based indicators to the provincial level by calculating the mean of the municipal scores within each province. The correlation between this aggregated indicator and the Provincial Organised Crime index was 0.75, indicating strong similarity in their geographical patterns.

To provide a more granular view of the correlation between the municipal and provincial indicators, we performed a split-sample analysis focusing on two dimensions. The first examines areas where media reporting is sparse, namely peripheral municipalities characterised by a smaller population. Specifically, we calculated the correlation separately for small municipalities (up to 5000 inhabitants, in accordance with Italian Law no. 158/2017) and for larger ones. Although the correlation is 68% for larger municipalities, it remains remarkably high for small municipalities (61%), suggesting that the index is robust even in areas characterised by lower media coverage. The second dimension focuses on the geographical distribution of the phenomenon, comparing traditional mafia regions (south) with non-traditional ones (centre-north). As expected, the correlation is stronger in the south (67%), where the objective indicators of organised crime (e.g., mafia-type murders) are most active. In the centre-north, the correlation is lower (29%) but remains positive and significant at the 5% level. This lower magnitude is attributable to the distinct nature of criminal infiltration in these areas, which is more economically oriented and often harder to distinguish from generic white-collar crime (Mocetti & Rizzica, 2024): while official statistics tend to aggregate mafia-related economic offenses with generic economic crimes, our news-based approach allows for a sharper distinction, isolating events explicitly linked to criminal organisations.

To complement the static geographical analysis, we examined the temporal dynamics of the phenomenon at the provincial level. Figure 8 illustrates the path of the Provincial Organised Crime index over the period considered.

The observed trends underscore the structural dichotomy that characterises the Italian criminal landscape. Southern Italy consistently shows a high and stable trajectory, as the indices range between 0.70 and 0.75, reflecting the deep-rooted and persistent nature of the phenomenon in traditional mafia territories. Unsurprisingly, central and northern Italy exhibit a distinct dynamic, settling on significantly lower values. Although there is some year-to-year variability, the well-known gap between the south and the rest of the country remains pronounced and constant throughout the decade. This evidence corroborates findings from municipal data, validating the existence of a consolidated dualism wherein the intensity of organised crime in its traditional strongholds follows a rigid, path-dependent evolution, distinct from the patterns observed in non-traditional areas.

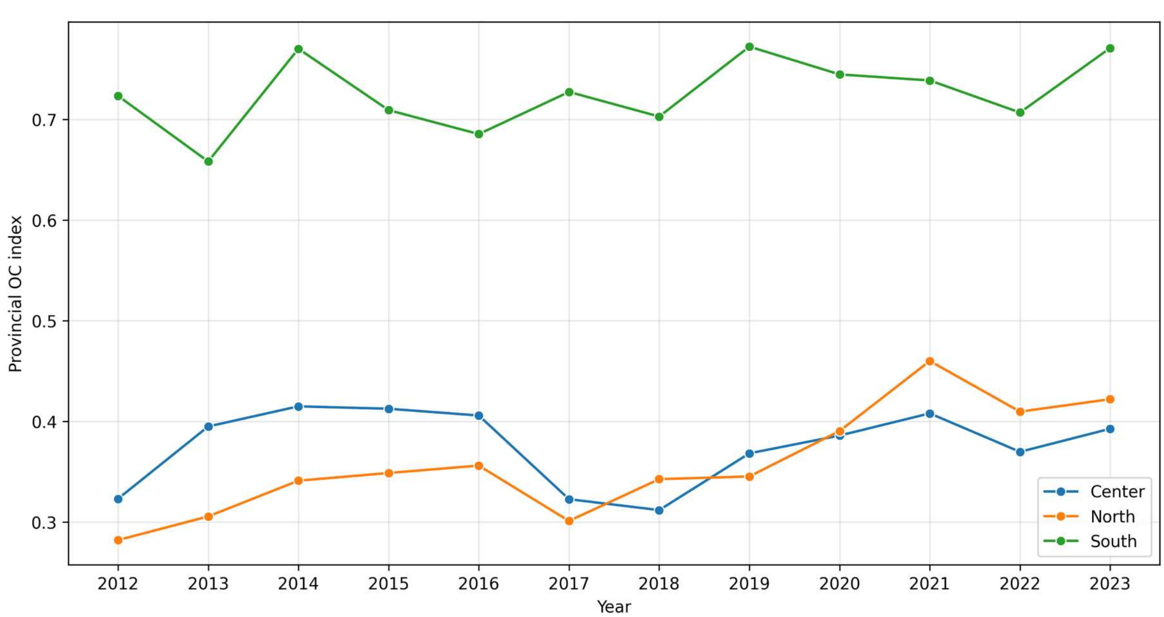

Figure 8. Time evolution of the provincial organised crime index by geographical area.

