ANAC
Autorità Nazionale Anticorruzione
Working Paper, 5

5

ANAC
Autorità Nazionale Anticorruzione

MATTEO TROÌA

DATA ANALYSIS E COSTRUZIONE DI INDICATORI DI RISCHIO DI CORRUZIONE PER LA BANCA DATI NAZIONALE DEI CONTRATTI PUBBLICI

M. TROÌA
Data analysis e costruzione di indicatori di rischio di corruzione per la Banca Dati Nazionale dei Contratti Pubblici

ESI

Edizioni Scientifiche Italiane

PREMESSA

Negli ultimi cinque anni l'Italia ha costantemente migliorato la sua posizione nella classifica pubblicata dall'organizzazione internazionale non governativa *Transparency International*. Questa classifica misura il livello di corruzione di 180 paesi del mondo, definendo un indice basato sull'opinione di una serie di esperti e sui risultati raccolti da una lunga serie di sondaggi, in riferimento alla corruzione percepita dal campione intervistato. Nel 2015, con un punteggio di 44/100, l'Italia occupava la posizione 61esima su 168. Secondo gli ultimi dati disponibili, riferiti all'anno 2019, l'Italia si colloca in posizione 51esima su 180, recuperando quasi dieci posizioni in cinque anni. Il *trend* appare dunque positivo, e i motivi riconducibili a questo miglioramento sono diversi. Sicuramente in concomitanza con l'inizio del periodo considerato, non va dimenticato l'accorpamento delle funzioni in capo all'Autorità per la vigilanza sui contratti pubblici di lavori, servizi e forniture (AVCP) all'Autorità Nazionale Anticorruzione (ANAC), che dal 2014 ha messo in campo una serie di politiche di lotta e di contrasto della corruzione che sicuramente hanno giovato al nostro Paese.

Il frequente utilizzo di indici *perception based* però, come quello di *Transparency International*, ha sempre reso difficile comprendere l'entità del "fenomeno corruzione". Gli indici basati sulla percezione degli intervistati o sulle loro personali opinioni, non restituiscono una misura oggettiva della diffusione della corruzione, sia che la si intenda indagare nella società privata, sia in riferimento a quella legata alla pubblica amministrazione. Pertanto, l'esigenza di strumenti di misurazione oggettiva del fenomeno, in grado non solo di affrontarlo dal punto di vista giuridico o qualitativo, ma di fornire risultati reali e precisi, diventa sempre più pressante, alla luce dell'estrema attualità e pervasività che la "questione corruzione" ricopre in Italia.

DATA ANALYSIS E COSTRUZIONE DI INDICATORI DI RISCHIO DI CORRUZIONE PER LA BANCA DATI NAZIONALE DEI CONTRATTI PUBBLICI

Matteo Troìa

1. GLI APPALTI PUBBLICI

1.1. Cosa sono e come funzionano gli appalti pubblici - La pubblica amministrazione, al fine di poter realizzare dei lavori, di acquistare dei beni o di erogare dei servizi, non sempre ha la possibilità di spendere del denaro in autonomia, così come lo potrebbe fare un'impresa privata, o come viene comunemente fatto dal privato cittadino. Le imprese private e i cittadini infatti, non solo spendono del denaro che gli appartiene, ma spendono o investono i loro soldi in acquisti che, di norma, hanno una ricaduta del tutto personale. Viceversa, ogni organizzazione pubblica, possedendo denaro pubblico, non solo è tenuta a rendicontare la propria attività quotidiana secondo il principio dell'«*accountability*» [9], ma ha altresì il compito di redistribuire la ricchezza che amministra alla cittadinanza.

Per attuare tutto questo, ogni organizzazione pubblica ha la necessità di indire delle gare d'appalto, ovvero di utilizzare degli strumenti tecnici e giuridici in grado di assicurare la trasparenza delle procedure nell'uso dei fondi, nonché un adeguato sistema di concorrenza e di meritocrazia. La gara d'appalto stabilisce l'*iter* fondamentale con cui la pubblica amministrazione spende il denaro pubblico. Il termine «gara» non è casuale: esso sottolinea l'aspetto competitivo, insito nella sua definizione. Una "gara d'appalto" è per l'appunto una competizione tra diversi soggetti economici che decidono di concorrere ad una determinata iniziativa.

Il processo appena descritto si riferisce esclusivamente alle fasi iniziali dell'intero svolgimento di un appalto pubblico. Un appalto infatti, è costituito da una fase iniziale (detta fase di aggiudicazione) e da una serie di altre fasi successive. Questa tesi si concentra esclusivamente sugli aspetti iniziali di un appalto, tralasciando le fasi successive all'aggiudicazione di una gara. Questo processo è esemplificato nella figura 1.1 che segue.

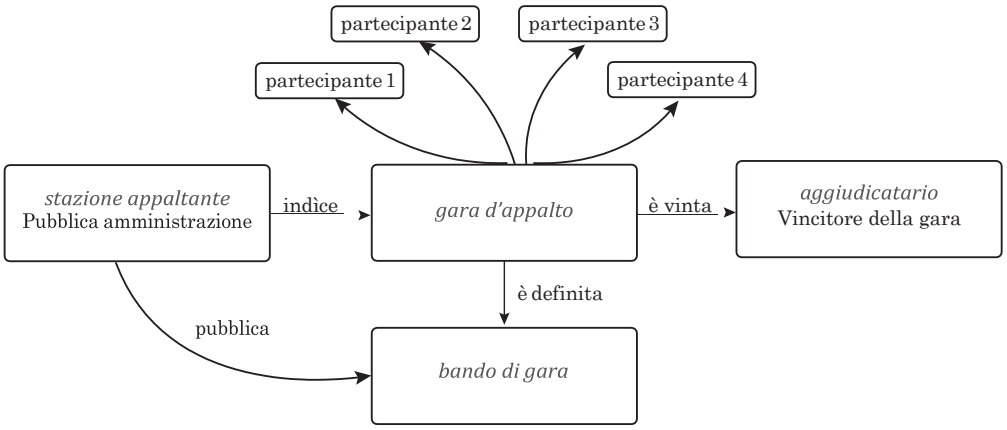

Figura 1.1: Schema generale della prima fase di un appalto pubblico

1.1.1. La Banca Dati Nazionale dei Contratti Pubblici - Il d.lg. 30 dicembre 2010, n. 235, che modifica ed integra il d.lg. 7 marzo 2005, n.82, noto come *Codice dell'Amministrazione Digitale (CAD)*, istituisce, tramite l'art. 62-bis, la *Banca Dati Nazionale dei Contratti Pubblici (BDNCP*). L'istituzione giuridica di questa banca dati, inaugura di fatto l'attività di raccolta e di gestione dei dati finanziari pubblici ad opera dell'Autorità per la vigilanza sui lavori pubblici, al fine di "*favorire la riduzione degli oneri amministrativi derivanti dagli obblighi informativi ed assicurare l'efficacia, la trasparenza e il controllo in tempo reale dell'azione amministrativa per l'allocazione della spesa pubblica in lavori, servizi e forniture, anche al fine del rispetto della legalità e del corretto agire della pubblica amministrazione e prevenire fenomeni di corruzione*.[1]"

La nascita di questa banca dati, che ad oggi conta dieci anni, sancisce una tappa di fondamentale importanza nell'evoluzione dei compiti spettanti alle autorità e alle istituzioni deputate al monitoraggio e al controllo dei flussi di denaro pubblico, poiché anzitutto rappresenta una presa di coscienza del valore e del potenziale dei dati da parte delle istituzioni italiane. La raccolta sistematica di ogni singolo passaggio del complesso processo con cui si sviluppa un appalto pubblico, ha rappresentato il primo fondamentale passo per rendere tempestivo ed efficace il controllo dei flussi finanziari da parte delle autorità competenti. Il monitoraggio dei flussi di denaro, non rappresenta l'unico motivo per cui è nata questa banca dati. Gli intenti scatenanti la costruzione e il mantenimento di questa piattaforma sono diversi, e sono rappresentati dalla volontà di raccogliere in un unico contenitore una vasta quantità di dati provenienti da enti pubblici geograficamente distanti e diversi per dimensioni, compiti, ambiti e competenze, oltre che dall'opportunità di analizzare i dati a supporto e beneficio del lavoro spettante l'Autorità, ma anche di quello di potenziali portatori di interesse. Ogni banca dati di fatto rappresenta un potente strumento di *decision making* per coloro che hanno la possibilità di accedere ai dati ed analizzarli.

[1] Art. 62 bis del Decreto Legislativo 7 marzo 2005, n. 82

La BDNCP in questo senso non fa eccezione. Oltre ad essere utilizzata dall'Autorità Nazionale Anticorruzione (ANAC), la BDNCP viene regolarmente utilizzata da altre importanti amministrazioni dello Stato, quali il Ministero dell'Economia e delle Finanze, la Ragioneria Generale dello Stato, la Corte dei Conti, il Ministero delle Infrastrutture, l'Istat, l'Ufficio Parlamentare di Bilancio, la Guardia di Finanza, i Carabinieri, le Procure della Repubblica, e diverse altre istituzioni.

In particolare, il funzionamento tecnico della banca dati viene così spiegato[2]: *"la banca dati raccoglie, integra e riconcilia i dati concernenti i contratti pubblici trasmessi dalle Stazioni Appaltanti per la disciplina e il controllo della materia dei contratti pubblici di lavori, servizi e forniture di qualsiasi importo e tipologia, senza restrizione alcuna, in ottemperanza ai principi di correttezza e trasparenza delle procedure di scelta del contraente, di economica ed efficiente esecuzione dei contratti e nel rispetto delle regole di concorrenza nelle singole procedure di gara."*[3]

Entrando brevemente nel merito di questa piattaforma, i numeri confermano la mole di informazione contenuta nella BDNCP. Secondo i dati aggiornati a marzo 2019 forniti da ANAC, la banca dati conta oggi 300.000 utenti attivi (ovvero le persone fisiche registrate che hanno effettuato almeno un accesso ai servizi online), 45.000 stazioni appaltanti, 250.000 operatori economici per un totale di 35 milioni di contratti. Nel 2018, il valore complessivo degli appalti di importo pari o superiore a 40.000 euro si è attestato attorno ai 139,5 miliardi di euro. Inoltre, secondo la più recente pubblicazione quadrimestrale [3] dell'Autorità Nazionale Anticorruzione, che fornisce dati aggregati e di sintesi aggiornati al mese di dicembre 2019, nei soli mesi compresi tra maggio ed agosto 2019 sono stati raccolti nella BDNCP contratti per 64.806.083.955e4. Le stazioni appaltanti che, nello stesso periodo, hanno bandito più contratti in riferimento all'importo delle gare, appartengono alla categoria delle regioni (1.728.256.877e) seguite dalla categoria delle università (1.458.838.156e).

