ISSN 1826-3534

federalismi.it
RIVISTA DI DIRITTO PUBBLICO ITALIANO, COMPARATO, EUROPEO

2 DICEMBRE 2020

La misurazione della corruzione attraverso le sentenze della magistratura: una proposta metodologica con strumenti di text mining

di Maria Francesca Romano
Istituto di Economia & EMBeDS
Scuola Superiore Sant'Anna

Gaetana Morgante
Istituto DIRPOLIS
Scuola Superiore Sant'Anna

Antonella Baldassarini
ISTAT

Giuseppe Di Vetta
Istituto DIRPOLIS
Scuola Superiore Sant'Anna

Pasquale Pavone
Istituto di Economia & EMBeDS
Scuola Superiore Sant'Anna

La misurazione della corruzione attraverso le sentenze della magistratura: una proposta metodologica con strumenti di text mining*

di M.F. Romano, G. Morgante, G. Di Vetta, P. Pavone (Scuola Superiore Sant'Anna Pisa) **e A. Baldassarini** (Istat)

Abstract [It]: La definizione dei metodi di misurazione dei fenomeni corruttivi rappresenta un tema centrale nel dibattito scientifico e, soprattutto, nei contesti multilaterali dove si elaborano le politiche di contrasto a livello globale. Le Autrici e gli Autori muovono dalla ricognizione delle metodiche sinora impiegate, evidenziando i limiti e le criticità che caratterizzano in particolar modo gli strumenti di misurazione c.d. soggettivi o percettivi. Su questa premessa, nell'articolo si propone un'innovativa metodologia, a carattere interdisciplinare, attraverso la quale raggiungere una rappresentazione oggettiva del fenomeno corruttivo, come emerge dalle fonti giudiziarie. In particolare, il metodo presentato si fonda sull'applicazione di strumenti di text mining e tecniche statistiche ad una base informativa, oggettiva e stabile, rappresentata da un campione significativo di sentenze pronunciate dalla Suprema Corte di Cassazione, nel periodo compreso tra il 2015 e il gennaio 2020, in materia di reati corruttivi (artt. 317, 318, 319, 319-quater, 321 del codice penale). Nell'articolo si offre una preliminare esposizione del metodo e se ne pongono in evidenza le potenzialità conoscitive e applicative, che non si esauriscono nella misurazione (quantitativa) del fenomeno osservato. L'analisi automatica, mediante text mining, consente infatti di interrogare a fondo la fonte giudiziaria (sentenze) per trarne informazioni estremamente significative nella prospettiva scientifica della descrizione qualitativa della variegata fenomenologia corruttiva. Il metodo a base oggettiva presentato, infine, rappresenta uno strumento promettente per lo sviluppo di politiche di prevenzione e contrasto evidence-based.

Abstract [En]: The definition of the methods of measurement of corruption is a central topic in the scientific debate and in multilateral contexts where the policies of contrast are developed at a global level. The Authors start from the recognition of the methods used up to now, highlighting the limits and criticalities that characterize the so-called subjective or perceptive measuring instruments. On this premise, the article proposes an innovative methodology, of an interdisciplinary nature, through which to achieve an objective representation of the corruption phenomenon, as emerges from the judicial sources. In particular, the method presented is based on the application of text mining and statistical techniques to an objective and stable information base, represented by a significant sample of sentences handed down by the Supreme Court of Cassation, in the period between 2015 and January 2020, regarding corruption offenses (arts. 317, 318, 319, 319-quater, 321 of the penal code). This article offers a preliminary exposition of the shown method and highlights its cognitive and applicative potential, which are not limited to the (quantitative) measurement of the observed phenomenon. The automatic analysis, through the text mining, allows to thoroughly interrogate the judicial source (judgments) in order to obtain extremely significant information from the scientific perspective of the qualitative description of the variegated corrupting phenomenology. Finally, the objective-based method presented represents a promising tool for the development of evidence-based policy.

* Articolo sottoposto a referaggio. Il lavoro scientifico svolto dalle Autrici e dagli Autori si è concretizzato in due articoli complementari: il presente Articolo sviluppa maggiormente prospettive giuridiche, criminologiche e di policy; l'altro, invece, affronta gli aspetti metodologici statistici sottesi ad entrambi gli articoli. Un ringraziamento per nulla convenzionale va al prof. Guido Rey che con lungimiranza ha saputo intravedere l'importanza di superare gli steccati disciplinari e proporre una prospettiva innovativa di indagine dei fenomeni sociali letti anche attraverso i documenti legali. Il suo entusiasmo e la sua energia hanno saputo dapprima costruire un gruppo di ricerca piccolo ma appassionato, poi la sua tenacia e il suo incoraggiamento costante hanno fatto in modo che il gruppo e la ricerca resistessero alle vicissitudini (anche finanziarie) della vita accademica. Per sua scelta non è Co-Autore ma a pieno diritto merita il titolo di Maestro.

Sommario: **1.** La misurazione della corruzione nella prospettiva (futuribile) di un'evidence-based legislation. **2.** Strumenti e indici di misurazione del fenomeno: criticità e limiti. **3.** Proposta di misurazione e assessment del fenomeno corruttivo attraverso l'analisi automatica dei testi delle sentenze. **4.** L'approccio empirico: dati, strumenti e metodi. **5.** Risultati. **6.** Discussione e potenziali sviluppi della ricerca. **7.** Appendice: le fonti dei dati. **7.1.** Le sentenze della Corte Suprema di Cassazione: il portale Italgiure Web. **7.2.** Il registro ASIA delle imprese attive. **7.3.** Il database ORBIS.

1. La misurazione della corruzione nella prospettiva (futuribile) di un'*evidence-based legislation*

La necessità di perfezionare approcci e metodologie sinora applicate per la misurazione del fenomeno corruttivo, e di elaborarne di nuove, costituisce una consapevolezza condivisa sia in ambito scientifico, sia nei contesti, soprattutto internazionali, dove si definiscono i programmi di contrasto e prevenzione della corruzione, e più in generale dei comportamenti di *maladministration* ad essa connessi o, comunque, assimilabili.

Il tema della misurazione della corruzione ha assunto una portata determinante in considerazione del fatto che la programmazione di efficaci strategie di contrasto non è realmente concepibile a prescindere dalla rilevazione della dimensione empirica del fenomeno e dalla considerazione delle sue correlazioni con altre "grandezze" e variabili rilevanti che possono riguardare sia il contesto istituzionale (politico, economico, sociale) nel suo complesso, sia specifici settori funzionali dell'apparato pubblico o privato. E' significativo, a tal proposito, come nell'ottica del perseguimento dell'Obiettivo 16 («*Peace and Security*») dei SDGs (*Sustainable Development Goals*), nel cui ambito il contrasto ad ogni forma di corruzione assume un ruolo preminente (16.5 dei SDGs), l'UNODC *Research* abbia adottato un autentico *Manual on corruption surveys*[1], cioè un complesso di *methodological guidelines* per la definizione e implementazione da parte degli Stati aderenti di innovativi e solidi strumenti di misurazione di variegate forme di corruzione, pubblica e privata. Questa opzione, adottata in seno al sistema UN di contrasto, conferma la straordinaria centralità della *measurement methodology* nell'agenda internazionale.

Lo sviluppo di un approccio *evidence-based* al tema in esame, che emerge in modo evidente sul piano internazionale, ha imponenti implicazioni anche di ordine teorico, soprattutto per ciò che attiene alla definizione dei "metodi" della politica-criminale e della formulazione delle opzioni legislative in materia penale. La validazione empirica delle politiche penali rappresenta un'esigenza non sempre tenuta nella dovuta considerazione nella legislazione interna, che risulta talvolta refrattaria a modelli *evidence-based*, in quanto fondati su acquisizioni e risultati empiricamente verificati.

L'esperienza legislativa più recente, maturata – non a caso – sul terreno del contrasto al fenomeno corruttivo conferma l'atteggiamento "contro-epistemico" del legislatore interno. È sufficiente considerare, in questo senso, il caso paradigmatico della c.d. *Legge Spazzacorrotti* (l. 09.01.2019, n. 3), con

[1] Pubblicato nel 2018 e consultabile in rete.

particolare riferimento all'estensione del regime penitenziario differenziato[2], previsto dall'art. 4-*bis* dell'ordinamento penitenziario, a taluni reati contro la pubblica amministrazione, comprese, naturalmente, le fattispecie di corruzione e concussione. La previsione per questi reati del doppio binario penitenziario – all'origine concepito per i detenuti, condannati per delitti in materia di "criminalità organizzata" e, poi, progressivamente esteso – è in realtà espressione – come osservato in letteratura[3] - di una precisa visione politico-criminale che tende, in buona sostanza, ad assimilare fenomenologie criminali, come corruzione e crimine organizzato, qualitativamente eterogenee.