Note: The plot shows the average Provincial Organised Crime index value across geographical macro regions.

5. Robustness checks

In this section, we perform a series of robustness checks to verify the reliability of our composite indicators across various construction strategies and sub-indicator subsets. Specifically, we first subject the MOC index to a sensitivity analysis involving distinct methodological assumptions. Given that our main RMDirBoD approach relies on endogenous weighting, we benchmark the results against those obtained using standard fixed weighting schemes, namely equal weights and factor analysis based weights.[9] In addition, we extend the analysis to normalisation techniques by comparing min-max scaling with Z-score standardisation.

Figure 9 shows the geographical distribution of different versions of the MOC index:

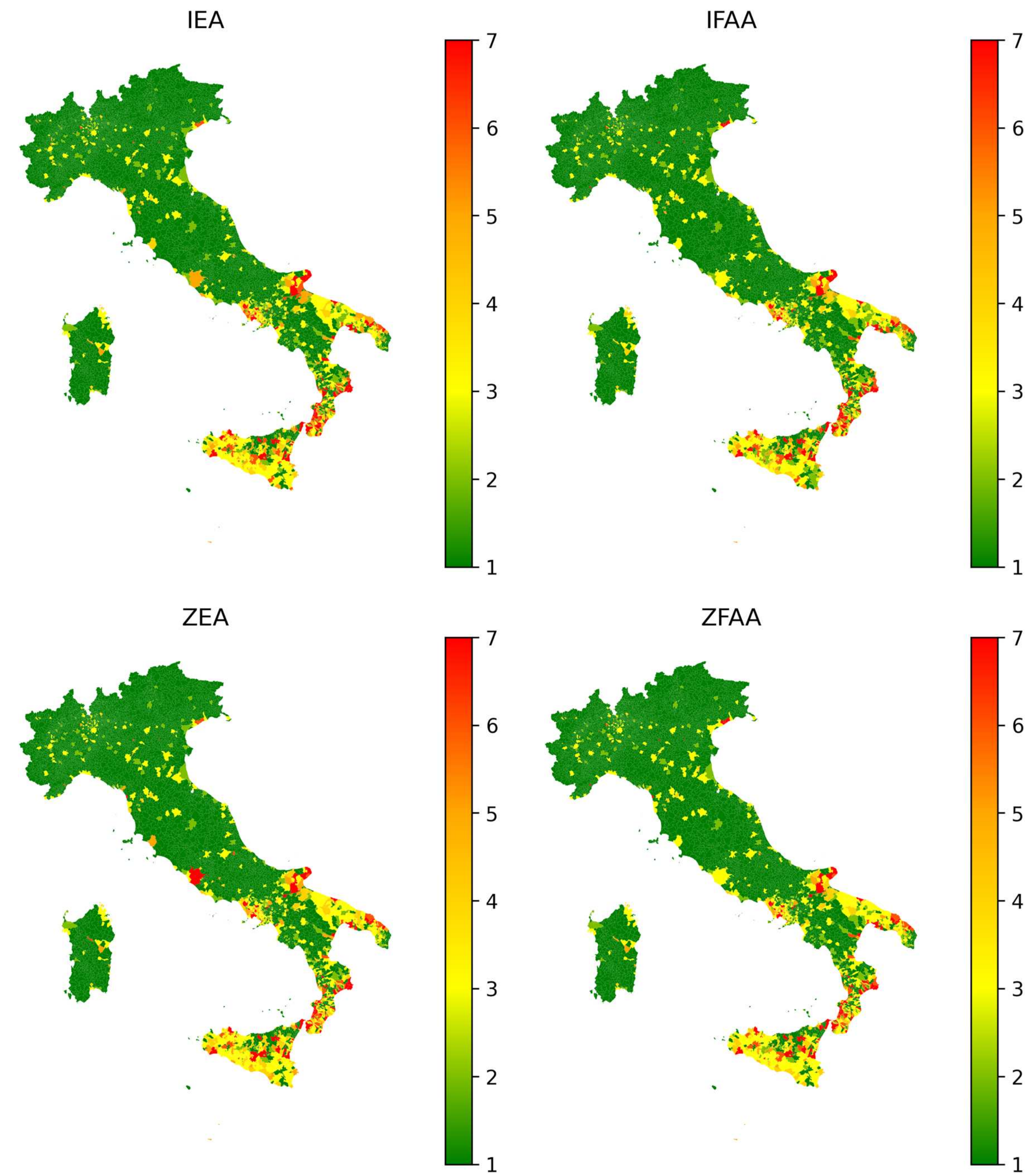

Figure 9. Different versions of the MOC index.