1.1.2. Perché questi dati hanno valore - Il monitoraggio degli appalti pubblici, e quindi dei flussi di denaro pubblico, rappresenta una sfida indispensabile per migliorare le attività di *procurement* delle pubbliche amministrazioni, che oggi possono beneficiare delle potenzialità dell'analisi dati per prendere delle decisioni precise e mirate. Analizzare questi dati, come verrà mostrato nel corso di questa tesi, aiuta a visualizzare delle informazioni che senza l'aiuto dell'informatica sarebbe difficile comprendere, data la loro quantità e la frequenza con cui cambiano, giorno dopo giorno.

[2] G.P. Sellitto, Anac - 2019

[3] https://synapta.it/wp-content/uploads/2019/02/2019-03-02-OPENDATA-DAY-Sellito.pdf

[4] Importo complessivo relativo a contratti per lavori servizi e forniture sia del settore ordinario sia del settore speciale.

A monte delle motivazioni che fanno di questa banca dati un contenitore di estremo valore, è utile riportare quanto emerge dalla lettura di [28] e [22], ovvero che *"l'importanza del tema è strettamente legata alla sua dimensione economica, sia nazionale che europea, e all'ammontare di risorse che la corruzione drena dal sistema economico. Il settore dei contratti pubblici muove circa il sette per cento del PIL nazionale e il sedici per cento del PIL europeo"*.[5] Secondo alcune stime, la corruzione "costa" al sistema economico italiano circa 60 miliardi di euro all'anno[6] [17].

I dati contenuti nella Banca Nazionale dei Contratti Pubblici dunque, hanno valore perché da essi è possibile comprendere nel dettaglio come vengono spesi i soldi pubblici, in quali opere, in quali servizi, attraverso quali e quanti diversi fornitori. La loro analisi sprona a condurre delle indagini di tipo geografico (come sono distribuite le stazioni appaltanti sul territorio nazionale in riferimento ad un determinato ambito di contratti?), indagini di tipo economico (quali stazioni appaltanti spendono di più? Con che frequenza? Per acquistare cosa?), indagini di tipo temporale (quanto tempo intercorre, in media, tra la pubblicazione di un bando di gara e la sua aggiudicazione?), fino ad indagini più sofisticate, mirate ad individuare dei pattern ricorrenti o a misurare degli indicatori utili all'individuazione della corruzione. Proprio su quest'ultima tipologia di indagine si focalizzeranno i capitoli 3 e 4 di questo testo.

[5] Per ulteriori approfondimenti si suggerisce anche l'Audizione del Presidente dell'Autorità per la vigilanza sui contratti pubblici alla Commissione parlamentare d'inchiesta sul fenomeno delle mafie e sulle altre associazioni criminali anche straniere, Roma, 25 maggio 2010.

[6] https://st.ilsole24ore.com/art/SoleOnLine4/Economia%20e%20Lavoro/2009/06/corte-conti-corruzione-rendiconto-generale-stato_PRN.shtml

2. APPALTI E CORRUZIONE

2.1. L'esigenza di misurare la corruzione e le motivazioni per farlo - In Italia si parla molto di corruzione. Il fenomeno è spesso oggetto di dibattito nei giornali e nei telegiornali, diventando ormai nell'immaginario collettivo una delle piaghe più significative del nostro Paese. Sentiamo parlare di corruzione quando viene scoperta l'elargizione di tangenti, quando imprese sempre più grandi tentano di avviare business all'estero in maniera poco trasparente o quando ci imbattiamo in casi di cattiva amministrazione, ad opera di funzionari o di intere organizzazioni che vengono meno ai loro fini istituzionali.

La corruzione è un fenomeno che non riguarda solamente chi ne è l'artefice, né chi sottoscrive per così dire il "patto illecito", ma coinvolge anche coloro che non ne sono direttamente coinvolti. Gli effetti indiretti della corruzione sul Paese sono significativi almeno tanto quanto quelli diretti, poiché ogni volta che si verifica un illecito, si genera un"'asimmetria finanziaria" in grado di arricchire un ridotto numero di soggetti, a discapito dell'intera cittadinanza.

La corruzione è un fenomeno complesso, articolato in diverse forme, che non riguarda, come scrivono Francesco Caringella e Raffaele Cantone [10] *"solo passaggi di denaro, ma giri vorticosi e smaterializzati di favori, piaceri, collusioni. Non più il classico accordo privato fra corruttore e corrotto, ma la creazione di un'organizzazione criminale attraverso cui politici, burocrati, imprenditori e mafiosi perseguono gli stessi obiettivi."* La conseguenza di questo complesso sistema di favoritismi, fa sì che la corruzione rappresenti in Italia una delle maggiori cause di sperpero di denaro pubblico, di inefficienza dei servizi erogati dalle pubbliche amministrazioni e di sfiducia da parte del cittadino nelle istituzioni democratiche, come viene sottolineato dalla stessa ANAC nella *Relazione annuale sull'attività svolta dall'Autorità nel 2018*[1], che apre il primo capitolo sottolineando come *"negli ultimi anni, il tema della corruzione si è indiscutibilmente imposto sempre più nello scenario internazionale, come fenomeno avvertito, anche da parte dei cittadini, nella sua gravità, per gli effetti negativi che determina sul tessuto sociale e sulla competitività del sistema economico ed in grado, altresì, di compromettere la legittimazione delle istituzioni democratiche."*

[1] http://www.anticorruzione.it/portal/public/classic/Attivitadocumentazione/Pubblicazioni/RelazioneParlamento_relazioni

Eppure, nonostante tutti parlino di corruzione, (ne parla spesso anche Papa Francesco, di cui divenne celebre la frase che pronunciò nel 2017 in visita a Scampia, *"un cristiano che lascia entrare dentro di sé la corruzione "spuzza". La corruzione spuzza, la società corrotta spuzza"*) pochi sono oggi in grado di quantificarla, di misurarla, e di conseguenza di individuarla in maniera sistematica. Misurare la corruzione, non darebbe però solo contezza di quanto il nostro Paese sia effettivamente corrotto, ma aiuterebbe a capire dove si annidano i fenomeni illeciti, chi coinvolgono e con quale frequenza, se esiste una loro particolare distribuzione geografica, se vi sono in corso degli effetti *lock-in*[2] tra stazioni appaltanti e aggiudicatari, e via dicendo.

Non vi è dubbio che negli anni il legislatore abbia fortemente concentrato i suoi sforzi nella costruzione di un *framework* giuridico unico, capace di prevenire e contrastare i fenomeni corruttivi del nostro Paese. Né si può negare, allo stesso tempo, gli sforzi e i contributi di cui oggi disponiamo da parte di rinomati gruppi di lavoro, che negli anni si sono concentrati su questo tema, cercando di definire la corruzione in maniera formale, e provando a proporre una serie di indicatori utili alla misurazione di un fenomeno che per natura sfugge ad una qualsivoglia definizione oggettiva; tuttavia va constatato che la maggior parte dei contributi e degli studi realizzati, provengono dalla letteratura economica, politica, sociologica, antropologica e senz'altro giuridica (si veda ad esempio [12] oppure [34]). Ciò che ancora fatica ad aumentare, è la percentuale di contributi tecnico-informatici-statistici sul tema, volti a mettere a disposizione gli strumenti che oggi l'informatica dispone, a beneficio del tema in esame. I contributi presenti in letteratura hanno senz'altro influenzato in maniera positiva le politiche pubbliche, fornendo ad esse importanti strumenti qualitativi. L'aspetto quantitativo del fenomeno però, ad oggi è ancora poco sondato. L'idea è che nuove ed originali indagini quantitative non possano fare altro che completare e supportare quelle qualitative. Solo in questa maniera, sottolinea Benedetto Ponti, PhD. in Diritto Pubblico e professore associato di Diritto Amministrativo presso il Dipartimento di Scienze politiche dell'Università degli studi di Perugia, *"è possibile porsi degli obiettivi specifici (anche perché quantificati), dimensionare, articolare, come pure indirizzare le misure di intervento a fini di prevenzione e contrasto), e - infine - di passare dalla comprensione del fenomeno (anche) alla sua individuazione."*

Vi è quindi l'esigenza di misurare la corruzione e non solo di parlarne. Proprio per questo, si ritiene che l'analisi degli appalti pubblici che questa tesi si prefigge di fare possa aiutare gli approcci orientati alla misurazione oggettiva del "fenomeno corruzione". In virtù di questo assunto, si è provato ad arricchire il dibattito in corso con un approccio *data driven*, nell'idea che la *data scence* sia un'ottima alleata per raggiungere gli scopri prefissati. Si ritiene di poter sostenere l'utilità di questa scelta per almeno tre motivi.

[2]Fenomeno che si verifica quando un agente, un insieme di agenti, o un intero settore sono intrappolati all'interno di una scelta o di un equilibrio economico dal quale è difficile uscire, anche se sono disponibili alternative potenzialmente più efficienti.

Il primo riguarda la scarsa attitudine a prendere decisioni e attuare strategie *data based*. Rispetto ai lavori che si sono occupati del rapporto "corruzione-appalti", è importante selezionare in particolar modo i lavori che utilizzano i dati in maniera sistematica [21]. L'approccio deve essere esattamente quello proposto dagli autori: *"(...) while there have been many qualitative accounts of high-level corruption in public contracting, it is only recently that quantitative indicators have become available. By making use of big data generated by governments on contracts, companies, and individuals, it is possible to develop a new generation of quantitative indicators which can be used to guide policy intervention and support control of corruption."* La carenza di misurazioni oggettive e di indicatori, nonché la grande quantità di dati pubblici oggi disponibili (*big data*), stimola a mettersi al lavoro in tal senso, a patto di una sincera collaborazione da parte della pubblica amministrazione a mettere a disposizione i dati che raccoglie e che possiede.

Il secondo motivo riguarda la possibilità, se non di individuare in maniera deterministica gli appalti soggetti a fenomeni di corruzione, perlomeno di evidenziare una serie di *pattern* da discutere con gli addetti della materia per ulteriori ed eventuali accertamenti. L'analisi di questi dati quindi, non vuole essere soltanto un'analisi volta alla storia passata degli appalti, ma si configura soprattutto come un'attività di prevenzione rivolta al futuro, grazie all'individuazione di elementi anomali ricorrenti, in grado di suggerire alle autorità competenti un approfondimento ulteriore sullo stato di un appalto al tempo presente e sui possibili stati che potrebbe assumere l'appalto nel futuro. Sull'idea di mettere in campo delle analisi atte a prevedere alcuni comportamenti futuri si ritiene importante citare il *machine learning* con le sue numerose tecniche previsionali, come area su cui sarebbe molto importante lavorare in riferimento al tema corruzione.