A prescindere da un esame critico del merito di questa opzione sanzionatoria, che ha un significativo retroterra storico e culturale[4] e si pone in linea di continuità con iniziative legislative che muovono nella medesima direzione di un superamento dell'autonoma *identità* criminologica della corruzione (concepita come tipica criminalità ascrivibile ai c.d. *white collars*), ciò che rileva è, in special modo, la sostanziale *autoreferenzialità* della scelta normativa: la carenza, da più parti rilevata, di un patrimonio conoscitivo, di un fondamento empirico-statistico, di un *pattern* teoretico e criminologico in grado di sostenere, in termini di ragionevolezza, la proiezione *tout court* dei "reati di corruzione" (per richiamare un lessico discutibile e assai diffuso) nell'orbita disciplinare e sanzionatoria destinata alla criminalità organizzata di tipo mafioso o, comunque, in questi termini teleologicamente connotata.

In ultima analisi, l'esperienza della l. *Spazzacorrotti* (per i profili connessi all'esecuzione penale), già in parte dichiarata incostituzionale con riguardo al regime intertemporale dell'estensione del regime differenziato, restituisce l'immagine di una politica-criminale tendenzialmente refrattaria alla verifica dei predicati epistemici e, in senso lato, scientifici che, secondo un modello *evidence-based*, dovrebbero sostenerla. In effetti – ma questa osservazione apre un campo critico più ampio, che esula da questo contributo -, anche il controllo costituzionale di ragionevolezza (art. 3 Cost.) risulta sensibilmente depotenziato per ciò che attiene al profilo della c.d. razionalità forte, che si sostanzia nella verifica della giustificazione empirico-scientifica delle scelte normative (ad es. dell'opzione di incriminazione) con riferimento alla generalità dei casi cui si riferisce.

---

[2] Sul punto, ormai in un'ampia bibliografia: V. MANES, *L'estensione dell'art. 4-bis ord. pen. ai delitti contro la P.A.: profili di illegittimità costituzionale,* in *Dir. pen. cont.,* 2, 2019; più in generale, sull'intervento legislativo in parola, il commento di T. PADOVANI, *La spazzacorrotti. Riforma delle illusioni e illusioni della riforma,* in *questa Rivista,* 2018, 1 ss.; V. MONGILLO, *La Legge "Spazzacorrotti": ultimo approdo del diritto penale emergenziale nel cantiere permanente dell'anticorruzione*, in *Dir. pen. cont.,* 5, 2019.

[3] In tal senso, V. MONGILLO, *Crimine organizzato e corruzione: dall'attrazione elettiva alle convergenze repressive,* in *Riv. trim. dir. pen. cont.,* 1, 20109, 158 ss.

[4] La visione politico-criminale, cui si è accennato nel testo, trae le sue premesse dalla stagione di Tangentopoli e, soprattutto, dalla consapevolezza che ne è emersa, anche in ambito scientifico: cfr. in tema i saggi raccolti in G. FORTI (a cura di), *Il prezzo della tangente. La corruzione come sistema a dieci anni da "mani pulite",* Milano, 2003 e, spec., i contributi di A. VANNUCCI e G. FORTI (*ivi* ampi riferimenti).

Le indicazioni che provengono dagli organismi e agenzie internazionali, in questa materia, propongono, invece un modello di *evidence-based policymaking* (EBPM), che muove dalla preliminare necessità di una misurazione obiettiva del fenomeno e di un successivo *assessment*, attuati mediante strumenti e metodologie validate. In buona sostanza, è sul terreno delle politiche "anti-corruzione" che sembra prospettarsi – forse in modo paradossale, se si considera che questo è tra gli ambiti più investiti dalle distorsioni "populiste"[5] – quell'apertura alle "scienze empiriche"[6] che costituisce la pre-condizione per una scienza penale integrata[7], al momento, tuttavia, ancora un mero programma teorico.

2. Strumenti e indici di misurazione del fenomeno: criticità e limiti

Tutto quanto premesso sull'opportunità di un approccio empirico-quantitativo alla rappresentazione di tratti distintivi e impatto economico della corruzione, le finalità della misurazione del fenomeno sono molteplici: consente, anzitutto, la rappresentazione congrua e relativamente affidabile ("verificabile") dell'oggetto in esame: segnatamente un fenomeno criminale estremamente complesso, soprattutto per la sua capacità di mimetizzazione sociale e ambientale, che ne rende estremamente ardua la rivelazione tramite metodi *diretti*, come indagini di vittimizzazione o, comunque, indicatori percettivi (es. *population surveys* e *user surveys*).

Questa rappresentazione attendibile (quanto meno alla luce di un determinato protocollo di indagine) contribuisce, in modo significativo, all'approfondimento e perfezionamento di modelli teorico-esplicativi (per esempio di estrazione criminologica) del fenomeno criminale; in altri termini, la formazione di un patrimonio conoscitivo oggettivo rende possibile la "falsificazione" empirica dei tentativi (qualitativi) di ricostruzione delle forme e dinamiche con cui si manifesta la corruzione (intesa in senso lato). In questa prospettiva, un tema particolarmente avvertito in letteratura concerne la verifica del nesso di interazione

[5] La corruzione è un "*totem*" in relazione al quale si esprimono, in modo singolarmente intenso, le tendenze "anti-elitarie" che connotano il populismo politico, nei suoi riverberi sul penale: S. ANASTASIA, M. ANSELMI, D. FALCINELLI, *Populismo penale. Una prospettiva italiana*, Padova, 2015, *passim*; M. DONINI, *Populismo e ragione pubblica. Il post-illuminismo penale tra lex e ius*, Modena, 2019, *passim*; si veda, inoltre, l'intenso dibattito, pubblicato in *Dir. pen. cont.,* 21.12.2016: AA. VV., *La società punitiva. Populismo, diritto penale simbolico e ruolo del penalista*. Più recente, il bel contributo di N. SELVAGGI, *Populism and Criminal Justice in Italy*, in G. DELLEDONNE, G. MARTINICO, M. MONTI, F. PACINI (a cura di), *Italian Populism and Constitutional Law,* Londra, 2020.

[6] È il fondamentale precipitato "metodologico" della riflessione roxiniana: cfr. C. ROXIN, *Sulla fondazione politico-criminale del sistema del diritto penale,* in ID., *Politica criminale e sistema del diritto penale,* Napoli, 2009, 177 ss.

[7] Su questo modello epistemologico, ci si limita ai riferimenti essenziali: F. VON LISZT, *Der Zweckgedanke im Strafrecht*, Berlino, 1905, trad. it. *La teoria dello scopo nel diritto penale,* a cura di A. A. Calvi, Milano, 1962; A. BARATTA, M. PAVARINI, *La frontiera mobile della penalità nei sistemi di controllo sociale della seconda metà del ventesimo secolo,* in *Dei delitti e delle pene,* 1, 1998, 7 ss.; più di recente, il tema è stato oggetto di una potente riflessione: M. DONINI, *La scienza penale integrale fra utopia e limiti garantistici*, in S. MOCCIA, A. CAVALIERE (a cura di), *Il modello integrato di scienza penale di fronte alle nuove questioni sociali*, Napoli, 2016, 7 ss.

tra corruzione e criminalità organizzata[8]: molteplici evidenze empiriche (nonché statistiche giudiziarie) registrano, in una pluralità di contesti, la sensibile contiguità funzionale dei due fenomeni criminali: il ricorso sistematico a pratiche corruttive rientra nel capitale strumentale di organizzazioni criminali, sostituendo o, comunque, affiancandosi, a metodi di azione (come la violenza, l'intimidazione, il controllo del territorio) che progressivamente disperdono il proprio carattere "tipizzante" (quanto meno sul piano dell'osservazione fenomenica).

La misurazione rende possibile, peraltro, istituire correlazioni significative tra determinate "dimensioni" (quantitative e qualitative) del fenomeno osservato e "grandezze economiche", sociali o, in senso lato, istituzionali. In tal senso, sono stati sviluppati indicatori complessi che indagano la relazione tra caratteristiche o fattori del *framework* "istituzionale"[9] (ad es. grado di tutela dei diritti umani; libertà di stampa; livello di sviluppo economico; grado di ineguaglianza sociale; livello di investimenti esteri; indipendenza degli organi giudiziari e via dicendo) e fenomeno corruttivo; si tratta di indici fortemente aggregati. Per questa ragione, assicurano un'ampia "comparabilità" con altri Paesi ma non sono in grado di fornire una stima diretta del livello di corruzione in un singolo contesto, limitandosi a registrare (essenzialmente sul piano qualitativo) mere correlazioni di rischio.

L'affinamento progressivo delle metodologie di "quantificazione" della corruzione risulta, soprattutto, funzionale al *policymaking*, come già osservato (cfr. *retro,* 1). L'elaborazione di politiche di prevenzione e contrasto, fondate su stime e valutazioni verificabili, è condizione funzionale per lo sviluppo di una legislazione coerente con la dimensione effettiva del fenomeno, secondo un modello di intervento *tailored*. In questo senso, misurare la corruzione significa anche monitorare l'impatto della legislazione e delle strategie di regolazione adottate con riferimento al contesto nazionale complessivo o a singoli e più specifici settori della *governance* pubblica[10].