Note: Each map shows the distribution of the MOC index for a different construction strategy, namely Indexing based on min-max normalisation, equal weights, and arithmetic mean (IEA); indexing based on min-max normalisation, factor analysis-based weights, and arithmetic mean (IFAA); Z-score normalisation, equal weights, and arithmetic mean (ZEA); and Z-score normalisation, factor analysis based weights, and arithmetic mean (ZFAA).

The geographical distribution of the composite indicator remains relatively consistent across the various construction strategies employed. Upon visual inspection, no substantial differences are evident between categories as the construction strategy changes; any variations that do occur tend to impact adjacent categories. Consequently, the original MOC index does not appear to be significantly influenced by the construction strategy, as both the distribution and intensity of mafia presence are largely similar across its different iterations. To corroborate this visual inspection with statistical evidence, we calculated the correlation coefficients between the RMDirBoD based MOC index and the four alternative versions (IEA, IFAA, ZEA, ZFAA). The analysis reveals high correlations (all exceeding 85%), confirming that our composite indicator is robust to the selection of construction strategies.[10]

We also assessed the divergence between composite indices derived from alternative weighting and standardisation methods and the index obtained using the RMDirBoD approach. Specifically, we classified each municipality into one of the seven categories, as defined in Figure 5. We then calculated the median of the various MOC indices partitioned into categories. For each category, we calculated the median of the municipal medians and computed the absolute difference between this value and the median of the original MOC index for that category.[11] The observed difference indicates the extent of variation within the category. The computation of median differences across the MOC index categories produced the results presented in Table 3.

The median change across MOC index categories is minimal, as municipalities typically remain within the same category or shift by no more than one adjacent category when the composite indicator is constructed using alternative strategies. These findings suggest that the composite indicator based on the RMDirBoD approach is robust, given that the extent of category variations is limited to one adjacent category when alternative versions of the composite indicator are considered.

As Bernardo et al. (2021) highlight, the inclusion of sub-indicators not directly associated with criminal organisations in the development of organised crime indicators may influence the assessed distribution of organised crime across regions. To address this, we calculated a restricted version of the MOC index and provincial organised crime index, utilising only sub-indicators that unequivocally pertain to organised crime (Type I crime measures), and compared these with versions that incorporate all sub-indicators. The findings reveal strong correlations of 94.9% and 81.4% for the MOC index and provincial organised crime index, respectively, indicating that the inclusion of Type II crime measures does not significantly alter the fundamental spatial distribution of organised crime presence.

To evaluate the sensitivity of the MOC index concerning specific sub-indicators, including those with weaker correlations with official statistics (e.g., arson), we performed a sensitivity analysis employing a leave-one-out approach, as outlined by Dugato et al. (2020). Specifically, we recalculated the MOC index nine times, each time excluding one sub-indicator. Subsequently, we computed the correlations between the complete MOC index and each of the recalculated versions (or leave-one-out iterations) to determine whether particular sub-indicators might disproportionately influence the final results. The findings are presented in Table 4.

The correlations between the full MOC index and each leave-one-out iteration demonstrated high values, indicating that the exclusion of individual sub-indicators did not significantly impact the overall measurement of organised crime presence. Similar results were observed when applying the leave-one-out approach at the provincial level, with correlations between the complete provincial organised crime index and its iterations exceeding 94%.

Additional robustness checks were performed using annual sensitivity analysis of the MOC index and its variants. Specifically, these variants include the composite indicator calculated using the standard MDirBoD

Table 3. Median category change across MOC index categories.

MOC index category	Median category change
1	0
2	0
3	0
4	0.5
5	1
6	1
7	0

Table 4. Correlations between the full MOC index and its leave-one-out iterations.

Excluded sub-indicator	Correlation
Mafia-type murders/attempted murders	0.968
Membership in mafia-type criminal association	0.982
Municipalities dissolved due to mafia	0.877
Number of mafia groups	0.947
Arson incidents	0.985
Attacks	0.971
Extortion	0.977
Money laundering	0.968
Usury	0.976

Table 5. Correlations between the MOC index and its variants by year.