Il terzo motivo riguarda invece un dato suggestivo, che rappresenta il legame tra la digitalizzazione di un Paese (misurata secondo l'indice DESI)3 e la corruzione percepita (misurata secondo l'indice CPI di Transparency International). La correlazione appare forte e positiva4, restituendo un valore pari all'88,6%. Nel costruire il grafico 2.1, sono state prese in esame 3 delle 5 sottodimensioni con cui viene calcolato il DESI, poiché considerate quelle maggiormente connesse al tema della corruzione. Le sottodimensioni prese in esame sono:

1. *Capitale umano*, che include al suo interno l'uso di internet e le competenze digitali di base e avanzate;

2. *Connettività*, che comprende la copertura della banda larga fissa, banda larga mobile, la velocità di connessione e i prezzi della banda larga;

3 *Servizi pubblici*, che riguarda l'efficacia dell'*e-government*.

[3] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=59913

[4] https://www.riparteilfuturo.it/assets/articles/images/report_Italia_interrotta_2018.pdf

Queste dimensioni sono state correlate all'Indice di Percezione della Corruzione (CPI). I dati sono stati trasformati in scala logaritmica prima di essere processati tramite l'indice di correlazione di Pearson [32]. Dai dati emerge che ad un incremento del 10% dello sviluppo digitale è associato un abbassamento del livello di corruzione pari al 9,8%.

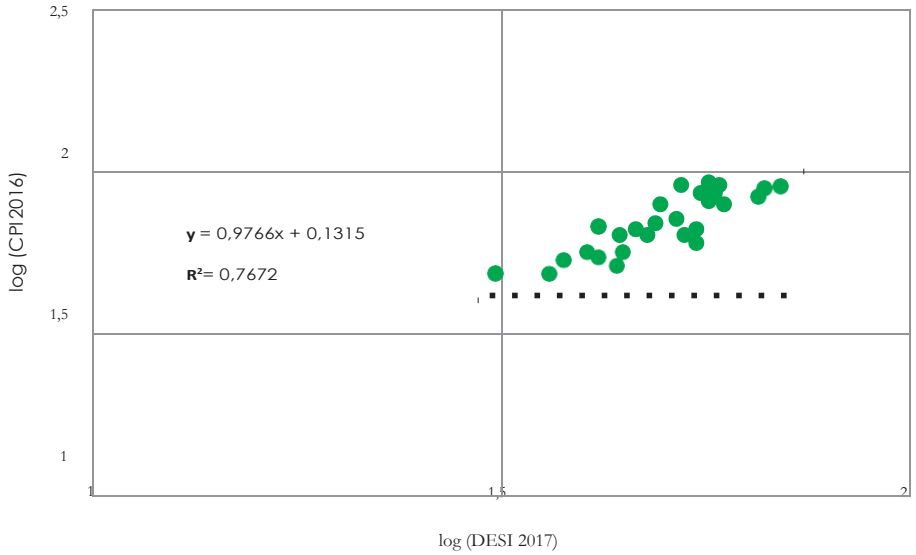

Figura 2.1: Elaborazioni I-Com su dati Transparency International e Digital Scoreboard

Il risultato riportato nel grafico 2.1 permette di affermare che dove il livello di digitalizzazione di un paese risulta elevato, il corrispondente livello di corruzione è basso, sebbene non si possa affermare che il primo fenomeno sia causa del secondo. La correlazione che intercorre tra le due variabili potrebbe rivelarsi una correlazione spuria [40]. Correlazioni di questo tipo tendenzialmente dipendono da una terza variabile in comune, che, nel caso in esame, non è nota. La correlazione tra corruzione e digitalizzazione suggerisce quindi non solo di continuare ad investire nella digitalizzazione delle organizzazioni pubbliche, ma di considerare la tecnologia lo strumento principale nella lotta al contrasto della corruzione.

In conclusione, la carenza di misurazioni oggettive e di indicatori di corruzione capaci di misurare in maniera sistematica il fenomeno, la possibilità di sfruttare la grande mole di dati oggi disponibile per identificare una serie di *pattern* da utilizzare a scopi preventivi, nonché la correlazione tra la digitalizzazione e la corruzione, stimolano a proseguire i lavori di ricerca su questa strada, nell'intento di elaborare strumenti sempre più efficaci, in grado di fare dei dati il cuore di questo nuovo approccio metodologico.

2.2. Misurare la corruzione.

2.2.1. Perché è difficile - Giuseppe Arbia, PhD all'Università di Cambridge e professore ordinario di Statistica economica all'Università Cattolica del Sacro Cuore di Roma, in [30] cita la poetessa britannica Christina Rossetti: "chi ha mai visto il vento? Né io né te. Ma quando gli alberi chinano le loro teste, il vento li ha attraversati." Questa breve poesia risulta piuttosto suggestiva in riferimento al tema in esame. La corruzione assomiglia proprio a questo vento: si fatica ad individuarne il passaggio, ma si riscontrano in maniera evidente i suoi effetti. Proprio le difficoltà intrinseche delle misurazioni dirette legate alla corruzione, hanno condotto alla costruzione e alla diffusione di indici indiretti essenzialmente basati sulla percezione. Questa famiglia di indicatori però presenta alcuni limiti.

Il primo limite è di natura metodologica. La critica in tal senso è riferita al fatto che le misurazioni basate sulla percezione non sono in grado di restituire risultati accurati e attendibili, e che spesso, il guadagno in termini di precisione statistica generata dall'aggregazione di diverse fonti di dati sulla corruzione, è molto più modesto di quanto viene affermato. La letteratura infatti [26], suggerisce che potrebbe essere più appropriato utilizzare i dati di una singola fonte, piuttosto che un indice composito che causa un'inevitabile perdita di precisione, dovuta alle operazioni di aggregazione di dati riferiti a tematiche differenti.

Ad oggi possediamo solamente delle stime sul "fenomeno corruzione", ma come scrive Raffaele Cantone, *"è diffusa l'idea che la corruzione in Italia abbia un costo di 60 miliardi di euro annui, una stima che però, per quanto ricorrente in vari documenti ufficiali, è frutto di una valutazione troppo sommaria per essere affidabile"*.

Il secondo limite è dovuto al successo di questi indici. Un lavoro che indaga ed approfondisce le dinamiche distorsive e *bias* cognitivi di cui sono affetti i risultati degli indici di percezione è riportato in [1], in cui dopo aver individuato alcune criticità dei più comuni indici di percezione (fondamentalmente le stesse qui riportate), viene illustrato come questi indici continuino ad esercitare una grande influenza sia sulla ricerca accademica sia nei confronti delle politiche legate all'anticorruzione, pena, talvolta, l'oggettività delle misurazioni.

Il terzo limite infine, riguarda la scala degli strumenti di misurazione messi in campo. Gran parte di essi sono stati costruiti per agevolare il confronto tra sistemi-Paese, e pongono quindi il Paese nel suo complesso come unità di base nella stesura dei *ranking* finali. Tra le attività dell'Autorità anticorruzione vi è senz'altro quella legata al confronto tra vari paesi europei ed extraeuropei, ma vi è soprattutto l'esigenza di avere dei dati di dettaglio: a livello regionale, provinciale e, all'occorrenza, comunale.

Non solo, nel caso dell'Italia e in riferimento alla BDNCP[5], potrebbe essere molto utile effettuare dei confronti i termini di stazioni appaltanti, di imprese offerenti o aggiudicatarie, di appalti aggiudicati sopra ad un determinato importo, di oggetti del bando e così via. Ma ancora una volta, gli indici di percezione non sono in grado di effettuare delle analisi verticali sui domini che analizzano. In conclusione, come scrive il professore Arbia, *"risulta evidente l'esigenza di disporre di strumenti di analisi e misurazione, funzionali all'implementazione di una politica nazionanale di contrasto ai fenomeni di corruzione amministrativa, diversi e ulteriori rispetto a quelli (più tradizionali e consolidati) basati sulla percezione. (...) Le misure finalizzate ad abilitare il confronto (e l'arbitraggio) tra sistemi-Paese, si rivelano insufficienti (se non, fuorvianti), quando si tratta invece di disporre di strumenti di analisi e di misurazione che siano adeguati a "leggere" e restituire lo specifico contesto in cui si sviluppano politiche di prevenzione e contrasto."*

2.2.2. Costruire nuovi indicatori di corruzione - Nel lavoro di costruzione di nuovi indicatori, si è ritenuto utile abbandonare la dicotomia tra indicatori soggettivi e indicatori oggettivi, che può condurre a ritenere meno importanti i primi e più importanti i secondi. In tal senso si è deciso di adottare la proposta avanzata da Michela Gnaldi, professoressa associata del Dipartimento di Scienze Politiche dell'Università degli Studi di Perugia, la quale suggerisce di orientare la scelta dell'indicatore più opportuno al solo oggetto della misurazione. Ne scaturisce così una classificazione suddivisa in cinque livelli, riportata nella tabella che segue:

Indicatori di corruzione	Oggetto della misurazione
Indicatori compositi di percezione della corruzione	Percezione della corruzione
Inchieste campionarie basate su esperienze dirette (*self-reported*)	Esperienze dirette della corruzione
Statistiche giudiziarie	Atti giudiziari (es: sentenze, denunce, condanne per reati di corruzione)
Inferenza statistica e misure di mercato	Confronto tra dati reali e modello teorico ipotizzato (in assenza di corruzione)
Misure di rischio e prevenzione	Presenza di anormalità che segnalano rischi di corruzione

La tabella, restituisce un quadro esaustivo dei più diffusi approcci con cui è possibile studiare la corruzione. Il lettore, giunto a questo punto, dovrebbe essere in grado di individuare, per ciascuna famiglia di indicatori, i vantaggi e gli svantaggi, gli elementi legati alla percezione del fenomeno e gli elementi oggettivamente misurabili.

[5]Banca Dati Nazionale Contratti Pubblici

L'utilizzo di una metodologia non esclude l'altra e anzi, probabilmente, solamente l'utilizzo di varie metodologie aiuta a consolidare i risultati finali.
D'ora in avanti, questa tesi si concentrerà esclusivamente sulla famiglia di indicatori che in tabella vengono classificati tra le *"misure di rischio e prevenzione"*, alla ricerca della presenza di anomalie nei dati che possano, potenzialmente, segnalare il rischio di corruzione. Si ritiene che l'approfondimento di questa particolare famiglia di metodologie, rappresenti la scelta più adatta in riferimento agli scopi di questo lavoro e soprattutto ai dati a disposizione. Le tecniche che perseguono questo approccio, generano una serie di campanelli d'allarme (*red flags*), che segnalano l'individuazione di specifici *pattern* anomali all'interno dei dati analizzati.
Naturalmente, la possibilità di condurre analisi efficaci è subordinata alla disponibilità di dati amministrativi accessibili e curati. In questo senso le istituzioni competenti hanno il dovere di rendere accessibili i dati amministrativi che le pubbliche amministrazioni producono. Di conseguenza, l'impegno dovrebbe essere quello di *"tradurre in strumenti concreti le potenzialità conoscitive connesse all'affermazione di un paradigma tecnologico che consente di produrre, conservare, far circolare ed elaborare l'informazione con costi, velocità e potenza che non hanno precedenti, e che pertanto consentono oggi, anche con riferimento alle "misure della corruzione", di rendere effettivamente attingibili strumenti che in precedenza non potevano esserlo"*, come sottolineato da Bendetto Ponti in [30].
Un utile approfondimento legato al tema della disponibilità dei dati pubblici e soprattutto alla loro qualità è riportato in [25].
I dati utilizzati in questa tesi e le analisi che verranno fatte nei capitoli successivi, esprimono da subito almeno tre elementi positivi:

1. permettono un'analisi *county-specific* che gli indicatori basati su scala mondiale non sono in grado fare;

2. tengono conto della normativa relativa ai contratti pubblici italiana, soddisfando così l'esigenza di realizzare un'analisi calata nel contesto giuridico del paese preso in esame;

3. rispondono alle indicazioni di ANAC[6], che individua nella BDNCP una delle principali risorse pubbliche per l'accesso a dati amministrativi dettagliati ed affidabili.