[8] Su questo nesso v. l'imponente studio di P. GOUNEV, T. BEZLOV, *Examining the links between organised crime and corruption,* 2010*,* ad accesso libero su www.ec.europa.eu; cfr. anche i saggi in P. GOUNEV, V. RUGGIERO (a cura di), *Corruption and Organized Crime in Europe. Illegal Partnership,* Londra, 2014; per una ricognizione complessiva, v. L. HOLMES (a cura di), *Terrorism, Organised Crime and Corruption. Networks and Linkages,* Londra, 2010.

[9] A tal proposito, si rinvia agli studi di D. TREISMAN, *What have we learned about the causes of corruption form ten years of cross-national empirical research?,* in *Annual Review of Political Science,* 2007, 10, 211 ss.; L. PELLEGRINI, R. GERLAGH, *Causes of corruption: a survey of cross-country analyses and extended results,* in *Economics of Governance*, 2008, 9, 245 ss.; J. LAMBSDORFF, *Consequences and causes of corruption: what do we know from a cross-section of country?*, Discussion Paper del Dipartimento di Economics, Passau University, 2005, Maggio.

[10] Sintetizzano queste finalità della "misurazione" A. M. DURANTE MANGONI, G. TARTAGLIA POLCINI, *La diplomazia giuridica,* Napoli, 2019, spec. 161 ss.

Senonché, sebbene le finalità di un approccio oggettivo al fenomeno siano molto chiare, la definizione di metodi e strumenti per realizzare questa misurazione è straordinariamente controversa in ambito scientifico[11].

La "prima generazione" di strumenti[12] è costituita da indicatori aggregati e compositi, che utilizzano come fonte principale *experts assessment* o sondaggi inerenti la percezione del fenomeno da parte di campioni generici di popolazione ovvero campioni settoriali (*surveys perception*). A questa classe di indicatori prevalentemente soggettivi o percettivi appartengono – esemplificando – il *Transparency International Corruption Index* (CPI) e il *World Bank's Governance Indicators* (WGBI). L'utilizzo per la misurazione del livello di corruzione in un dato contesto (ad es. nazionale o settoriale) di indici soggettivi o percettivi è stato oggetto, negli ultimi anni, di un'approfondita revisione critica, sia in ambito accademico sia in seno agli organismi internazionali (a tal proposito, è ancora significativo il *Manual on corruption surveys* UNODC). Le criticità si concentrano, come intuibile, sull'«ontologica imprecisione»[13] di questi indicatori, in quanto fondati su un patrimonio informativo prevalentemente costituito dall'elaborazione della percezione, da parte dell'opinione pubblica o del campione intervistato, della presenza del fenomeno e del grado di compromissione del contesto istituzionale di riferimento. La fonte percettiva, in questo senso, non può essere impiegata come "indicatore" del reale livello di radicamento della corruzione, atteso che l'opinione pubblica e quella individuale sono soggette alla sensibile influenza di molteplici fattori[14]. Più in generale, gli indicatori percettivi non restituiscono la dimensione *reale* del fenomeno[15], poiché muovono esattamente dall'esperienza soggettiva della realtà che pretendono di indagare: cioè dalla sua rappresentazione individuale e collettiva, soggetta, come noto, a incontrollabili dinamiche di conformazione e condizionamento mediatico.

Tra i limiti epistemici dei modelli di misurazione soggettivi (cioè fondati su *percezioni,* non già sull'esperienza diretta del fenomeno[16]), si osserva, peraltro, un autentico paradosso: un maggior livello di

---

[11] Per un affresco, in una vasta letteratura, cfr. F. HEINRICH, R. HODESS, *Measuring corruption*, in A. GRAYCAR, R. G. SMITH (a cura di), *Hanbook of Global Research and Practice in Corruption,* Cheltenham (UK)-Northampton (US), 2012, 18 ss.; N. HELLER, *Defining and measuring corruption: where have we come from, where are we now, and what matters for the future?,* in R.I. ROTBERG (a cura di), *Corruption, Global Security, World Order,* Baltimore (US), 2009, 47 ss.; L. HOLMES, *Corruption. A Very Short Introduction,* Oxford, 2015, 36 ss.

[12] Propongono questa categorizzazione "generazionale" degli strumenti di misurazione della corruzione, sinora sviluppati: F. HEINRICH, R. HODESS, *Measuring corruption,* cit., 19 ss.

[13] Così A. M. DURANTE MANGONI, G. TARTAGLIA POLCINI, *La diplomazia giuridica,* cit., 163.

[14] In questi termini, tra gli altri, S. ANDERSSON, P.M. HEYWOOD, *The politics of perception: use and abuse of Transparency International's approach to measuring corruption*, in *Political Studies*, 2009, 57, 746 ss.; J. P. LYNCH, *Problems and promise of victimization surveys for cross-national research,* in *Crime and Justice,* 2006, 34. Discute il tema E. CARLONI, *Misurare la corruzione? Indicatori di corruzione e politiche di prevenzione,* in *Politica del diritto,* 2017, 3, 445 ss.

[15] Valorizza criticamente questo aspetto B. A. OLKEN, *Corruption perceptions vs. corruption reality,* in *Journal of Public Economy,* 2009, 93, 950 ss.

[16] Gli *experience indicators* sono metodi di rivelazione diretta del fenomeno che, certo, scontano limiti noti (cfr. *infra,* nel testo) e, tuttavia, sono considerati, insieme alle statistiche giudiziarie e altre fonti formali (ad es. dati provenienti da

enforcement delle politiche anti-corruzione (repressive o preventive) è correlato ad un incremento del grado di percezione collettiva e individuale di questa fenomenologia criminosa in quanto motivato dalla maggiore visibilità – *rectius strepitus fori* – del fenomeno[17].

Nonostante queste criticità, gli indici "percettivi" sono in grado di sanare quei limiti che, invece, caratterizzano gli strumenti di misurazione oggettivi, elaborati a partire dalla rivelazione dell'esperienza diretta di episodi di corruzione (in senso lato, di *maladministration*) ovvero dalle statistiche giudiziarie. Limiti, come noto, connessi alla dinamica della "cifra oscura" (*dark figure*) che caratterizza l'emersione giudiziaria degli episodi penalmente rilevanti di corruzione, in ragione della natura (necessariamente) clandestina di questo genere di attività illecita e della scarsissima propensione alla denuncia o, più in generale, al *reporting* da parte dei soggetti coinvolti in un accordo, strutturalmente fondato sul reciproco vantaggio dell'omertà[18]. Gli indici percettivi superano il problema del sensibile livello di *underreporting*, sostituendo al racconto dell'esperienza diretta la comunicazione della percezione (indiretta e sovente mediata) del fenomeno. Al contempo, in quanto elaborati per aggregazione di dati, consentono «comparazioni sia geografiche che temporali e quindi la produzione di «classifiche» e *rankings* che a loro volta sono fattori di trasformazione guidando, se non imponendo, risposte attraverso apposite politiche pubbliche»[19].

D'altronde, l'impiego di questa tipologia di indicatori per la formulazione del *country rankings* ha imponenti riverberi negativi, a livello nazionale, sul piano macroeconomico, in termini sia di ridotta attrattività del sistema Paese per gli investimenti esteri sia, direttamente, su talune variabili economiche legate, ad es., al debito pubblico. Si spiega, in tal senso, l'iniziativa – molto ambiziosa, anzitutto da un punto di vista teorico -, intrapresa dal Governo italiano nell'ambito della presidenza del G7, volta a promuovere una revisione costruttiva delle metodologie sinora applicate nella misurazione della corruzione[20].

Gli indici oggettivi (*experience-based* o giudiziari)[21], malgrado i limiti che li connotano (ora in termini di "cifra oscura", ora in termini di ridotta propensione al *reporting*, atteso che si tratta di tipici reati *victimless*),

pubbliche amministrazioni), sicuramente attendibili nella misurazione della consistenza reale del fenomeno in esame: a tal proposito, cfr. UNODOC RESEARCH, *Manual on corruption surveys,* 23 ss., dove si sottolinea il valore degli "experience-based surveys" nell'elaborazione e programmazione di "evidence-based anti-corruption policies".

[17] Su questo paradosso cfr. G. TARTAGLIA POLCINI, *Il paradosso di Trocadero,* in *Dir. pen. della globalizzazione,* 22.10.2017, 1 ss.

[18] Approfondiscono questo aspetto P. DAVIGO, G. MANNOZZI, *La corruzione in Italia. Percezione sociale e controllo penale,* Bari, 2007, 63 ss.

[19] Così E. CARLONI, *Misurare la corruzione? Indicatori di corruzione e politiche di prevenzione,* cit., 452.

[20] In merito a questa iniziativa, A. M. DURANTE MANGONI, G. TARTAGLIA POLCINI, *La diplomazia giuridica,* cit., 165 ss.