	MDirBoD	IFAA	Type I	LOO1	LOO2	LOO3	LOO4	LOO5	LOO6	LOO7	LOO8	LOO9
2012	0.863	0.775	0.878	0.988	0.983	0.988	0.986	0.976	0.953	0.982	0.945	0.852
2013	0.919	0.867	0.851	0.992	0.958	0.986	0.965	0.967	0.977	0.965	0.939	0.865
2014	0.933	0.843	0.875	0.975	0.981	0.990	0.976	0.968	0.983	0.975	0.950	0.942
2015	0.886	0.739	0.828	0.986	0.947	0.978	0.971	0.974	0.966	0.990	0.936	0.890
2016	0.875	0.809	0.795	0.981	0.967	0.987	0.954	0.968	0.981	0.959	0.931	0.893
2017	0.929	0.897	0.897	0.971	0.985	0.993	0.973	0.960	0.963	0.983	0.952	0.846
2018	0.951	0.826	0.831	0.980	0.985	0.970	0.978	0.987	0.958	0.966	0.959	0.827
2019	0.903	0.803	0.836	0.986	0.950	0.986	0.979	0.954	0.966	0.982	0.947	0.854
2020	0.898	0.845	0.854	0.990	0.981	0.989	0.970	0.960	0.960	0.977	0.949	0.906
2021	0.917	0.797	0.905	0.993	0.970	0.991	0.990	0.959	0.974	0.974	0.959	0.855
2022	0.927	0.843	0.827	0.965	0.975	0.981	0.987	0.971	0.965	0.993	0.952	0.884
2023	0.874	0.804	0.837	0.982	0.967	0.975	0.989	0.973	0.977	0.960	0.940	0.911

Note: The table presents the correlations between the MOC index, as discussed in Section 4.2, and its variants across different years. Specifically, these variants include the composite indicator calculated using the standard MDirBoD approach, which does not incorporate the robust optimisation techniques employed in the RMDirBoD model, the index based on factor analysis weights (IFAA), the index limited to Type I crime measures, and the composite indicators derived through the leave-one-out strategy (columns from LOO1 to LOO9). The labels LOO1 to LOO9 correspond to the indices excluding the following sub-indicators: membership in mafia-type criminal association, attacks, arson incidents, extortion, Mafia-type murders/attempted murders, money laundering, usury, number of mafia groups and municipalities dissolved due to mafia.

approach (i.e., excluding the robust optimisation techniques employed in the RMDirBoD model), the index derived from factor analysis weights, the index restricted to Type I crime measures, and the composite indicators generated via the leave-one-out strategy. The correlations presented in Table 5 indicate a high degree of stability for each year, with an average correlation of 0.936, ranging from 0.739 to 0.993.

Similarly, the provincial organised crime index exhibits substantial stability over the years when compared with its variants, showing an average correlation of 0.922, spanning from 0.623 to 0.998. Table C.1, containing all the correlations, is provided in the Appendix in the online supplemental data.

Although criminal infiltration is acknowledged as a dynamic institutional pathology (Rodríguez-Pose, 2013), the phenomenon of the mafia is nonetheless marked by significant persistence. This persistence implies that the regions most susceptible to mafia influence today are largely the same as those that have been historically affected (Mocetti & Rizzica, 2024). Consequently, although some year-to-year variation in the annual iterations of the MOC index was anticipated, a high degree of temporal consistency was expected. To assess the structural persistence of organised crime presence, particularly whether the hierarchy of affected municipalities remains stable over time, Spearman's rank correlation coefficient was calculated between consecutive years. The findings indicate a strong correlation with an average value of 85%, ranging from a minimum of 77% to a maximum of 93%. Similar results were observed for the provincial organised crime index, which exhibited an average correlation of 70%, with values ranging from 64% to 81%. A table detailing the correlations for each pair of adjacent years is provided in the Appendix in the online supplemental data (Table C.2).

6. Conclusions

This study introduces a novel dual data-driven framework to map organised crime in Italy, addressing the challenges posed by the lack of granular and reliable data on mafia-related activities. By constructing and cross-validating two complementary indices – one at the municipal level using advanced text mining and computer vision techniques and another at the provincial level using official records – we provide a comprehensive and robust measure of the presence of organised crime.