Le potenzialità delle misure di rischio e prevenzione, assieme agli elementi positivi che invitano ad utilizzare i dati della "Banca Dati Nazionale dei Contratti Pubblici", spronano a testare alcuni indicatori sui dati disponibili. Nel paragrafo successivo verranno elencati gli indicatori individuati dall'Autorità

[6] https://bit.ly/38rAN9G

Nazionale Anticorruzione, che rappresenteranno il riferimento da cui sono cominciate le analisi. Per ciascun indicatore verrà riportata una sintetica descrizione.

2.2.3. Gli indicatori dell'Autorità Anticorruzione - Nel documento *Analisi istruttoria per l'individuazione di indicatori di rischio corruzione e di prevenzione e contrasto nelle amministrazioni pubbliche coinvolte nella politica di coesione*[7] sono riportati, tra i documenti allegati, una lista di indicatori di contrasto della corruzione riferiti agli appalti pubblici, classificati come "indicatori oggettivi di rischio di corruzione". La lista presenta 19 proposte di indagine, da cui si è scelto di estrarne 5, di seguito riportate così come descritte da ANAC. Gli indicatori esclusi sono stati scartati per assenza di dati utili a poterne effettuare il calcolo. Si ritiene viceversa che gli indicatori selezionati misurino gran parte dei diversi aspetti della fase di aggiudicazione di un appalto.

i) Indicatore sulle procedure che utilizzano il criterio dell'OEPV - I_{oepv}

L'offerta economicamente più vantaggiosa (*OEPV*), sebbene trovi anche con l'introduzione delle nuove direttive uno spazio sempre maggiore come criterio di scelta da utilizzare, presenta un più alto rischio di discrezionalità rispetto al criterio del prezzo più basso. Sotto il profilo della letteratura economica l'utilizzo dell'OEPV sarebbe più indicato per appalti complessi mentre il criterio del prezzo più basso sarebbe da preferire per appalti con componenti standardizzate. L'indicatore può essere espresso come segue:

$$I_{oepv} = \frac{NTPOEPV_{i,t}}{NTP_{i,t}}$$

dove il termine $NTPOEPV_{i,t}$ rappresenta il numero dei bandi della *i*-esima stazione appaltante al tempo *t* che utilizzano il criterio dell'offerta economicamente più vantaggiosa ed $NTP_{i,t}$ è il numero totale delle procedure di appalto utilizzate dalla *i*-esima stazione appaltante al tempo *t*.

ii) Indicatore sul numero delle procedure non aperte - I_{npna}

L'indicatore ha lo scopo di valutare la percentuale di procedure non aperte (procedure negoziate con o senza previa pubblicazione di un bando, affidamenti diretti, cottimi fiduciari, ecc.) sul totale delle procedure utilizzate da una medesima stazione appaltante in un determinato arco temporale. L'indicatore di per se stesso non segnala illegittimità poiché è possibile che le procedure prescelte da una stazione appaltante diverse da quelle aperte o ristrette rispettino tutti i requisiti imposti dalla normativa vigente. Tuttavia, una elevata percentuale di affidamenti non concorrenziali insieme ad altri indicatori potrebbe segnalare una patologia da monitorare in maniera specifica.

[7] https://bit.ly/2vvUvm9

L'indicatore può essere calcolato come segue:

$$I_{npna} = \frac{NT\,PNA_{i,t}}{NT\,P_{i,t}}$$

dove il termine $NT\,PNA_{i,t}$ è il numero delle procedure di appalto non aperte o ristrette utilizzate dalla *i*-esima stazione appaltante al tempo *t*, e $NT\,P_{i,t}$ è il numero totale delle procedure di appalto utilizzate dalla *i*-esima stazione appaltante al tempo *t*.

iii) Indicatore sul valore delle procedure non aperte - I_{vpna}

Questo indicatore è analogo al precedente, con l'unica differenza di considerare il valore delle procedure non aperte sul valore totale delle procedure attivate da una medesima stazione appaltante in un determinato periodo. Il presente indicatore andrebbe letto congiuntamente con l'indicatore I_{npna}. L'indicatore può essere calcolato come segue:

$$I_{vpna} = \frac{VT\,PNA_{i,t}}{VT\,P_{i,t}}$$

dove il termine $VT\,PNA_{i,t}$ è il valore totale delle procedure non aperte attivate dalla *i*- esima stazione appaltante al tempo *t*, $VT\,P_{i,t}$ è il valore totale delle procedure attivate dalla *i*-esima stazione appaltante al tempo *t*.

iv) Indicatore sul numero di procedure con un'unica offerta - I_{uo}

L'indicatore consiste in un semplice conteggio di bandi per i quali è stata ricevuta una sola offerta. Questo indicatore andrebbe valutato guardando il mercato di riferimento del servizio/prodotto. Per alcuni beni e servizi, infatti, l'analisi dei dati effettuata ha mostrato molte gare con la presenza di un'offerta singola. L'indicatore è calcolato come segue:

$$I_{uo} = \frac{N\,TA1_{i,t}}{N\,TA_{i,t}}$$

dove $NTA1_{i,t}$ è il numero delle procedure aggiudicate della *i*-esima stazione appaltante al tempo *t* con un numero dei partecipanti uguale ad uno. $N\,TA_{i,t}$ è il numero totale delle procedure di appalto aggiudicate dalla *i*-esima stazione appaltante al tempo *t*.

v) Indicatore sul tempo medio richiesto per la presentazione delle offerte - I_{tmpo}

L'indicatore misura l'adeguatezza della tempistica necessaria alla presentazione delle offerte.

Tempi molto ristretti che intercorrono tra la data di pubblicazione del bando e la data di scadenza per la presentazione delle offerte possono essere un indice di favoritismo nei confronti di un particolare operatore economico e comunque un segnale di restringimento del grado di concorrenza potenziale. L'indicatore si calcola come segue:

$$I_{tmpo} = \frac{\sum_{k=1}^{NTA_{it}} (DSPO_{ik} - DPB_{ik})}{NTA_{i,t}}$$

dove $DSPO_{i,k}$ è la data di scadenza di presentazione delle offerte per l'amministrazione *i*-esima e per l'affidamento *k*-esimo, $DPB_{i,k}$ è la data di pubblicazione del bando per l'amministrazione *i*-esima e per l'affidamento *k*-esimo. Infine, $NTA_{i,t}$ è il numero totale delle procedure di appalto aggiudicate dalla *i*-esima stazione appaltante al tempo *t*.

3. DESCRIVERE E PREPARARE I DATI

3.1. I dati a disposizione - I dati utilizzati in questo lavoro di tesi, sono stati ottenuti attraverso il "regolamento concernente l'accessibilità dei dati raccolti nella Banca Dati Nazionale dei Contratti Pubblici"[1] pubblicato nella Gazzetta Ufficiale n. 80 del 6 aprile 2018. All'articolo 4 del suddetto regolamento, ANAC specifica che "chiunque può accedere ai dati, nel rispetto della normativa in materia di trattamento dei dati personali, attraverso apposite modalità intese quali servizi di consultazione disponibili sul portale dell'ANAC, la quale ne disciplina le caratteristiche tecniche". L'articolo 7 invece, afferma che "fino al momento della completa disponibilità dei servizi di cui all'art. 4, le richieste riguardanti dati non già liberamente accessibili attraverso il portale dell'Autorità sono formulate utilizzando l'apposita modulistica messa a disposizione sul sito dell'ANAC".

Tra gli scopi di questa tesi, vi è anche quello di valutare le modalità di accessibilità ai dati, nonché la loro effettiva disponibilità.

La richiesta dei dati utilizzati nelle analisi, è stata inoltrata ad ANAC tramite la compilazione dell'apposito modulo[2] a cui fa riferimento il regolamento. Tramite questo modulo è stato possibile specificare la profondità dei dati che si intendeva richiedere all'Autorità. La richiesta iniziale riguardava l'intera banca dati, nell'intento di poter testare così l'effettiva possibilità di accedere all'intero patrimonio informativo in possesso dell'Autorità Anticorruzione. La richiesta non è andata a buon fine, poiché "non era attualmente possibile dare corso alla stessa, in quanto si trattava di una richiesta massiva di tutti i dati della BDNCP[3]". Appurato che l'intera banca dati non era accessibile, si è provveduto ad inoltrare una nuova richiesta, sempre attraverso il modulo proposto dal regolamento citato.

[1] https://bit.ly/2P747KS [2] https://bit.ly/32bdNt7

[3] Risposta ufficiale data da ANAC a seguito della richiesta di accesso all'intero patrimonio informativo contenuto nella BDNCP

Questa seconda richiesta chiedeva ad ANAC la massima partizione possibile dei dati contenuti nella BDNCP, senza distinzione tra le categorie dei lavori, dei servizi o delle forniture, e senza alcun tipo di vincolo. La richiesta così formulata è stata esaudita attraverso l'invio dei dati sugli appalti relativi agli anni 2015, 2016 e 2017. Tutti i dati ricevuti sono in formato csv, e seguono lo schema con cui verranno pubblicati nella Piattaforma Digitale Nazionale Dati[4], al quale, secondo ANAC, si dovrà fare riferimento per ulteriori *dataset* ed eventuali aggiornamenti. Ritenuto che la quantità di dati ricevuti fosse sufficiente agli scopi prefissati, si è proceduto alla loro analisi, senza ulteriori richieste all'Autorità. Nonostante sia stato impossibile accedere all'intera banca dati così come previsto per legge, si è comunque potuto avviare un lavoro d'analisi.