[21] Rientrano in questo ambito, ad es., l'International Crime Victim Survey (ICVS), il Transparency International's Global Corruption Barometer.

sono considerati, anche a livello internazionale[22], gli strumenti più affidabili, da un punto di vista metodologico, per la formazione di un patrimonio di dati capace orientare il *policymaking* così come le prassi operative delle agenzie pubbliche di *enforcement* (in primo luogo, la magistratura). Le indagini di vittimizzazione, soprattutto se settoriali (*context-based*: relativi a determinati ambiti funzionali della Pubblica amministrazione), sono utili per l'identificazione delle aree, delle posizioni e delle procedure maggiormente esposte a rischio così come per il monitoraggio, nel tempo, dell'impatto delle politiche anti-corruzione messe in atto[23].

Si basano su un approccio metodologico di tipo essenzialmente *oggettivo* anche gli strumenti che assumono come parametri rilevanti alcune "grandezze" misurabili, inerenti lo specifico settore di riferimento[24]. Questi parametri sono individuati in funzione di un nesso di correlazione con la fenomenologia corruttiva; si tratta, in buona sostanza, di «*proxy indicators*», cioè indici che, al variare di determinate grandezze, assunte come *benchmark*, esprimono probabili interazioni con il fenomeno in esame: ad es. la differenza tra i costi sostenuti dalla Pubblica amministrazione per l'acquisto di determinati beni o servizi e i costi di mercato; il numero di "nuove" imprese che acquisiscono contratti pubblici, in un dato segmento economico rilevante[25]. I *proxy indicators* non garantiscono risultati comparabili con altri Paesi: si tratta, infatti, di indici relativi a ben determinati e circoscritti contesti funzionali. Sono, però, utilmente impiegabili per valutare l'impatto di specifiche misure anti-corruzione applicate in uno specifico ambito (*context-based*). L'aspetto più delicato di questa categoria di indicatori concerne, come è agevole intuire, la selezione del parametro di riferimento e, in special modo, la valutazione (di ordine essenzialmente qualitativo) che è ad essa sottesa: cioè l'individuazione di un collegamento o, comunque, di un rapporto di inferenza tra una precisa dimensione del fenomeno criminale e una definita "grandezza" (se, esemplificando, un differenziale più o meno significativo tra costi per l'acquisto di un bene, sostenuto dalla P.A., e prezzo di mercato, sia correlato – e in quali termini – ad un tasso di *maladministration* nel contesto preso in esame o ad altri fattori, per così dire, non penalmente rilevanti).

Negli ultimi anni, la prospettiva di sviluppo di nuove metodologie di "misurazione" e *assessment* della corruzione è radicalmente mutata; gli sforzi della ricerca si concentrano, in special modo, nella definizione e validazione di "indici dinamici" di rischio (o indicatori predittivi), in grado di identificare, in uno

[22] Cfr. *retro,* nt. 20.
[23] Cfr. UNODOC RESEARCH, *Manual on corruption surveys,* spec. 26 ss.
[24] In tema, per una puntuale ricognizione, cfr. J. JOHNSØN, P. MASON, *The Proxy Challenge: Why bespoke proxy indicators can help solve the anti-corruption measurement problem*, U4 Brief, Bergen: U4 Anti-Corruption Resource Centre, 2013.
[25] Un modello *proxy* è proposto da M.A. GOLDEN-L. PICCI, *Proposal for a new measure of corruption, illustrated with data*, in *Economics and Politics*, 17, 1, 2005, 37-75.

specifico contesto, determinate anomalie (c.d. *red flags*), correlate (in termini di probabilità o possibilità) al verificarsi di episodi di corruzione o *maladministration*.

Lo sviluppo di questi indici di anomalia corrisponde, in realtà, al mutare progressivo delle linee programmatiche e delle strategie di contrasto al fenomeno corruttivo, le quali hanno assunto, sempre più, un orientamento spiccatamente preventivo. Un aspetto particolarmente significativo, che merita di essere segnalato con riguardo a questi indici dinamici è rappresentato dalle potenzialità di implementazione. Non si tratta, infatti, di strumenti di mera "misurazione": piuttosto, gli indici di anomalia rappresentano, anche nel loro insieme, un arsenale operativo attraverso il quale e le pubbliche amministrazioni e gli operatori economici privati sono in grado di perfezionare e approfondire i rispettivi modelli di *risk assessment* e *risk managament*, al fine di garantire più elevati *standards* preventivi nei differenti contesti funzionali di riferimento. Allo stesso tempo, gli indici di rischio potrebbero essere impiegati come strumenti per orientare efficacemente le attività istruttorie e di investigazione condotte dalle agenzie di *enforcement*. La prospettiva di sviluppo ulteriore è rappresentata, pertanto, dalla formalizzazione di autentici cataloghi di indicatori, *context-oriented*, empiricamente validati e messi a disposizione dei soggetti che concorrono attivamente (ivi comprese le imprese) nel sistema integrato di prevenzione del fenomeno corruttivo.

3. Proposta di misurazione e *assessment* del fenomeno corruttivo attraverso l'analisi automatica dei testi delle sentenze

L'analisi del "dato giudiziario" (cioè dei molteplici episodi di corruzione, come ricostruiti e accertati nel corso dei gradi di giudizio) consente, anzitutto, di *estrarre* elementi significativi del contesto e dei tratti caratterizzanti la variegata fenomenologia della dinamica corruttiva.

L'affidabilità della fonte giudiziaria (sentenze) consente, inoltre, di isolare le *condizioni di contesto* che, nei vari casi, possono aver agevolato o, comunque, facilitato il verificarsi del fatto corruttivo: l'estrazione di questi elementi è fondamentale per l'individuazione di *proxy indicators,* fortemente contestualizzati.

Conditio sine qua non è il ricorso a tecniche automatiche di testi in linguaggio naturale, per poter trattare non solo un numero rilevante di sentenze ma soprattutto informazioni estratte con gli stessi criteri ed in modo automatico[26]. Le "sentenze", una volta trattate mediante *text mining*[27], possono restituire molteplici informazioni su luoghi, modalità, interazioni tra soggetti e condizioni del contesto economico e

[26] Cfr. S. BOLASCO, *L'analisi automatica dei testi: fare ricerca con il text mining*, 2013; L. LEBART-A. SALEM, *Exploring Textual Data*, 2011; C.C. AGGARWAL-C.X. ZHAI, (a cura di), *Mining text data*. Springer Science & Business Media, 2012.

[27] La metodologia impiegata viene descritta in dettaglio nell'articolo M.F. ROMANO-A. BALDASSARINI-P. PAVONE-G. MORGANTE-G. DI VETTA, *Linkage of statistical archives and judicial decisions. Methodology and results on corruption* (attualmente sottomesso in *review* per la rivista scientifica *Social Indicators Research*).

territoriale in cui si è verificato il fatto giudizialmente accertato. Interrogando il "dato giudiziario", è quindi possibile procedere ad una topografia del fenomeno corruttivo "emerso", non soltanto limitata all'indicazione territoriale (dove maggiormente si verificano fatti di corruzione), bensì estesa alla "geografia" dei contesti funzionali (in quali *specifici* settori o procedure della Pubblica amministrazione si concentra maggiormente il fenomeno).

In questa prospettiva, il *text mining,* applicato come metodo di analisi dei dati giudiziari, può rappresentare un efficace strumento per l'elaborazione di strategie e politiche di contrasto, anche fortemente settorializzate. Nonostante il limite della c.d. cifra oscura, che affligge tutti gli indicatori oggettivi-giudiziari, la metodologia proposta non risente in alcun modo delle criticità che invece emergono con riferimento agli indici soggettivi.

La straordinaria potenzialità delle tecniche di estrazione "testuale" da una fonte giurisprudenziale, assicura, peraltro, la possibilità di "validare", alla luce del dato reale, cristallizzato nelle sentenze, i modelli esplicativi del fenomeno proposti in letteratura (così, ad es., sono suscettibili di verifica, attraverso questo strumento, talune acquisizioni qualitative circa le modalità di manifestazione in forma "associativa" o "plurisoggettiva" della corruzione) consentendo di valorizzare il richiamato modello dell'*evidence-based policymaking* a beneficio della maggiore efficienza delle strategie di contrasto .

In questo articolo, gli autori illustrano i risultati preliminari ottenuti dalla applicazione di una metodologia automatica di interazione tra dati e informazioni[28] su un numero limitato di sentenze emesse dalla Corte di Cassazione[29].

Nel paragrafo 4 vengono brevemente descritti i dati e gli strumenti utilizzati, nel paragrafo 5 si riportano alcuni risultati ottenuti su di un numero limitato, ma non irrilevante, di documenti; infine nel paragrafo 6 si discuteranno i punti di forza e di debolezza del metodo proposto - evidenziandone sia le difficoltà operative ancora da superare, sia le potenzialità conoscitive - sia alcuni possibili sviluppi del percorso di ricerca[30], con l'ambizione di dimostrare come l'interazione tra documenti della PA ed altri archivi di dati possa efficacemente applicarsi in svariati contesti e potenzialmente riferibili ad ambiti che vanno dalla ricerca all'applicazione pratica.