The municipal level index, derived from over a decade of ANSA news articles and DIA investigative reports, offers unprecedented granularity, uncovering hyperlocal patterns of criminal activity that are often missed in coarser data sets. This index captures the intensity of organised crime at the municipal level and identifies specific hotspots, enabling targeted interventions. Meanwhile, the provincial level index synthesises official data on mafia-related offenses and administrative dissolutions, providing a normalised benchmark that accounts for population differences and allows meaningful comparisons across regions.

A significant outcome of this study was the substantial alignment between the two indices, as evidenced by a 75% spatial alignment. Furthermore, components that are unequivocally associated with mafia activity exhibit correlations near 90%, thereby validating the use of media narratives as a reliable proxy for underreported criminal activity. This alignment highlights the effectiveness of integrating unstructured data sources, such as news articles and investigative reports, with traditional structured data to address the limitations of official data. Moreover, the implementation of the RMDirBoD technique ensures that the composite indices are robust against common statistical challenges, such as variability and the presence of outliers, thereby rendering them a dependable tool for policymakers.

The dual approach employed in this study is particularly relevant as it combines different perspectives, news reports and official records to detect organised crime, a phenomenon that often operates in silence and remains underreported. By integrating these diverse data sources, the analysis provides a more comprehensive and accurate picture of mafia-related activities, which is crucial for understanding their true extent and impact on society. This comprehensive view is valuable for law enforcement agencies, enabling them to concentrate resources on identified hotspots of organised crime, and for institutions tasked with implementing preventive measures. These measures are essential for eradicating the sources that sustain organised crime, which is recognised as a major factor contributing to the underdevelopment of certain areas in Italy. From a broader regional development perspective, these high-resolution metrics address a significant gap by applying the concept of institutional traps at a detailed level. Our mafia indices enable researchers to empirically identify specific local contexts in which criminal governance undermines the returns on public investment and impedes the economic convergence expected by place-based policies.

The implications of this study are significant. Policymakers can leverage these indices to assess the prevalence of organised crime at both the micro and macro levels, facilitating the design of tailored law enforcement strategies and the restoration of legal conditions essential for community and economic activities. Granular insights from the municipal index can guide targeted interventions in specific municipalities, whereas the provincial index can inform broader policy decisions and resource allocation.

This study opens several avenues for future research. The methodologies developed here could be applied to other types of crime or extended to different geographical contexts, potentially enhancing our understanding of global criminal ecosystems. In this regard, future research could advance the current two-stage NLP pipeline by replacing the human-in-the-loop strategy employed in the second stage with a fully automated approach based on transformer-based models, specifically fine-tuned BERT models. This enhancement would be particularly valuable for processing large amounts of data, where direct human involvement is no longer feasible, at the cost of accepting a trade-off between data scalability and classification accuracy. In general, further refinement of text mining and computer vision techniques could improve the accuracy and scope of data extraction. Given that economic decisions are often driven by individual perceptions, future research could integrate subjective indicators, as done by Mocetti and Rizzica (2024) at the provincial level, by utilising survey data. For our composite indicator, this subjective component could be captured via sentiment analysis of ANSA news reports. Furthermore, granular data at the municipal and provincial levels could be employed to investigate the socio-economic impacts of organised crime at a micro level. For example, future research might explore whether the presence of organised crime acts as a structural factor influencing corruption or serves as an intermediary variable.

In conclusion, this study demonstrates the power of combining advanced data analytics with traditional data sources to address complex social issues. By providing a detailed and reliable map of organised crime in Italy, this study not only advances academic understanding but also offers practical tools for policymakers to effectively combat mafia-related activities.

Notes

1. The data employed in this textual analysis were previously used in Forgione and Migliardo (2025) to construct a municipal-level mafia index. However, the latter index was developed by integrating textual data with additional structural data, such as firms and real estate assets seized because of their association with organised crime, as well as municipalities dissolved due to mafia infiltration. Furthermore, this indicator possesses a cross-sectional structure and pertains to a more limited temporal span.
2. This index was developed through collaboration between the Catholic University of the Sacred Heart–Transcrime and the Italian Ministry of the Interior.
3. A comprehensive survey conducted by Taha (2025) provides an in-depth analysis of machine learning and deep learning techniques employed in the development of crime prediction algorithms. This framework highlights the growing importance of text mining in extracting crime-related information from unstructured data sources, such as news articles.
4. NER can be defined as the task of locating and classifying named entities mentioned in unstructured texts into predefined classes, such as personal names, organisations and locations.
5. We duplicated each news item that referred to multiple municipalities, obtaining one observation for each municipality to which the news referred.
6. Arson is quite common, especially in southern Italy, where the peak of this crime is recorded during the summer due to forest fires, which are rarely linked to organised crime.
7. It is worth noting that the Mocetti and Rizzica's (2024) framework also includes a fourth dimension concerning subjective indicators (e.g., perception of security). However, owing to the unavailability of reliable survey data at the municipal level, we restricted our analysis to the first three domains.
8. A comparison between the two composite indicators is provided in the Appendix in the online supplemental data.
9. Factor analysis based weights is a common weighting scheme in which component weights are derived from the factor loadings obtained in factor analysis; see Nardo et al. (2005) for details on how this weighting scheme works.
10. Similarly, the provincial organised crime index exhibits high methodological stability, as the correlation between the RMDirBoD based version and the one derived using factor analysis based weights is 86%.
11. We opted to use the median rather than the mean for two primary reasons. Firstly, the median demonstrates greater robustness in the presence of outliers. Secondly, calculating an average for a variable that represents categories is methodologically inappropriate, as categories constitute qualitative data lacking uniform numerical intervals.