3.1.1. La modellazione dei dati ricevuti - Secondo Hadley Wickham [38], la preparazione dei dati non è solamente il primo *step* di un'analisi, ma è un'attività che va ripetuta più volte nel corso del tempo man mano che emergono nuove problematiche e mentre si collezionano nuove basi di dati. Il processo di pulizia (*data cleaning*) effettuato sui dati ricevuti ha riguardato principalmente la loro strutturazione in una serie di nuove tabelle. Le operazioni di pulizia effettuate in R, non rappresentano altro che l'implementazione di alcuni concetti chiave della teoria delle basi di dati e dell'algebra relazionale (si veda a tal proposito [14]).
I dati ricevuti da ANAC sono stati suddivisi in quattro diverse tabelle, ciascuna racchiusa all'interno di una cartella che ne specifica l'anno di riferimento rispetto a quelli concordati per via burocratica. I quattro *dataset* ricevuti sono così suddivisi:

- *appalti senza oggetto*, è formato da 48 diversi attributi, che esprimono diversi aspetti della fase di aggiudicazione degli appalti (dall'elenco delle stazioni appaltanti, alle modalità di scelta del contraente, dalla data di pubblicazione della gara alla data di aggiudicazione e così via);
- *gara lotto oggetti*, contiene invece un elenco di codici identificativi delle gare (Cig), corredati dall'oggetto di ciascuna gara e dall'oggetto di ciascun lotto in cui una gara può essere suddivisa;
- *cig cup*, contiene l'elenco di Cig e dei Cup[5] riferiti all'anno a cui il dataset si riferisce;
- *aggiudicatari*, raccoglie invece i dati inerenti agli aggiudicatari delle gare, riportandone il loro codice fiscale, la loro denominazione, il loro ruolo nella gara e altre informazioni di corredo.

[4] https://pdnd.italia.it/
[5] Il codice CUP (Codice Unico di Progetto) è un codice univoco di 15 caratteri alfanumerici che identificano un progetto d'investimento pubblico.

L'intera lista degli attributi che caratterizzano questi quattro *dataset* è riportata in figura 3.1, dove risulta evidente che la maggior parte dell'informazione è contenuta nel *dataset* "*appalti senza oggetto*", mentre i *dataset* "*Gare lotto oggetti*" e "*Cig cup*" condividono con gli altri *dataset* alcune informazioni (come l'elenco dei Cig). Tutto ciò che riguarda gli aggiudicatari è invece raccolto nella relativa tabella.

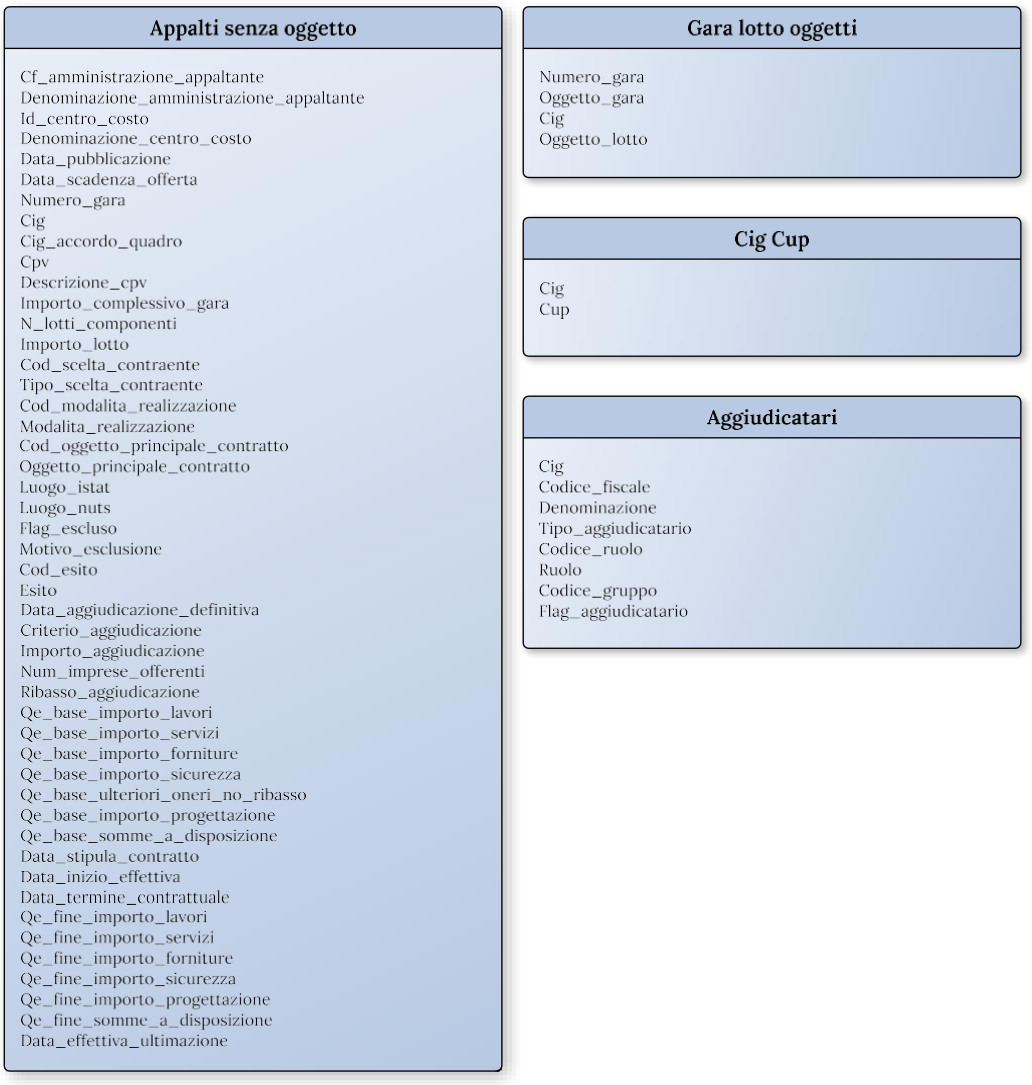

Figura 3.1: *I quattro dataset iniziali e i loro attributi*

L'organizzazione dell'informazione così come presentata da ANAC, non risulta molto utile alle analisi previste, né rispecchia in maniera adeguata la fase di aggiudicazione di un appalto pubblico. Pertanto si è provveduto a progettare una nuova suddivisione degli attributi, che potesse far emergere le entità presenti nei dati, e le relazioni che intercorrono fra di esse. Il lavoro di progettazione concettuale è consistito nella costruzione del diagramma entità-relazione (E-R) riportato in figura 3.2. Il diagramma segue i principi riportati in [13], [5] e [20]. A partire dai dati ricevuti, si è pertanto provveduto a ricavare tre nuove

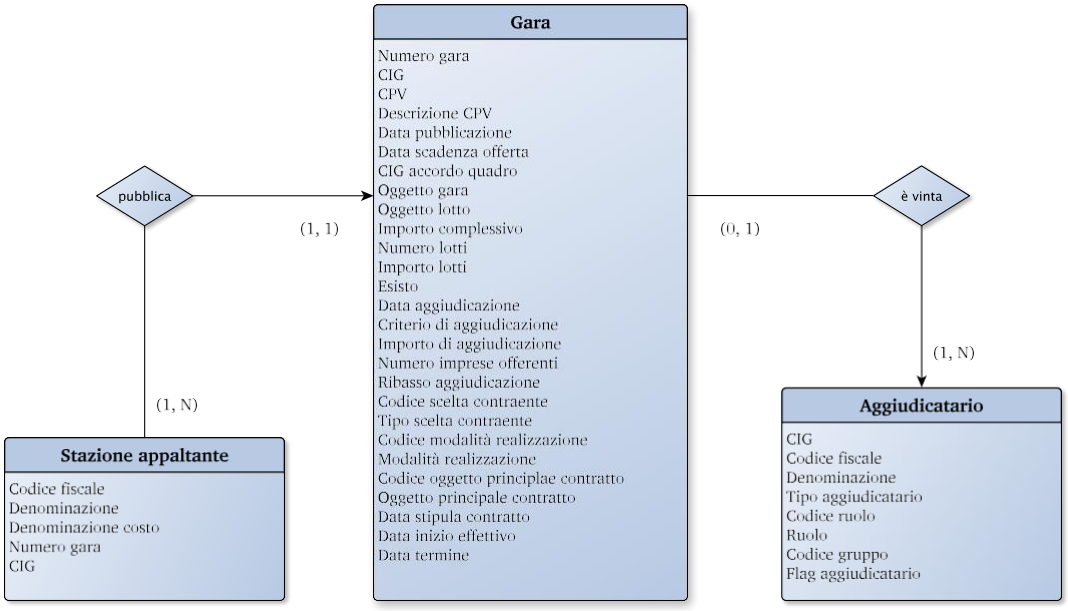

Figura 3.2: *Il diagramma E-R*

entità: *stazione appaltante*, *gara* e *aggiudicatario*, collegate dalle relazioni *"pubblica"* e *"è vinta"*. Ciascuna di queste tre nuove tabelle rappresenta un dominio specifico, così come illustrato di seguito:

- la tabella "*Stazione appaltante*" definisce le organizzazioni pubbliche coinvolte in una o più gare d'appalto;
- la tabella "*Gara*" riassume tutti i dettagli tecnici e organizzativi delle gare effettuate;
- la tabella "*Aggiudicatari*" elenca tutti gli enti aggiudicatari di una o più gare d'appalto.

I vincoli di cardinalità posizionati lungo gli archi del diagramma, aiutano a modellare le relazioni che intercorrono tra le diverse entità raffigurate. Leggendo la figura 3.2 da sinistra a destra possiamo osservare che:

- Il vincolo *(1, N)* afferma che una stazione appaltante può pubblicare da 1 a *n* gare nel corso dell'anno per sopperire a diverse esigenze. Naturalmente una stazione appaltante potrebbe anche non pubblicare alcuna gara in un determinato anno, ma questo caso è escluso in riferimento agli scopi di questo lavoro.
- il vincolo *(1, 1)* afferma che una gara esiste come entità se e solo se esiste un'organizzazione pubblica responsabile di quella gara. Allo stesso tempo ciascuna gara può essere indetta al massimo da una e una sola stazione appaltante. L'entità "gara" dunque, esiste solo in funzione di una stazione appaltante;

- il vincolo (0, 1) che collega l'entità "*gara*" all'entità *"aggiudicatario"* dalla relazione *"è vinta*", afferma che una gara può non essere stata vinta da alcun soggetto economico; di conseguenza non avrà alcun aggiudicatario (è il caso in cui una gara va deserta o non è stata ancora pubblicata nell'intervallo di tempo *t* considerato). Viceversa se una gara è stata vinta, allora deve esistere almeno un aggiudicatario;

- il vincolo (1, N) infine, afferma che ogni aggiudicatario, per essere tale, deve aver vinto almeno una gara, ma allo stesso tempo è possibile che un soggetto economico sia aggiudicatario di più di una sola gara, da cui il vincolo *N* come margine superiore.