[28] Si tratta, come viene descritto in dettaglio nell'articolo M.F. ROMANO-A. BALDASSARINI-P. PAVONE-G. MORGANTE-G. DI VETTA, *op. cit.,* di una sequenza di analisi automatiche che, nel loro complesso, tracciano una metodologia riproponibile su altri insiemi di sentenze, con limiti dettati solo dalle capacità informatiche della piattaforma utilizzata.

[29] La scelta di utilizzare solo le sentenze emesse dalla Corte di Cassazione deriva dalla non disponibilità di sentenze per gradi precedenti di giudizio. In verità la banca dati DeJure permette di avere i testi delle sentenze di gradi precedenti di giudizio, e rimandiamo a M. F. ROMANO, *op. cit.,* 124 ss. per ulteriori approfondimenti.

[30] L'idea di utilizzare le sentenze come fonte di dati e informazioni da integrare progressivamente con altre fonti (ufficiali e non) era una possibile risposta pe la quantificazione economica delle attività illegali (si veda, in tema, G. REY, *op. cit.*).

4. L'approccio empirico: dati, strumenti e metodi

Le sentenze della magistratura sono redatte come documenti di testo, interpretabili dai lettori a cui sono destinate, per cui si è dovuto immaginare, costruire e testare una metodologia più complessa per identificare e classificare/quantificare in modo non ambiguo le informazioni contenute nelle sentenze considerate "utili" per misurare l'evento criminoso, in questo caso quello corruttivo. Merita sottolineare che l'esito del processo e le "statuizioni" indicate non rientrano nel nostro interesse, dedicato esclusivamente a ricostruire gli eventi corruttivi. E a questo scopo l'accertamento di costanti linguistiche permette di svelare le connessioni tra eventi e soggetti coinvolti in una vicenda corruttiva, distinguendo tra persone fisiche (imputati) e giuridiche (enti pubblici e/o aziende), nel loro ruolo (imputati, parti offese, parti civili), individuare il luogo geografico dell'evento, il periodo temporale di riferimento e l'eventuale valore economico della corruzione.

La finalità dell'analisi lessico-testuale svolta sui testi delle sentenze è appunto quella di estrarre le informazioni "utili" per mappare il fenomeno criminoso oggetto d'interesse. La disponibilità dei testi completi delle sentenze, non oscurati da Omissis, è possibile solo per quelle emesse dalla Corte di Cassazione. La descrizione dei fatti può quindi essere meno dettagliata.

Descriviamo in questa sede alcuni risultati della sua applicazione su un insieme di circa 700 sentenze (684 sentenze[31] per la precisione) emesse dalla Corte di Cassazione tra il 2015 ed il gennaio 2020[32].

Occorre sottolineare che l'insieme delle sentenze non è da considerarsi un campione rappresentativo del totale delle sentenze, e che anche i criteri della loro selezione possono apparire arbitrari[33].

Le sentenze utilizzate per le analisi sono molto variabili quanto a numero di parole: la più breve è stampabile in meno di 1 pagina, mentre per la più lunga ne occorrerebbero quasi 270; tuttavia questa variabilità può essere considerata un aspetto positivo dell'esercizio proposto: una ricchezza lessicale omogenea tra le sentenze potrebbe influenzare la capacità di estrazione dei contenuti informativi e quindi potrebbe non costituire un'adeguata esemplificazione della metodologia proposta[34].

[31] Non sono state prese in considerazione 12 sentenze, considerate "anomale" per le finalità del presente lavoro: si tratta, infatti, di decisioni della Corte di Cassazione che hanno ad oggetto questioni di ordine meramente processuale.

[32] Si rimanda a M. F. ROMANO, *op. cit.,* i lettori interessati ad altri risultati, oltre che per una dettagliata presentazione della metodologia.

[33] La selezione delle sentenze è stata effettuata ricercando la presenza dei termini "corruzione" o "concussione"; le sentenze selezionate sono state comunque oggetto di numerose verifiche a campione.

[34] In linea generale, le sentenze troppo brevi possono contenere meno riferimenti dettagliati agli eventi criminosi, dando quindi meno spazio alla descrizione dei fatti; tuttavia, anche in questi casi, la presenza di persone giuridiche (aziende private ed enti pubblici) come parti civili, il loro nome comparirà anche in sentenze molto brevi.

I testi importati in TaLTaC costituiscono un Vocabolario di 75.436 differenti parole[35] (FG o MW) per un totale di 3.135.050 parole complessive.

La Tabella 1 presenta la ripartizione delle sentenze per presenza (congiunta) delle parole chiave (concussione e corruzione) e la Tabella 2 ne illustra la ripartizione per anno di emissione.

Tabella 1 Classificazione delle sentenze per presenza delle parole chiave

Presenza della parola chiave	Totale	
	num	%col
Concussione	60	8,7
Corruzione	393	57,5
Corruzione E Concussione	231	33,8
Totale	684	100,0

Tabella 2 Classificazione delle sentenze per anno di emissione

Anno di emissione della sentenza	Totale	
	num	%col
2015	106	15,5
2016	146	21,3
2017	148	21,6
2018	151	22,1
2019	123	18,0
2020	10	1,5
Totale	684	100,0

Dopo aver trattato le sentenze con il software TaLTaC[36], abbiamo potuto "taggare" le sequenze di parole con lo stesso significato semantico e poi conteggiarne la loro presenza o sequenza di esse per ciascuna sentenza.

Per ogni sentenza ne possiamo descrivere caratteristiche giudiziarie (presenza del D.lgs. n. 231/2001, sede giudiziaria iniziale, presenza di parti civili distinte per tipo (persone fisiche, enti pubblici, aziende private), ma anche la presenza di organizzazioni criminali (con più *tag* semantici), del ruolo professionale e/ o politico delle persone fisiche coinvolte, identificandone anche il ruolo nel processo (parti civili o

[35] Il termine "parola" (meglio definita come Forma Grafica oFG) indica non solo le parole singole, ma anche le unità lessicali composte da più parole (amministratore delegato, carta di credito, assessorato ai lavori pubblici, per fare qualche esempio, oppure locuzioni grammaticali tipo "di fronte a", "alla luce di", e similari. Per ciascuna Forma Grafica è disponibile un identificativo grammaticale (verbo, aggettivo, nome proprio, etc) ed anche il lemma.

[36] "TaLTaC è un software per l'analisi di una collezione di testi, finalizzata a descrivere e interpretare il suo contenuto e/o alcune sue proprietà. L'impianto generale adottato nel programma è noto in letteratura come "approccio lessicometrico" in quanto consente lo studio diretto di qualsiasi insieme di dati espressi in linguaggio naturale, da documenti a interviste, da rassegne stampa a messaggi, secondo i principi della "statistica testuale" e della Analisi Automatica dei Testi." www.taltac.com. Cfr. S. BOLASCO-G. DE GASPERIS, *TaLTaC 3.0. A Multi-level Web Platform for Textual Big Data in the Social Sciences*, 2017.

coinvolti) sia per le persone fisiche che per quelle giuridiche. Le imprese identificate nelle sentenze sono state distinte in funzione del ruolo processuale e, quindi, sostanziale rivestito nella fattispecie concreta; a tal fine, è stata adottata una definizione stipulativa di «impresa coinvolta», rilevante ai fini della ricerca: si tratta di imprese incolpate ai sensi del D.lgs. n. 231/2001 ovvero di imprese i cui amministratori e/o dipendenti sono a loro volta coinvolti nell'episodio corruttivo, in quanto imputati.

A titolo esemplificativo, riportiamo di seguito il risultato dell'applicazione di 5 "tagging semantici" per individuare la presenza di organizzazioni criminali:

Tag semantico	**Descrizione**	**Unità lessicali**	**Numero di unità lessicali identificate**	**Esempi**
OrgCrim	nome dell'organizzazione criminale	98	1.555	associazione mafiosa ndrangheta cosca
AmbOC	ambiente di tipo criminale	63	720	natura mafiosa connotazione mafiosa camorristico appoggio mafioso contesto mafioso
AttivOC	attività dell'organizzazione criminale	38	111	attuale operatività della cosca attività ausiliatrici della cosca cosca di monopolio sul settore degli appalti boschivi esteriorizzazione dei poteri di tipo mafioso
MetodiOC	metodi criminali	24	52	modalità mafiose protezione mafiosa metodologia di tipo mafioso metodi tipicamente mafiosi modello di stampo mafioso metodica delinquenziale delle mafie storiche metodi camorristici mafia silente
personaOC	ruolo di persone nell'OC	23	40	capomafia capo mafia esponenti mafiosi referenti della cosca intraneo alla cosca esponente mafioso capo della cosca capicosca

Per ogni sentenza sono state create più variabili (in questo caso 5) contenenti il numero di tag semantici identificati all'interno del testo; se il valore complessivo dei tag semantici era superiore all'unità, la sentenza viene complessivamente classificata come "presenza di organizzazioni criminali", come si vedrà dalle tabelle presentate nel paragrafo successivo.