Acknowledgments

The authors express their gratitude to the anonymous reviewers for their insightful comments and constructive suggestions, which have substantially enhanced the quality and robustness of this work. Their recommendations concerning the validation of text mining techniques, the analysis of temporal dynamics, and the additional robustness checks were particularly instrumental in refining the methodological framework. The authors would like to thank the participants of the conferences where earlier versions of this paper were presented for their valuable comments and suggestions. These include the conference 'Manifesto and Research Frontiers for a Renaissance in Economics' (University of Perugia, Italy) and the 65th SIE Annual Conference (University of Urbino, Italy).

Author contributions

All authors contributed equally to the work.

Ethical standards

The authors declare that the research was conducted in accordance with ethical standards. No human participants, animals, or personal data were involved in this study. The study complies with the relevant laws and regulations.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research project received funding from the European Union – NextGenerationEU – National Recovery and Resilience Plan (NRRP) – Mission 4 Component 2, Investment N. 1.1, CALL PRIN 2022 PNRR D.D. 1409 14-09-2022 – entitled

‘The effect of organized crime on firm technical efficiency and R&D investments’, ID P20227XY5N, CUP J53D23016850001 (University of Messina), P20227XY5N_001.

Data availability statement

The data used in this study were obtained from Lexis Nexis (Nexis Uni), the Italian Ministry of the Interior (2012–2023) and the Anti-Mafia Investigation Directorate (DIA). The derived indices generated during the study are available from the corresponding author on reasonable request.

ORCID

Antonio Fabio Forgione http://orcid.org/0000-0003-4015-4939
Carlo Migliardo http://orcid.org/0000-0002-4055-0782
Paolo Mustica http://orcid.org/0000-0002-3218-5084

References

Alesina, A., Piccolo, S., & Pinotti, P. (2019). Organized crime, violence, and politics. *The Review of Economic Studies*, *86*(2), 457–499. https://doi.org/10.1093/restud/rdy036

Aprigliano, V., Emiliozzi, S., Guaitoli, G., Luciani, A., Marcucci, J., & Monteforte, L. (2023). The power of text-based indicators in forecasting Italian economic activity. *International Journal of Forecasting*, *39*(2), 791–808. https://doi.org/10.1016/j.ijforecast.2022.02.006

Arbolino, R., & Boffardi, R. (2024). Organized crime and corruption: What are the consequences for Italian cohesion policy investments? *Regional Studies*, *58*(1), 178–193. https://doi.org/10.1080/00343404.2023.2188181

Arulanandam, R., Savarimuthu, B. T. R., & Purvis, M. A. (2014). Extracting crime information from online newspaper articles. In *Proceedings of the Second Australasian Web Conference-Volume 155* (pp. 31–38).

Barbero, J., Mandras, G., Rodríguez-Crespo, E., & Rodríguez-Pose, A. (2021). Quality of government and regional trade: Evidence from European union regions. *Regional Studies*, *55*(7), 1240–1251. https://doi.org/10.1080/00343404.2021.1873934

Battisti, M., Bernardo, G., Konstantinidi, A., Kourtellos, A., & Lavezzi, A. M. (2020). Socio-economic inequalities and organized crime: An empirical analysis. In *Understanding recruitment to organized crime and terrorism* (pp. 205–239). Springer.

Bernardo, G., Brunetti, I., Pinar, M., & Stengos, T. (2021). Measuring the presence of organized crime across Italian provinces: A sensitivity analysis. *European Journal of Law and Economics*, *51*(1), 31–95. https://doi.org/10.1007/s10657-020-09676-0

Bonisoli, G., Rollo, F., & Po, L. (2021). Using word embeddings for Italian crime news categorization. In *2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS)* (pp. 461–470). IEEE.