Il lavoro di progettazione concettuale appena illustrato, ha permesso di organizzare meglio l'informazione ricevuta, e ha rappresentato il punto di partenza per le successive analisi in R, attraverso le quali si è provveduto ad implementare il diagramma E-R riportato. L'aderenza dei dati ricevuti ad uno schema che ANAC intende mantenere, anche in virtù di un possibile trasferimento dei dati verso la Piattaforma Digitale Nazionale Dati citata in precedenza, si rivela un'informazione importante, che permette alla modellazione proposta in questo capitolo e alle analisi effettuate nei capitoli successivi di rimanere coerenti con gli schemi dell'Autorità anche per il prossimo futuro. Di seguito, il lettore verrà accompagnato alla scoperta di ulteriori modi per modellare questi dati, descrivendo alcune tecniche di modellazione avanzata che attualmente sono solo in corso di sperimentazione da parte di alcuni tecnici e di alcune startup[6]. Inoltre, si è ritenuto importante dedicare qualche pagina alla modellazione *linked data*, poiché rappresenta la possibilità di modellare i dati non solo dal punto di vista strutturale ma semantico, liberando così tutto il potenziale che l'interoperabilità semantica è in grado di generare.

3.2. VERSO UNA MODELLAZIONE LINKED DATA

3.2.1. I contratti pubblici nel web semantico - Nel 2001, Tim Berners Lee pubblicò un articolo che negli anni a venire divenne fondamentale per lo sviluppo dell'informatica e in particolar modo del *web*. Nel famoso *"The Semantic Web"* [6], Berners Lee assieme ai colleghi James Hendler e Ora Lassila, immaginava *"a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities."* Dietro a questa idea di costruire una "nuova forma di *web*", vi era (come vi è tutt'ora), l'idea di rendere il web capace non solo di essere compreso dagli esseri umani, ma di far sì che imparasse un linguaggio ad uso esclusivo delle macchine, in modo da rendere quest'ultime in grado di comprendere il significato delle informazioni che avrebbero dovuto gestire. *"Most of the Web's content today is designed for humans to read, not for computer programs to manipulate meaningfully.*

[6] https://synapta.it/

Computers can adeptly parse Web pages for layout and routine processing—here a header, there a link to another page—but in general, computers have no reliable way to process the semantics."

Per rispondere a queste esigenze nel corso degli anni sono nate e si sono sviluppate delle tecnologie in grado di dare una semantica all'informazione contenuta nel web. Trale più importanti, sicuramente vi è il linguaggio XML (eXtensible Markup Language) e il framework RDF (Resource Description Framework), lo strumento proposto dal W3C[7] per la codifica, lo scambio e il riutilizzo di metadati strutturati in modo da consentire l'interoperabilità semantica tra applicazioni che condividono informazioni attraverso il web. L'idea proposta nel lavoro citato in apertura, ha così dato vita al cosiddetto *"semantic web"* (web semantico), che ha spronato l'informatica a costruire delle tecniche sempre più sofisticate di *knowledge representation*.

Ai fini di questo lavoro, si ritiene utile illustrare una tecnica di *knowledge representation*, che prende il nome di ontologia. Un'ontologia è una rappresentazione formale e condivisa di un particolare dominio di interesse, strettamente imparentata con la logica al prim'ordine. Infatti un'ontologia è anche descrivibile come una teoria assiomatica del prim'ordine, ovvero come un insieme di assiomi da cui dedurre delle conseguenze. La costruzione di ontologie in grado di modellare specifici domini alimentano il web semantico, aumentando di conseguenza la capacità dei calcolatori di modellare nuova conoscenza. In questo modo, le macchine sono in grado di trarre delle conclusioni sotto forma di nuova conoscenza (attività solitamente tipica della mente umana). Per il teorema di completezza di Gödel [4], lavorare alla costruzione di ontologie in grado di modellare in maniera opportuna un particolare dominio del mondo, permette ai calcolatori di fare delle deduzioni in maniera automatica sui dati che processano.

Un'ontologia suddivide una porzione di conoscenza in classi, proprietà e restrizioni, similmente a quello che fa la modellazione di un *database* in uno schema entità - relazione. In questo modo, la modellazione ontologica di un determinato dominio, alimentato da una corposa banca dati come la BDNCP, aiuta ad organizzare i dati sulla base del ruolo giocato all'interno del dominio, permettendo così alle macchine di distinguere la natura dell'informazione processata. Inoltre, essendo le ontologie dei modelli condivisi, la modellazione della BDNCP secondo questi principi potrebbe valorizzare i dati all'interno di un modello condiviso ad esempio a livello europeo.

L'esigenza di uno standard ontologico condiviso per i dati pubblici, non ha solamente l'intento, seppur prioritario e fondamentale, di costruire un linguaggio di descrizione dei dati comune e riconosciuto da tutte le organizzazioni pubbliche, ma anche quello di garantire l'interoperabilità semantica tra diverse collezioni di dati. L'esigenza di un'interoperabilità semantica è stata sottolineata anche in Europa: la ritroviamo citata in diversi documenti tra cui il manuale *"e-Government Core*

[7]World Wide Web Consortium

Vocabularies handbook - Using horizontal data standards for promoting interoperability" [8] (si veda anche [8]), che evidenzia la facilità con cui oggi è possibile valicare i confini nazionali, per motivi personali o di lavoro (*" (...) in Europe, citizens and businesses increasingly live, work, and conduct business across borders."*) con la conseguente necessità di dovere utilizzare servizi pubblici in grado di dialogare tra paesi differenti (*" (...) their increased mobility must be supported by cross-border public ser- vices, such as the registration of a foreign branch, obtaining a licence to conduct business in another country, or getting a birth certificate."*)

Proprio per questo, la pubblica amministrazione di un paese, dev'essere in grado di recuperare agevolmente i dati di un cittadino che proviene da un paese diverso, attraverso l'utilizzo di applicativi software che pur dovendo elaborare informazioni provenienti da fonti variegate, condividono degli standard comuni.

Alla luce di queste premesse, anche l'Italia ha cominciato a lavorare alla creazione di una serie di ontologie per la pubblica amministrazione italiana, nell'intento di agevolare il lavoro di descrizione della sua organizzazione, ma soprattutto dei dati che essa possiede. È nata così OntoPiA[9], la rete di ontologie realizzata dal Governo, in stretta collaborazione con l'Agenzia per l'Italia Digitale (AgID) e il Team per la Trasformazione Digitale della Presidenza del Consiglio dei Ministri.

All'interno di questa rete di ontologie, è possibile individuare l'ontologia dei contratti pubblici (*Public Contracts Ontology*), che modella il dominio dei contratti pubblici italiani, fornendo una rappresentazione semantica di tutto il processo che definisce un appalto pubblico. Pertanto, un frammento di questa ontologia definisce anche i dati a disposizione di questa tesi.

3.2.2. OntoPiA: la rete di ontologie della pubblica amministrazione - OntoPiA è una collezione di ontologie e di vocabolari controllati basata sulle indicazioni dettate dal Piano Triennale per l'Informatica della Pubblica Amministrazione[10]. In particolare, il documento "Elenco di basi dati chiave"[11] sostiene che *"il primo passo da compiere nel percorso riguarda l'individuazione di basi di dati chiave da valorizzare per rispondere a bisogni della collettività, rendendole disponibili sotto forma di open data, facilmente ottenibili "in bulk" e/o interrogabili attraverso Application Programming Interface (API) e descritte sia a livello di metadatazione generale che a livello di dati con chiari modelli condivisi, allineati ad altri già esistenti a livello Europeo e nel Web."*

La necessità di aderire ad uno standard adatto alla modellazione delle basi di dati pubbliche italiane, nasce dalla presa di coscienza da parte delle istituzioni che una frammentazione tecnologica come quella attuale non sia più sostenibile.

[8] https://op.europa.eu/en/publication-detail/-/publication/b3e2b008-f516-431a-8d22-38aa99a74360
[9] https://github.com/italia/daf-ontologie-vocabolari-controllati/blob/master/README.md
[10] https://pianotriennale-ict.italia.it/
[11] https://docs.italia.it/italia/daf/pianotri-elencobasidatichiave/it/stabile/index.html

"La maggior parte delle basi di dati pubbliche oggi esistenti è stata progettata e realizzata in modo distinto, senza il supporto di una visione d'insieme utile a indirizzare azioni normative e tecniche in grado di favorire la qualità dei dati. Questa caratteristica ha prodotto nel tempo la frammentazione del patrimonio informativo della Pubblica amministrazione in veri e propri silos informativi: "contenitori" in cui i dati sono spesso replicati e memorizzati in modo disomogeneo o addirittura incoerenti e disallineati tra loro."[12]

OntoPiA si propone dunque di colmare questa assenza di uniformità, proponendo un *framework* di modelli concettuali con cui descrivere le basi di dati più significative della pubblica amministrazione italiana.

I principi su cui si basa OntoPiA sono quattro, e sono:

1. facilitare lo sviluppo di nuovi sistemi informativi, beneficiando di un modello di base da cui partire senza reinventarsi ogni volta la ruota;
2. agevolare lo scambio di dati, beneficiando di un linguaggio comune;
3. abilitare l'integrazione tra dati provenienti da sorgenti diverse;
4. standardizzare i dati.

Inoltre, la rete di ontologie presenti all'interno di OntoPiA, si basa sui cosiddetti principi *fair*, i principi di "giustizia" europei[13] recepiti di recente anche dall'Italia. In questo senso le ontologie sviluppate all'interno di OntoPiA dovranno essere:

1. *findable*: attraverso l'utilizzo di URI permanenti per identificare concetti e relazioni nella rete di ontologie e termini nei vocabolari controllati;
2. *accessibile*: attraverso l'utilizzo di protocolli standard aperti per l'accesso sul Web (come ad esempio l'HTTP(S) o, per l'interrogazione dei dati SPARQL*);
3. *interoperable*: attraverso l'utilizzo di protocolli standard aperti per modellare i dati come ad esempio RDF o OWL;
4. *reusable*: tutte le ontologie e i vocabolari devono essere pubblici, con licenza aperta (CC-BY 4.0) e collegati ad altre ontologie disponibili nel "web dei dati".

La rete di ontologie raccolte in OntoPiA è suddivisa in più livelli.

- Le ontologie di *livello core* identificano le ontologie delle persone, dei luoghi, e delle organizzazioni. Ogni realtà pubblica, tendenzialmente, necessita di tutte o alcune di queste ontologie.

[12] https://pianotriennale-ict.italia.it/

[13] https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_0.pdf

- Le ontologie verticali, rappresentano le cosiddette *ontologie di dominio*: fanno parte di questa categorie i luoghi della cultura, gli eventi, le strutture ricettive, i contratti pubblici e tante altre.

- Infine, vi è un *livello di supporto*: fanno parte di questo livello le ontologie sulle unità di misura, sulla modellazione del tempo, sui ruoli delle persone, sulle lingue ed altro.