Per le persone giuridiche identificate come aziende private sono disponibili ulteriori dati e informazioni, frutto di un linkage con altri database[37], e inoltre è stato possibile classificare ogni sentenza per la presenza/assenza di aziende private.

5. Risultati

Riportiamo innanzitutto alcune tabelle che hanno l'obiettivo di far emergere la capacità di descrizione delle sentenze dopo il "trattamento" dei testi con le tecniche di text mining: i risultati sono presentati per tre insiemi di sentenze: a) il totale (n=684), b) le sentenze con riferimenti alla 231 (n=31), 3) le sentenze che contengono traccia della presenza di organizzazioni criminali (n=115).

Tabella 3 Sentenze per origine giudiziaria e presenza organizzazioni criminali, DLG231

Sede di origine giudiziaria	Totale sentenze		Presenza OC		Presenza 231	
	num	%col	num	%col	num	%col
NordOv	119	17,4	6	7,7	12	33,3
NordEs	52	7,6	1	1,3	3	8,3
Centro	156	22,8	8	10,3	9	25,0
Sud	333	48,7	63	80,8	11	30,6
n.d.	24	3,5		0,0	1	2,8
Totale	684	100,0	78	100,0	36	100,0

Tabella 4 Sentenze per presenza organizzazioni criminali e DLG231

organizzazioni criminali	Totale sentenze		Presenza 231	
	num	%col	num	%col
non presente	573	88,4	33	91,7
presente	75	11,6	3	8,3
Tot	648	100,0	36	100,0

Nell'80% delle sentenze sono stati identificati nomi di aziende, e la loro distribuzione per sede di origine giudiziaria è riportato nella Tabella 5:

Tabella 5 Sentenze per sede di origine giudiziaria e presenza di aziende

Sede di origine giudiziaria	Sentenze						Tot		
	Con Aziende coinvolte			Senza Aziende coinvolte					
	num	%col	%riga	num	%col	%riga	num	%col	%riga
NordOvest	103	19,0	86,6	16	11,2	13,4	119	17,4	100,0
NordEst	39	7,2	75,0	13	9,1	25,0	52	7,6	100,0
Centro	118	21,8	75,6	38	26,6	24,4	156	22,8	100,0
Sud	266	49,2	79,9	67	46,9	20,1	333	48,7	100,0
n.d.	15	2,8	62,5	9	6,3	37,5	24	3,5	100,0
Tot	541	100,0	79,1	143	100,0	20,9	684	100,0	100,0

[37] Il linkage, la cui metodologia è descritta in M. F ROMANO, *op. cit.*, è stato effettuato con l'archivio Orbis (fonte non ufficiale) e Asia (fonte statistica ufficiale ISTAT) ed ha permesso l'aggiunta di dati descrittivi (sede legale dell'azienda, forma giuridica, status) e quantitativi (numero di addetti, anno di costituzione) attendibili in quanto provenienti da una fonte statistica ufficiale (ISTAT).

Per le 541 sentenze in cui ci sono aziende coinvolte, solo in 28 il loro ruolo nel processo è quello di parte civile: la loro distribuzione per sede di origine giudiziaria si legge nella Tab. 6.

La Tabella 7 mostra che quando è presente nella sentenza un riferimento ad organizzazioni criminali è quasi sempre coinvolta un'azienda (94,9% nella tabella 5).

Tabella 6 Sentenze con presenza di aziende per sede di origine giudiziaria e ruolo nel processo

Sede_origine giudiziaria	con ruolo						Totale		
	coinvolte[38]			parte civile					
	num	%col	%rig	num	%col	%rig	num	%col	%rig
NordOv	94	18,3	91,3	9	32,1	8,7	103	17,4	100,0
NordEs	36	7,0	92,3	3	10,7	7,7	39	7,6	100,0
Centro	111	21,6	94,1	7	25,0	5,9	118	22,8	100,0
Sud	258	50,3	97,0	8	28,6	3,0	266	48,7	100,0
n.d.	10	1,9	100,0				10	2,8	100,0
misto	4	0,8	80,0	1	3,6	20,0	5	0,7	100,0
Totale	513	100,0	94,8	28	100,0	5,2	541	100,0	100,0

Tabella 7 Sentenze per presenza di organizzazioni criminali e presenza di aziende

Organizzazioni criminali	Sentenze con aziende						Totale		
	coinvolte			non coinvolte					
	num	%col	%rig	num	%col	%rig	num	%col	%rig
non presenti	467	86,3	77,1	139	97,2	22,9	606	88,6	100,0
presenti	74	13,7	94,9	4	2,8	5,1	78	11,4	100,0
Totale	541	100,0	79,1	143	100,0	20,9	684	100,0	100,0

Rimangono comunque non poche (139) le sentenze in cui non sono state identificate né imprese né organizzazioni criminali, e il tagging semantico effettuato ci consente di individuare i sottogruppi di sentenze per presenza / assenza dei risultati dell'analisi lessicale.

Nella Tabella 8 riportiamo la descrizione dei gruppi di sentenze simili per le variabili individuate dall'analisi lessico-testuale, limitando la lista ai soli 12 gruppi con un numero di almeno 25 sentenze per un totale complessivo di 436 sentenze. Si noti che i gruppi sono ulteriormente aggregabili se non si tiene conto della sede di origine giuridica.

[38] Per la definizione di «impresa coinvolta», ai fini della presente ricerca, v. *retro,* nel testo.

Tabella 8 Gruppi di Sentenze per presenza di categorie semantiche

gruppo	Imprese	Sede_origine	PAstrutture	TagOrgCrim	RuoloPol	N	N1
A	*presente*	Sud	non presente	non presente	*presente*	78	147
	presente	NordOv	non presente	non presente	*presente*	26	
	presente	Centro	non presente	non presente	*presente*	43	
B	*presente*	Sud	*presente*	non presente	*presente*	52	106
	presente	Centro	*presente*	non presente	*presente*	26	
	presente	NordOv	*presente*	non presente	*presente*	28	
C	*presente*	Sud	non presente	non presente	non presente	46	97
	presente	NordOv	non presente	non presente	non presente	26	
	presente	Centro	non presente	non presente	non presente	25	
D	*presente*	Sud	non presente	presente	*presente*	30	30
E	*presente*	Sud	*presente*	non presente	non presente	29	29
F	non presente	Sud	non presente	non presente	non presente	27	27

Non meraviglia il constatare che tra questi 12 gruppi ben 11 siano composti da sentenze in cui ci siano imprese coinvolte, dato che la loro presenza è stata individuata in 541 sentenze (cfr. Tab.5); inoltre la presenza di organizzazioni criminali si trova solo nel gruppo E (30 sentenze).

Si noti che i gruppi sono ulteriormente aggregabili se non si tiene conto della sede di origine giuridica: il gruppo A è costituito da 147 sentenze in cui sono state identificate aziende, sono presenti persone con ruoli politici (a tutti i livelli) e mancano riferimenti a organizzazioni criminali e a strutture della Pubblica Amministrazione.

Tabella 9 Imprese (coinvolte e parti civili) per settore economico: numero e media di addetti

Settore economico (classificazione ATECO)	**Addetti**		
	N	**Somma**	**Media**
G - Commercio ingrosso e dettaglio	248	55.752	225
F - Costruzioni	149	5.134	34
C - Attività Manifatturiere	149	18.476	124
I - Servizi di alloggio e di ristorazione	80	23.056	288
L - Attività Immobiliari	79	6.846	87
M - Attività Professionali, Scientifiche e Tecniche	59	3.406	58
E - Fornitura di Acqua; Reti Fognarie; Att Gest Rifiuti e Risanam	43	10.761	250
N - Noleggio, Agenzie Di Viaggio, Serv Supporto alle imprese	39	28.773	738
H - Trasporto e Magazzinaggio	35	1.474	42
Q - Sanita e Assistenza Sociale	22	8.535	388
J - Servizi di Informazione e Comunicazione	15	1.729	115
S - Altre Attività di Servizi	14	1.698	121
K - Attività Finanziarie e Assicurative	14	73	5
D - Fornitura di Energia	11	6.071	552
B - Estrazione di Minerali	9	116.847	12.983
R - Att Artistiche, Sportive, Intrattenimento e Divertimento	9	355	39
Totale	975	288.986	296

Tabella 10 Imprese (coinvolte e parti civili) per forma giuridica: numero e media di addetti

forma giuridica	N	Somma addetti	Media addetti
Società di cui SRL	602	35.308	59
Società di cui SPA	195	171.205	878
Società di cui SRL con un unico socio	81	7.293	90
Società di persone	43	53.126	1.235
Società cooperative	42	21.303	507
Consorzio	12	751	63
Totale	975	288.986	296

Per quanto riguarda la distribuzione geografica delle imprese per sede legale[39] (Tabella 11) l'elevato valore medio di addetti delle imprese con sede locale nel Centro Italia dipende dalla presenza di alcune grandi imprese la cui sede legale è nella Capitale.