Boukabous, M., & Azizi, M. (2022). Multimodal sentiment analysis using audio and text for crime detection. In *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)* (pp. 1–5). IEEE.

Buscaglia, E., & Van Dijk, J. (2003). Controlling organized crime and corruption in the public sector. *Forum on Crime and Society*, *3*(1-2), 3–34.

Calderoni, F. (2011). Where is the mafia in Italy? Measuring the presence of the mafia across Italian provinces. *Global Crime*, *12*(1), 41–69. https://doi.org/10.1080/17440572.2011.548962

Cavallaro, L., Ficara, A., De Meo, P., Fiumara, G., Catanese, S., Bagdasar, O., Song, W., & Liotta, A. (2020). Disrupting resilient criminal networks through data analysis: The case of sicilian mafia. *PLoS ONE*, *15*(8), e0236476. https://doi.org/10.1371/journal.pone.0236476

Charron, N., Dijkstra, L., & Lapuente, V. (2014). Regional governance matters: Quality of government within European union member states. *Regional Studies*, *48*(1), 68–90. https://doi.org/10.1080/00343404.2013.770141

Cinelli, M., Spada, M., Kim, W., Zhang, Y., & Burgherr, P. (2021). Mcda index tool: An interactive software to develop indices and rankings. *Environment Systems and Decisions*, *41*(1), 82–109. https://doi.org/10.1007/s10669-020-09784-x

Coniglio, N. D., Celi, G., Scagliusi, C., et al. (2010). Organized crime, migration and human capital formation: Evidence from the South of Italy. *Southern Europe Research in Economic Studies*, Working Paper No. 28.

Crescenzi, R., Di Cataldo, M., & Rodríguez-Pose, A. (2016). Government quality and the economic returns of transport infrastructure investment in European regions. *Journal of Regional Science*, *56*(4), 555–582. https://doi.org/10.1111/jors.12264

Daniele, V., & Marani, U. (2011). Organized crime, the quality of local institutions and FDI in Italy: A panel data analysis. *European Journal of Political Economy*, *27*(1), 132–142. https://doi.org/10.1016/j.ejpoleco.2010.04.003

Das, P., Das, A. K., Nayak, J., Pelusi, D., & Ding, W. (2021). Incremental classifier in crime prediction using bi-objective particle swarm optimization. *Information Sciences*, *562*, 279–303. https://doi.org/10.1016/j.ins.2021.02.002

Direzione Investigativa Antimafia. (2012–2023). Relazione del ministro dell’interno al parlamento sull’attività svolta e sui risultati conseguiti dalla direzione investigativa antimafia. Report, Ministero dell’Interno.