Tutte queste ontologie utilizzano dei vocabolari controllati che sono metadatati secondo standard europei. L'utilizzo di vocabolari controllati diventa fondamentale per rendere più efficaci le analisi. Un vocabolario controllato su cui si sta lavorando in riferimento all'ontologia dei contratti pubblici è quello riferito al vocabolario comune per gli appalti pubblici (CPV)[14], un sistema di classificazione che permette la descrizione dell'oggetto degli appalti. Costruire un elenco di parole chiave e definire di conseguenza un modo univoco e condiviso per descrivere l'oggetto dei bandi diventa molto utile per eventuali elaborazioni o attività di *Natural Language Processing* (NLP).
Il contenuto di OntoPiA inoltre, è stato messo a disposizione sulla piattaforma GitHub[15], in modo tale che chiunque possa contribuire al miglioramento delle ontologie già presenti, oltre a proporne di nuove.
La strada intrapresa da questo ambizioso progetto mira a consolidare, alimentare ed arricchire quello che viene comunemente chiamato grafo della conoscenza (*knowloedge graph*) dei dati della pubblica amministrazione.

3.2.3. L'ontologia dei contratti pubblici - All'interno di OntoPiA è descritta l'ontologia dei contratti pubblici, chiamata "PublicContract-AP_IT". Questa ontologia aderisce agli standard nazionali, ed è costitutita da triple RDF che la rendono conforme al paradigma *linked data*. La BDNCP potrebbe essere quindi modellata da questa ontologia, che, a differenza della modellazione entità-relazione proposta nelle pagine precedenti, conferirebbe un carattere semantico ai dati in esame.
L'ontologia è caratterizzata da qualche decina di *classi*, di cui segue un elenco esemplificativo non esaustivo: "Accordo Quadro", "Aggiudicazione", "Appalto", "Bando di Gara", "Concessione", "Documento di Offerta", "Documento di Variante", "Gara", "Lotto", "Pagamento", "Procedura di Scelta del Contraente", "Pubblicazione", "Stato avanzamento lavori" e "Variante". Le classi rappresentano le entità del modello. Vi è poi un elenco di *object property*, come "Allega offerta", "Ha accordo quadro", "Ha CPV", "Ha modalità di scelta contraente" e altre. Le *object property* collegano fra loro diverse classi stabilendone una determinata relazione. Infine, sono elencate e descritte le *data property*, come ad esempio "Attuale importo lavori", "Aumento importo", "Data di pubblicazione", "Data variante", "Diminuzione importo", "Importo base d'asta", "Percentuale di ribasso" e così via.

[14] https://simap.ted.europa.eu/it/cpv
[15] https://github.com/italia/daf-ontologie-vocabolari-controllati

Attraverso le *data property* vengono specificati gli attributi delle classi descritte dall'ontologia. La *data property* "importo base d'asta", ad esempio, ha come dominio la classe "Progetto di approvigionamento", e come codominio il tipo di dato *float*, per indicare che un dato importo è riferito ad un progetto (o equivalentemente ad una gara) ed è determinato da un valore di tipo *float*.
La classe "Procedura" riportata in figura 3.3 è specializza nella classi "Concessioni", "Amministrazione diretta", "Accordo quadro", "Partnership pubblico privato" e "Appalto". La procedura è messa in relazione con una stazione appaltante definita dalla classe "Agente". La classe "Agente" si collega ad un'altra ontologia definita a livello più basso dello *stack*.

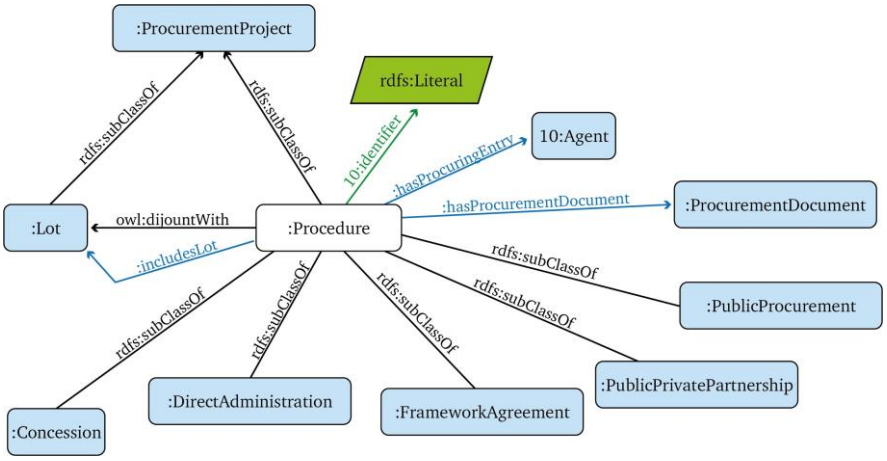

Figura 3.3: *La classe "Procedura" e le sue relazioni*

La classe "Lotto" rappresentata dalla figura 3.4, è definita da un CIG, da un CPV ed è associata ad una procedura. Si ricorda che il lotto è un frazionamento di una gara.

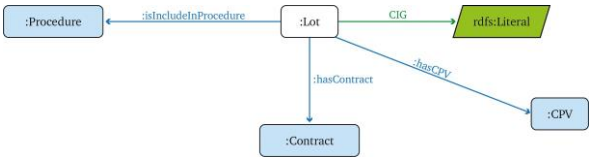

Figura 3.4: *La classe "Lotto" e le sue relazioni*

Infine, si riporta il frammento di ontologia riferito alla classe "Contratto":

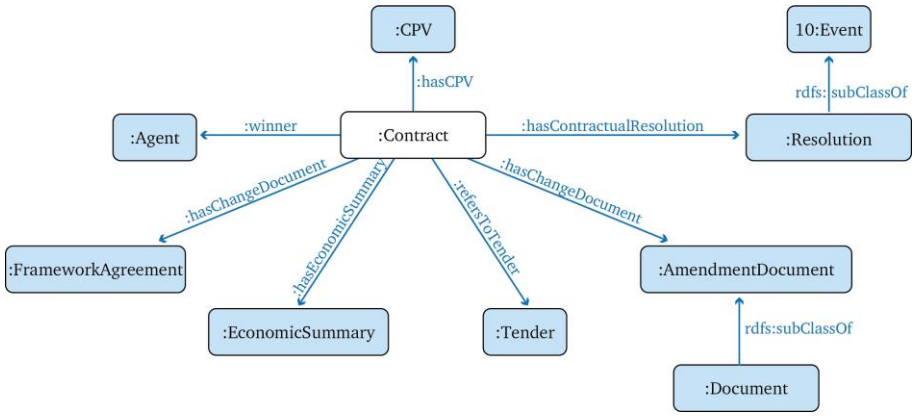

Figura 3.5: *La classe "Contratto" e le sue relazioni*

Nell'immagine 3.6 riportata in calce di questo capitolo è rappresentata l'intera ontologia "PublicContract-AP_IT" disegnata con il *tool* automatico WebVOWL - Web-based Visualization of Ontologies[16]. L'immagine non permette di visualizzare nel dettaglio il contenuto dei nodi e le relazioni distribuite lungo gli archi (lo si può fare però collegandosi all'indirizzo riportato nel piè di pagina[17]), tuttavia mostra l'elevato numero di legami (gli archi) che collegano le varie classi dei contratti pubblici (i nodi).

3.3 Prospettive future - L'ontologia dei contratti pubblici qui presentata non è ancora diventata uno standard riconosciuto e utilizzato da ANAC per modellare i suoi dati, tuttavia avrebbe bisogno solo di essere stabilizzata, ovvero di essere riconosciuta come riferimento ufficiale, per poi essere utilizzata su progetti e dati reali. Tra i lavori più importanti che ANAC potrebbe prendere come riferimento vi è sicuramente quello rappresentato dall'*Open Contracting Data Standard*, un'iniziativa di sviluppo di standard emessa dalla rete Omidyar[18], che si definisce una "società di investimento filantropica". *Open Contracting* rappresenta uno standard per la pubblicazione di informazioni strutturate in modalità *open data* su tutte le fasi di un processo di appalto o di investimento pubblico: dalla pianificazione alla sua realizzazione finale.

In chiusura di questo capitolo si ritiene importante riportare la *mission* dell'organizzazione *Open Contracting Partnership*[19], ritenendola adeguata a chiudere questa parte di trattazione, nella speranza di stimolare le autorità competenti a intraprendere con convinzione la strada dei *linked data*.

[16]http://vowl.visualdataweb.org/webvowl.html
[17]http://www.visualdataweb.de/webvowl/#iri=https://w3id.org/italia/onto/PublicContract [18]https://www.omidyar.com/
[19]https://www.open-contracting.org/

"One in every three dollars spent by government is on a contract with a company. Public contracting is the world's largest marketplace, covering $10 trillion of spending every year. It is the bricks and mortar of public benefit where the vital goods, works, and services for us all are purchased. Yet, too many governments don't seem to know what they are buying and selling, for how much, when and with whom they are dealing. And it's government's number one corruption risk. Open contracting can change all this. We can transform how business is done by engaging stakeholders across government, business and civil society to collaborate on reforms, engage users, respond to feedback and to create open data & tools to drive systematic change. A modern economy needs a smart, data-driven government contracting ecosystem. We bring governments, businesses, and citizens together to build one."