Tabella 11 Imprese (coinvolte e parti civili) per circoscrizione della sede legale dell'impresa: numero e media di addetti

CIRCOSCRIZIONE	NUM IMPRESE	SOMMA ADDETTI	MEDIA ADDETTI
NORD OVEST	260	67.626	260,1
NORD EST	198	46.202	233,3
CENTRO	262	146.565	559,4
SUD	255	28.593	112,1
TOTALE	975	288.986	296,4

Il confronto tra sede di origine del percorso giudiziario e sede legale dell'azienda permette di evidenziare aspetti molto interessanti (Tabella 12).

Tabella 12 Imprese coinvolte nelle sentenze per sede legale dell'azienda e sede di origine giudiziaria

Sede origine giudiziaria	*sede impresa*				*Totale*
	NOVEST	NEST	CENTRO	SUD	
NOvest	102	53	44	37	236
NEst	11	28	15	14	68
Centro	50	39	97	29	215
Sud	93	77	104	173	447
Totale	256	197	260	253	966

La prima informazione che emerge è che le aziende sono spesso coinvolte in sentenze con origine territoriale *differente* da quella della loro sede legale: 400 aziende (41,4%) delle 966 totali (individuate) sono coinvolte all'interno della propria area, mentre il restante 59% (566 aziende) è coinvolto in sentenze che hanno origine in altre aree; inoltre si evidenzia uno sbilanciamento territoriale: infatti, una lettura più

[39] Si è scelto di raggruppare la sede legale per circoscrizione, dal momento che le imprese individuate hanno sede legale in quasi tutte le regioni italiane (sole eccezioni: Val d'Aosta e Molise).

attenta evidenzia che la propensione ad essere coinvolte fuori area è maggiore soprattutto per le imprese con sede legale nel Nord Est (85,8%) rispetto alle aziende del Sud (31.6%) (Tabella 13).

Tabella 13 Aziende coinvolte in sentenze con sede di origine giudiziaria diversa dalla sede legale dell'azienda

	NOVEST	NEST	CENTRO	SUD
Num aziende fuori area	154	169	163	80
Totale aziende	256	197	260	253
% fuori area	60,2	85,8	62,7	31,6

Concentrando l'attenzione sulla sede di origine del processo (Tabella 14), si nota un differente comportamento tra le sentenze di origine Nord (sia Ovest che Est) e quelle del Centro e Sud: nel primo caso tra il 57 ed il 59% delle aziende ha sede legale in altre aree, mentre per il Centro e Sud la percentuale sale all'82% circa.

Tabella 14 Aziende coinvolte per sede di origine giudiziaria diversa dalla sede legale dell'azienda

Sede origine giudiziaria	*Aziende da altre aree*	*Totale aziende*	*%*
NordOvest	134	236	56,8
NordEst	40	68	58,8
Centro	176	215	81,9
Sud	370	447	82,8

Il dato geografico che si trae da queste preliminari elaborazioni è potenzialmente molto significativo: sembrano emergere, infatti, peculiari movimenti di "migrazione" Nord(E)-Sud. Tale elemento dovrà essere oggetto di ulteriori approfondimenti e verifiche, anche al fine di accertarne la consistenza e individuarne le spiegazioni sostanziali.

Nelle tabelle seguenti viene invece riportato il numero di sentenze in cui siano coinvolte persone con ruoli politici (dal livello comunale in su). Le sentenze sono suddivise per presenza di organizzazioni criminali, poi per presenza di aziende e infine per sede di origine giuridica.

Presenza di organizzazioni criminali	**Persone con ruoli politici**						**Tot**		
	coinvolte			**non coinvolte**					
	num	**%col**	**%rig**	**num**	**%col**	**%rig**	**num**	**%col**	**%rig**
non presente	325	80,6	57,1	244	86,8	42,9	569	83,2	100,0
presente	78	19,4	67,8	37	13,2	32,2	115	16,8	100
Tot	403	100,0	58,9	281	100,0	41,1	684	100,0	100,0

Aziende	Persone con ruoli politici						Tot		
	coinvolte			non coinvolte					
	num	%col	%rig	num	%col	%rig	num	%col	%rig
presenti	344	85,4	63,6	197	70,1	36,4	541	79,1	100,0
non presenti	59	14,6	41,3	84	29,9	58,7	143	20,9	100,0
Tot	403	100,0	58,9	281	100,0	41,1	684	100,0	100,0

Sede_origine	Persone con ruoli politici						Tot		
	coinvolte			non coinvolte					
	num	%col	%rig	num	%col	%rig	num	%col	%rig
NordOvest	61	15,1	51,3	58	20,6	48,7	119	17,4	100,0
NordEst	30	7,4	57,7	22	7,8	42,3	52	7,6	100,0
Centro	95	23,6	60,9	61	21,7	39,1	156	22,8	100,0
Sud	206	51,1	61,9	127	45,2	38,1	333	48,7	100,0
n.d.	8	2,0	42,1	11	3,9	57,9	19	2,8	100,0
misto	3	0,7	60,0	2	0,7	40,0	5	0,7	100,0
Tot	403	100,0	58,9	281	100,0	41,1	684	100,0	100,0

6. Discussione e potenziali sviluppi della ricerca

L'applicazione del *text mining* alla fonte giudiziaria può realmente rappresentare un metodo di misurazione in grado di fornire una rappresentazione oggettiva del fenomeno corruttivo, così come di altre fenomenologie criminali. Oggettività e verificabilità dei risultati derivano dalla natura del metodo applicato e dal patrimonio di informazioni che ne costituisce la base conoscitiva: la fonte giudiziaria. La nostra proposta si inserisce, in tal senso, nel più ampio dibattito in ordine alla definizione e validazione di strumenti di misurazione e indicatori del fenomeno corruttivo[40].

Lo strumento messo a punto ha vaste potenzialità applicative anzitutto nell'ottica di perfezionare la caratterizzazione quantitativa e, soprattutto, qualitativa dell'evento in senso lato corruttivo, così come cristallizzato nella fonte giudiziaria. Nel corso del presente articolo, sono state già illustrate alcune applicazioni del *text mining* al fine di osservare precisi profili "dimensionali" della fenomenologia in esame. In particolare:

i. *profilo dell'impresa coinvolta nell'evento corruttivo*: l'analisi statistico-testuale – come dimostrato - consente di trarre dalle fonti giudiziarie, in modo automatico e attraverso il linkage con i registri riconosciuti, informazioni relative al numero di dipendenti, al volume di affari e al fatturato, nonché al settore economico di riferimento (cfr. Tabella 9), alla composizione e alla forma societaria (Tabella 10), così come all'eventuale partecipazione dell'impresa in aggregati societari stabili (gruppi di imprese) o non strutturati (associazioni temporanee di imprese).

[40] Cfr. *retro,* par. 1.

Queste informazioni, una volta aggregate, sono necessarie per verificare alcune ipotesi significative, formulate in letteratura, come ad es.: i) se sussiste una correlazione tra maggior frequenza di eventi corruttivi e settore economico di riferimento delle imprese coinvolte; ii) se è empiricamente fondata la relazione tra dimensioni e caratteri dell'impresa (es. forma societaria; composizione della proprietà azionaria) e dimensioni della pubblica amministrazione coinvolta o, comunque, del settore di transazione pubblico-privato interessato dall'evento corruttivo: in tal senso, l'applicazione della metodologia in esame potrebbe confermare o smentire un rapporto tra dimensioni dell'impresa coinvolta e caratteristiche e/o tipologia dell'evento corruttivo (sussiste una correlazione positiva tra episodi di *grand corruption* e profili dimensionali dei soggetti coinvolti?; le società di capitali, caratterizzate da una proprietà diffusa, sono più esposte al rischio che l'attività del *management* degeneri in episodi corruttivi? È possibile enucleare, allora, «indici di rischio» relativi ai caratteri delle imprese coinvolte?).

ii. profilo geografico dell'evento corruttivo: i dati che emergono dalle fonti giudiziarie sono estremamente rilevanti per realizzare un'indagine topografica sul fenomeno corruttivo (in senso lato). Il *text mining* consente di elaborare automaticamente un complesso di informazioni relative non soltanto – come dimostrato – alla *localizzazione* dell'episodio illegale ma anche di intersecare questo dato con le informazioni "geografiche" concernenti le imprese coinvolte (cfr. Tabelle 11-14). L'interazione tra queste informazioni, secondo il gruppo di ricerca, è fondamentale per approfondire la dinamica geografica del fenomeno corruttivo e dei soggetti in esso coinvolti.

I risultati ottenuti incoraggiano il gruppo di ricerca a proseguire l'attività, ampliando il numero delle sentenze analizzate e migliorando le fasi della categorizzazione semantica, base per il *text mining*. La collaborazione di esperti del linguaggio giuridico ed economico permetterà di predisporre un lessico economico per migliorare l'individuazione e la caratterizzazione delle imprese, e di identificare termini ed espressioni peculiari del linguaggio giuridico per chiarire il ruolo sostanziale e processuale dei soggetti coinvolti. La presenza di ricercatori ISTAT nel gruppo di ricerca ha già consentito di realizzare il linkage con il registro ASIA, ed è presupposto per ulteriori approfondimenti[41].