Dugato, M., Calderoni, F., & Campedelli, G. M. (2020). Measuring organised crime presence at the municipal level. *Social Indicators Research*, *147*(1), 237–261. https://doi.org/10.1007/s11205-019-02151-7
Feuerriegel, S., & Gordon, J. (2018). Long-term stock index forecasting based on text mining of regulatory disclosures. *Decision Support Systems*, *112*, 88–97. https://doi.org/10.1016/j.dss.2018.06.008
Ficara, A., Cavallaro, L., Curreri, F., Fiumara, G., De Meo, P., Bagdasar, O., Song, W., & Liotta, A. (2021). Criminal networks analysis in missing data scenarios through graph distances. *PLoS ONE*, *16*(8), e0255067. https://doi.org/10.1371/journal.pone.0255067
Fioroni, T., Lavezzi, A. M., & Trovato, G. (2025). Organized crime, corruption, and economic growth. *Journal of Regional Science*, *65*(2), 535–560. https://doi.org/10.1111/jors.12751
Forgione, A. F., & Migliardo, C. (2025). The mafia's economic grip: Firm efficiency and a composite indicator of organized crime. *Journal of Regional Science*, *65*(3), 759–779. https://doi.org/10.1111/jors.12761
Freunek, M., & Niggli, M. (2023). Introducing DynaPTI–constructing a dynamic patent technology indicator using text mining and machine learning. *Frontiers in Artificial Intelligence*, *6*, 1136846. https://doi.org/10.3389/frai.2023.1136846
Fusco, E. (2023). Potential improvements approach in composite indicators construction: The multi-directional benefit of the doubt model. *Socio-Economic Planning Sciences*, *85*, 101447. https://doi.org/10.1016/j.seps.2022.101447
Ganau, R., & Rodríguez-Pose, A. (2018). Industrial clusters, organized crime, and productivity growth in Italian SMEs. *Journal of Regional Science*, *58*(2), 363–385. https://doi.org/10.1111/jors.12354
Hassan, M., & Rahman, M. Z. (2017). Crime news analysis: Location and story detection. In *2017 20th International Conference of Computer and Information Technology (ICCIT)* (pp. 1–6). IEEE.
Jacinto, R., Reis, E., & Ferrão, J. (2020). Indicators for the assessment of social resilience in flood-affected communities – A text mining-based methodology. *Science of the Total Environment*, *744*, 140973. https://doi.org/10.1016/j.scitotenv.2020.140973
Jayaweera, I., Sajeewa, C., Liyanage, S., Wijewardane, T., Perera, I., & Wijayasiri, A. (2015). Crime analytics: Analysis of crimes through newspaper articles. In *2015 moratuwa Engineering Research Conference (MERCon)* (pp. 277–282). IEEE.
LexisNexis (2024). Nexis Uni database (ANSA news archive 2012-2023). https://www.lexisnexis.com/
Marselli, R., & Vannini, M. (1997). Estimating a crime equation in the presence of organized crime: Evidence from Italy. *International Review of law and Economics*, *17*(1), 89–113. https://doi.org/10.1016/S0144-8188(96)00060-9
Mendez, C., & Bachtler, J. (2024). The quality of government and administrative performance: Explaining cohesion policy compliance, absorption and achievements across eu regions. *Regional Studies*, *58*(4), 690–703. https://doi.org/10.1080/00343404.2022.2083593
Ministero dell'Interno – Dipartimento della Pubblica Sicurezza (2024). Numero dei delitti commessi e denunciati da tutte le Forze di Polizia all'Autorità Giudiziaria (rilevazione cod. ISTAT INT00062). Data provided upon request. Rome, Italy.
Mocetti, S., & Rizzica, L. (2024). Organized crime in Italy: An economic analysis. *Italian Economic Journal*, *10*(3), 1339–1360. https://doi.org/10.1007/s40797-023-00236-4
Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2005). *Handbook on constructing composite indicators: Methodology and user guide*. OECD.
Park, K., & Kremer, G. E. O. (2017). Text mining-based categorization and user perspective analysis of environmental sustainability indicators for manufacturing and service systems. *Ecological Indicators*, *72*, 803–820. https://doi.org/10.1016/j.ecolind.2016.08.027
Pinotti, P. (2015). The economic costs of organised crime: Evidence from Southern Italy. *The Economic Journal*, *125*(586), F203–F232. https://doi.org/10.1111/ecoj.12235
Rodríguez-Pose, A. (2013). Do institutions matter for regional development? *Regional Studies*, *47*(7), 1034–1047. https://doi.org/10.1080/00343404.2012.748978
Rodríguez-Pose, A., & Garcilazo, E. (2018). Quality of government and the returns of investment: Examining the impact of cohesion expenditure in European regions. In *Place-based economic development and the New EU cohesion policy* (pp. 34–50). Routledge.
Salvatore, C., Biffignandi, S., & Bianchi, A. (2024). Augmenting business statistics information by combining traditional data with textual data: A composite indicator approach. *Metron*, *82*(1), 71–91. https://doi.org/10.1007/s40300-023-00261-4
Solivetti, L. M. (2016). Crime patterns between tradition and change: A territorial analysis of the Italian provinces. *Social Indicators Research*, *128*(2), 531–558. https://doi.org/10.1007/s11205-015-1043-1
Taha, K. (2025). Empirical and experimental insights into data mining techniques for crime prediction: A comprehensive survey. *ACM Transactions on Intelligent Systems and Technology*, *16*(2), 1–75. https://doi.org/10.1145/3699515
Union, E., & Centre, J. R. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. OECD publishing.
Van Dijk, J. (2007). Mafia markers: Assessing organized crime and its impact upon societies. *Trends in Organized Crime*, *10*(4), 39–56. https://doi.org/10.1007/s12117-007-9013-x
Vidoli, F., Fusco, E., Pignataro, G., & Guccio, C. (2024). Multi-directional robust benefit of the doubt model: An application to the measurement of the quality of acute care services in OECD countries. *Socio-Economic Planning Sciences*, *93*, 101877. https://doi.org/10.1016/j.seps.2024.101877