In questa prospettiva, sono ancora molti i "profili" dimensionali del fenomeno in relazione ai quali sperimentare le potenzialità del metodo proposto; soltanto a titolo di esempio:

i. profilo della pubblica amministrazione interessata dall'evento corruttivo: l'estrazione automatica dalle fonti giudiziarie di informazioni relative alle "pubbliche amministrazioni" coinvolte nell'episodio criminoso risulta essenziale per operare un'ulteriore caratterizzazione dell'evento corruttivo, ad esempio ponendo in correlazione il livello "territoriale" o "funzionale" dell'amministrazione interessata, la localizzazione

[41] Nell'articolo M.F. ROMANO-A. BALDASSARINI-P. PAVONE-G. MORGANTE-G. DI VETTA, *op. cit.,* gli Autori si soffermano maggiormente sulle possibilità emerse dal linkage con altri archivi / database gestiti da ISTAT.

geografica e le modalità del fenomeno corruttivo in considerazione; in tal modo, è anche possibile verificare il fondamento empirico del rilievo, ricorrente in letteratura, secondo il quale il fatto corruttivo assume determinati caratteri e modalità soprattutto in funzione del livello cui si colloca l'amministrazione pubblica cui appartiene il pubblico agente "infedele" (si tratta, in buona sostanza, del rilievo su cui si fonda la distinzione qualitativa, molto diffusa, tra *grand corruption* e *petty corruption*).

ii. profilo delle modalità dell'evento corruttivo: è il versante più ambizioso e complesso del programma di ricerca, le cui linee essenziali sono state presentate in questo articolo. Lo strumento di analisi testuale assicura l'estrazione di molteplici dati concernenti le modalità realizzative dei fatti di corruzione (in senso lato), accertati nelle fonti giudiziarie. Come illustrato, queste informazioni sono poi oggetto di un'elaborazione che ne assicura l'aggregazione; e in tal modo si possono trarre elementi significativi sul piano dell'osservazione *qualitativa* del fenomeno corruttivo (come si estrinseca empiricamente; quali sono le modalità realizzative ricorrenti, per come "descritte" e "qualificate" nella decisione giudiziaria). Più nel dettaglio, questo tipo di applicazione dell'analisi testuale sulla fonte giudiziaria è particolarmente utile per verificare anche la dimensione quantitativa di quelle forme di manifestazione del fenomeno corruttivo, ampiamente indagate in letteratura, come la «corruzione sistemica», la «corruzione ambientale» e via dicendo. Al contempo, la metodologia in esame consente di approfondire i nessi empirici che intercorrono tra i reati di corruzione e altri illeciti penali, come quelli tributari (specialmente, artt. 2 e 3 D.lgs. n. 74/2000) o quelli associativi. A quest'ultimo proposito, il tema – vivacemente discusso in dottrina – di un'interazione dinamica, sempre più significativa, tra reati di corruzione e criminalità organizzata (anche di tipo mafioso) può essere oggetto di un percorso di verifica oggettivo e attendibile, proprio applicando il metodo di analisi testuale prospettato alla fonte giudiziaria.

Gli sviluppi futuri del lavoro di ricerca possono quindi essere molteplici. Le sue applicazioni, naturalmente, dipendono in larga misura dalla natura dei fruitori del metodo, dal contesto e dalle finalità operative di impiego. In questo articolo, a ben vedere, ci si è limitati ad offrire una prima panoramica circa le potenzialità di questa metodologia sul piano dell'osservazione scientifica del fenomeno in esame. È intuibile che il metodo proposto possa rappresentare, se sviluppato in questo senso, un utile strumento anche per le attività di *enforcement,* in primo luogo per gli organi inquirenti. Un'applicazione della metodologia sulle sentenze emesse nei precedenti gradi di giudizio permetterebbe di disporre di un maggiore "contenuto informativo" sui fenomeni corruttivi (ma anche delle attività / eventi economici connessi); e, inoltre, porterebbero ad analizzare eventi o atti illeciti effettuati in periodi più vicini al tempo di analisi. A queste sentenze è, tuttavia, possibile accedere per il solo tramite del personale interno delle strutture giudiziarie o attraverso convenzioni specifiche.

Inoltre, il risultato più promettente ci appare l'incremento conoscitivo ottenuto dalla interazione tra più fonti di dati. Sono molte le altre fonti di informazione e basi di dati che potrebbero essere prese in esame: su eventi illegali in materia di corruzione l'archivio della Corte dei Conti[42] ed il database AVCP, gestito da ANAC. Il linkage delle sentenze può essere effettuato (oltre che con il Registro ASIA ed il database Orbis come per il presente lavoro) con altre fonti pubbliche ufficiali: si pensi al Censimento permanente delle istituzioni pubbliche e al Registro ASIA delle Amministrazioni Pubbliche, per estendere le analisi al settore pubblico.

In questa ottica, i risultati del linkage con altri database (dati oggettivi e informazioni sul fenomeno corruttivo identificati nelle sentenze) potranno definire alcuni indicatori di rischio, fondamentali per accrescere la consapevolezza sulle peculiarità dei territori, dei settori economici e la tipologia di soggetti coinvolti.

7. Appendice: le fonti dei dati

7.1. Le sentenze della Corte di Cassazione: il portale Italgiure Web

I documenti emessi dalla Corte di Cassazione sono disponibili da un decennio sul portale Italgiure per interrogazioni su uno o più più archivi di dati. Gli archivi sono aggiornati in modo continuo, e sono visualizzabili i 6 anni precedenti all'anno di accesso

Dal sito ***www.italgiure.giustizia.it/sncass/*** si possono selezionare i documenti attraverso più filtri (anno, sentenze o ordinanze, sezione, civile o penale), e chiedere inoltre di limitare la selezione ai soli documenti che contengano parole o riferimenti ad articoli di legge contenuti nel documento. L'accesso è libero ed è possibile scaricare i documenti selezionati sia in formato PDF sia nella loro trasformazione in testo[43].

Per le esigenze del gruppo di ricerca, attraverso interrogazioni sul portale sono stati effettuati vari download, sono state selezionate le sole sentenze penali di tutte le sezioni della Corte di Cassazione, indicando come criterio di ulteriore selezione la ricerca di una o più delle parole: "corruzione", "concussione", "turbativa" e "appalto".

7.2. Il registro ASIA delle imprese attive

Il registro statistico delle imprese attive (ASIA) è aggiornato annualmente dall'Istat ed è costituito dalle unità economiche che svolgono attività industriali, commerciali e di servizio alle imprese e alle famiglie.

[42] ZULIANI A.- AURISICCHIO -G., CANZONETTI A., *Un'analisi statistica delle sentenze della Corte dei Conti: prime evidenze*, *in Rivista trimestrale di diritto pubblico,* 2009.

[43] Ulteriori dettagli sono disponibili nel "Sistema Italgiure Web. Manuale utente", disponibile sul sito (alla data del 15 settembre 2020 è disponibile la Versione 21.0 del novembre 2018).

Il registro contiene diverse informazioni sulla demografia e sulla struttura dell'impresa. I dati di struttura riguardano l'attività economica, il numero degli addetti, dipendenti e indipendenti, la forma giuridica, la data di inizio e fine dell'attività, il fatturato. Il registro rappresenta l'universo di riferimento delle indagini sulle imprese condotte dall'Istat. L'aggiornamento del registro è effettuato attraverso un processo di integrazione delle informazioni provenienti da fonti di diversa natura. La disponibilità di un registro unico di unità economiche consente di standardizzare e unificare le informazioni economiche sulle imprese. Il registro, in particolare, rappresenta la base informativa di tutte le indagini Istat sulle imprese, utilizzato per le stime di Contabilità nazionale ed individua la popolazione di riferimento per i piani di campionamento e per il loro riporto all'universo.
ed in particolare:

- la ragione sociale così come dichiarata alle camere di commercio,
- la localizzazione (regione/provincia),
- il settore di attività economica,
- la forma giuridica,
- il numero dei dipendenti,
- il valore aggiunto.

Sono, inoltre, disponibili anche altre informazioni di natura anagrafica relative all'impresa (data di inizio ed eventuale cessazione dell'attività), nonché altri dati su procedure in atto (ad esempio, amministrazione controllata, fallimento, liquidazione e altre).

7.3. Il Database Orbis

Il database Orbis "ha informazioni su circa 300 milioni di aziende in tutto il mondo". "Orbis è lo strumento ideale per verificare l'esistenza di una società e per generare rapporti sulle società: è molto più probabile che trovare un rapporto aziendale su Orbis che su qualsiasi altro database aziendale".
Tra le molte informazioni disponibili per singola impresa stati utilizzati finora i seguenti:

- Identificatori univoci (partita IVA o codice fiscale
- LEI (*Legal Entity Identifier*) e altri numeri ufficiali dell'azienda
- Codici di settore economico
- Sede dell'impresa
- Numero di addetti