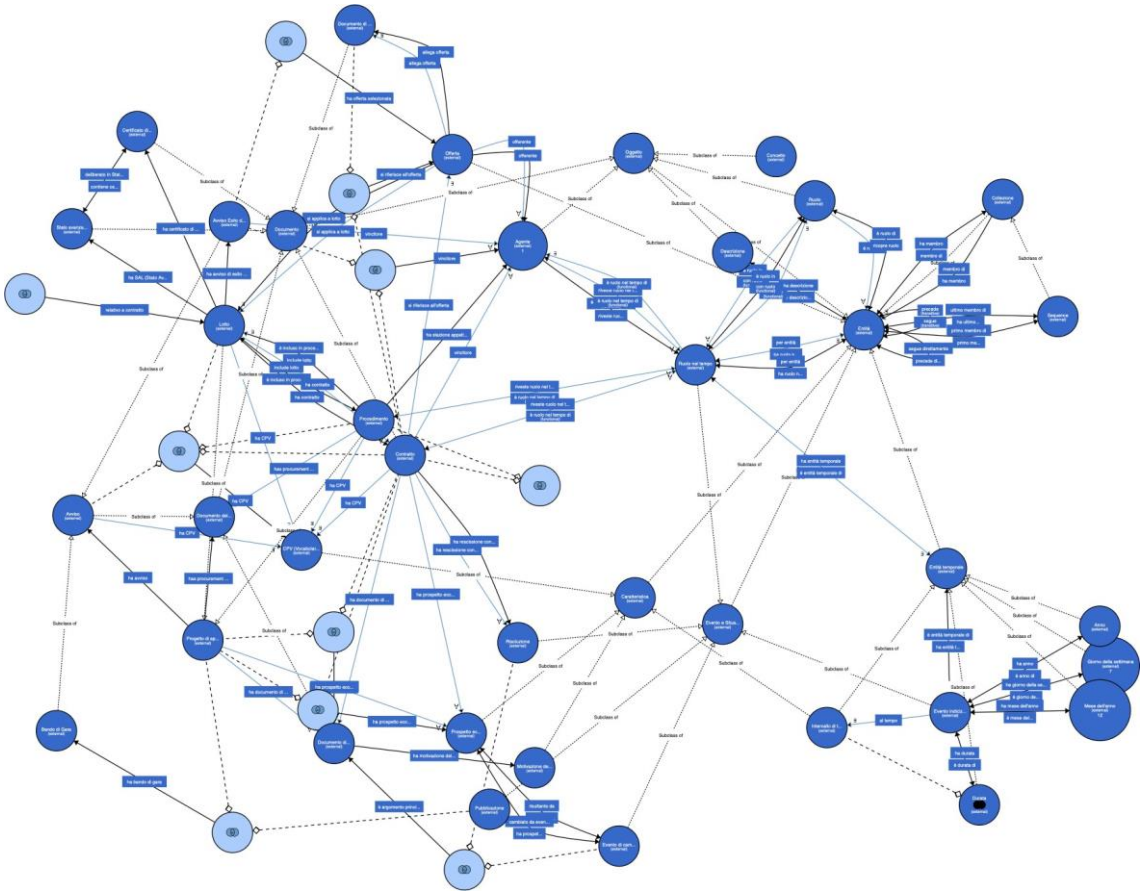

Figura 3.6: *L'intera ontologia dei contratti pubblici italiani*

4. ANALIZZARE I DATI

4.1. Analisi preliminari - In questo capitolo il lettore verrà guidato nel percorso di *data analysis* condotta sui dati in esame. Le analisi che seguono sono state eseguite con il linguaggio R, all'interno dell'ambiente di sviluppo RStudio. La versione di R utilizzata è la 3.6.2, rilasciata in data 12.12.2019. Le prime operazioni svolte sui dati hanno riguardato la loro pulizia, seguendo parte di quanto suggerito in [39]. L'analisi statistica condotta, segue lo schema proposto in [15] e riportato in figura 4.1.

I dati ricevuti sono stati sottoposti a diverse operazioni di pulizia. In particolare, si è rivelato necessario effettuare diversi *type checking* sui dati, per poterli elaborare efficaciemente tramite le funzioni e le librerie di R. Un'ulteriore attività di *data cleaning* condotta sui dati, ha riguardato l'esplorazione dei dati mancanti (*missing data*), la quale ha evidenziato delle aree critiche nelle tabelle di input. Per mostrare al lettore queste aree critiche, attraverso il pacchetto visdat[1] di R, è stato possibile costruire le *heatmap* riportate in figura 4.2 e 4.3. Le elaborazioni proposte nelle due figure, riportano l'analisi condotta sui dati grezzi, e non sulle tre tabelle da essi ricavate. D'ora in avanti si farà riferimento ai dati nella seguente maniera: "ds_appalti", per il *file* "appalti_senza_oggetto.csv", "ds_cigcup", per il *file* "cig_cup.csv", "ds_oggetto gare", per il *file* "gara_lotto_oggetti.csv" e infine "ds_aggiudicatari" per il *file* "aggiudicatari.csv".
Nella figura 4.2 viene visualizzata la distribuzione dei valori "NULL" all'interno del *dataset* "ds_appalti". Sono state colorate di viola le celle che *non* contengono un valore uguale a "NULL". Il numero di celle valorizzate con un valore diverso da "NULL", ammontano, in percentuale, al 53,8%. Viceversa, le celle colorate di arancione, rappresentano celle contenenti il valore "NULL". Esse ammontano al 46,2% del totale. Infine, la mappa evidenzia una piccola presenza di celle contenenti il *placeholder* "NA".

[1] https://cran.r-project.org/web/packages/visdat/index.html

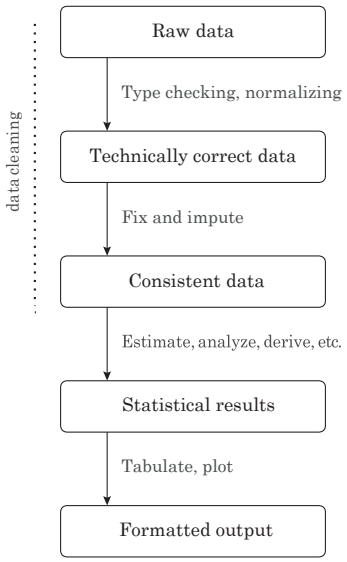

Figura 4.1: *Statistical analysis value chain*

Come si evince dalla figura, i valori "NA" sono maggiormente concentrati lungo la colonna dell'attributo "CIGAccordoQuadro", mentre i valori "NULL" sono distribuiti in zone diverse della tabella, diffondendosi, laddove presenti, lungo tutta la relativa colonna.

La stessa analisi condotta sui *dataset* "ds_cigcup"e "ds_oggettogare", restituiscono una totale assenza di valori "NULL" o "NA". Si omette pertanto la loro rappresentazione grafica.

Per quanto riguarda invece il dataset "ds_aggiudicatari" riportato in figura 4.3, la percentuale di dati "NULL" ammonta al 12,1%, mentre la colonna "Ruolo" è totalmente caratterizzata da valori mancanti "NA".

In conclusione, il *dataset* "ds_appalti" è quello in cui l'informazione mancante è presente in percentuale maggiore.

I dati a cui fanno riferimento le visualizzazioni appena descritte, corrispondono a quelli che in figura 4.1 vengono etichettati come *raw data*. Per passare ai *technically correct data* è stata implementata in R la modellazione concettuale descritta dal diagramma E- R del paragrafo 3.2, permettendo così l'ottenimento delle tre nuove tabelle denominate "tab_staz_appaltanti", "tab_gare" e "tab_aggiudicatari". La disponibilità di ulteriori dati ufficiali in formato aperto, all'interno del portale "Indice delle Pubbliche Amministrazioni (iPA)"[2], ha permesso di arricchire la tabella relativa alle stazioni appaltanti di una serie di campi utili. In particolare, per ciascuna stazione appaltante presente nei dati d'origine, è stato aggiunto il suo codice d'avviamento postale, il comune, la provincia e la regione dove ciascuna stazione appaltante è dislocata, oltre al nome e cognome della figura responsabile. La possibilità di accedere a questi attributi, ha permesso di ricavare alcune visualizzazioni, come quella di figura 4.4, dove sono riportate le 10 stazioni appaltanti che hanno contratto più gare nell'anno 2015.

[2]indicepa.gov.it

Distribuzione dei valori "NULL" e "NA" nel dataset "ds_appalti"
In arancione i valori "NULL", in viola i valori presenti.

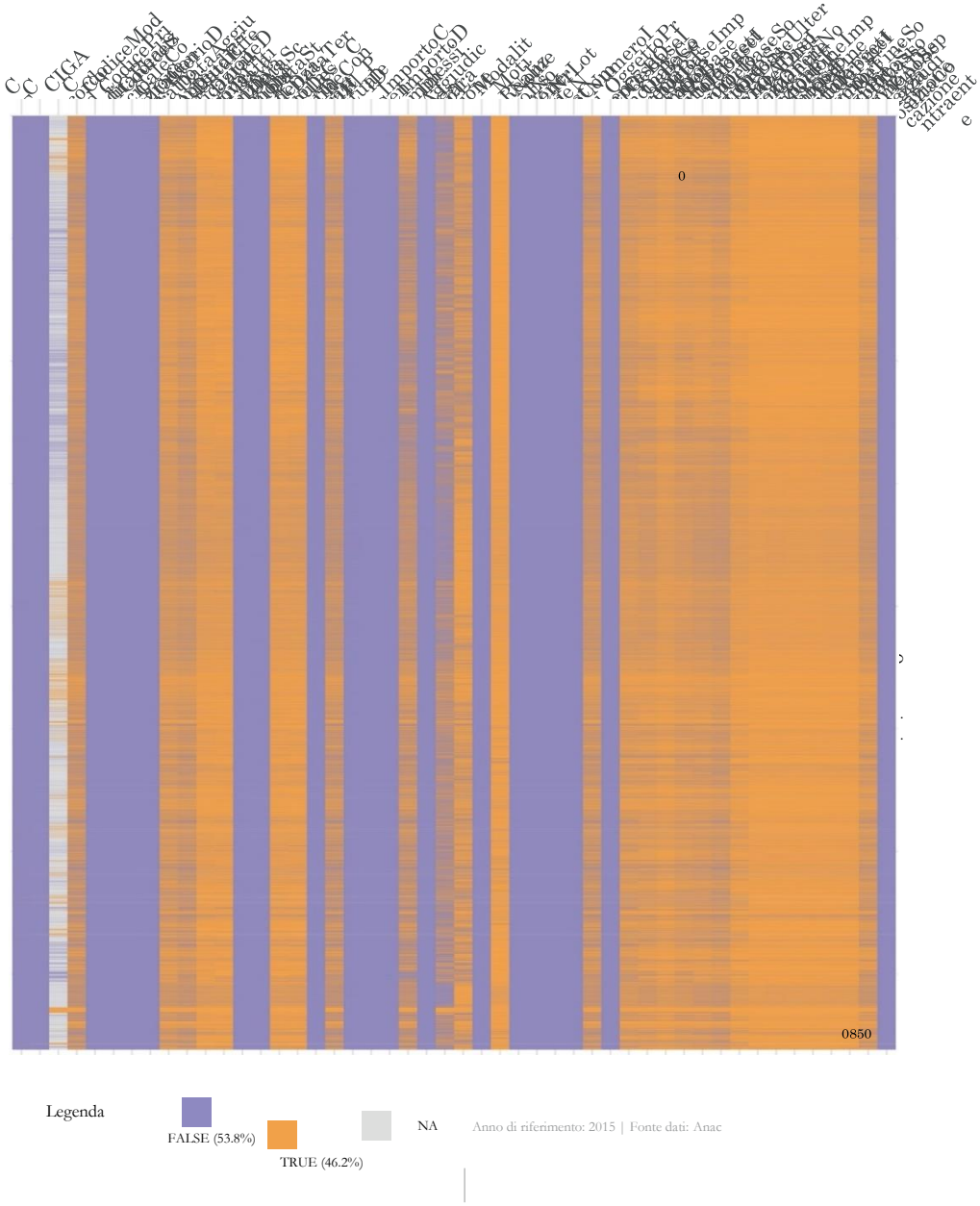

Figura 4.2: *Distribuzione dei dati mancanti nel dataset "ds_appalti"*

Distribuzione dei valori "NULL" e "NA" nel dataset "ds_aggiudicatari"
In arancione i valori "NULL", in viola i valori presenti.

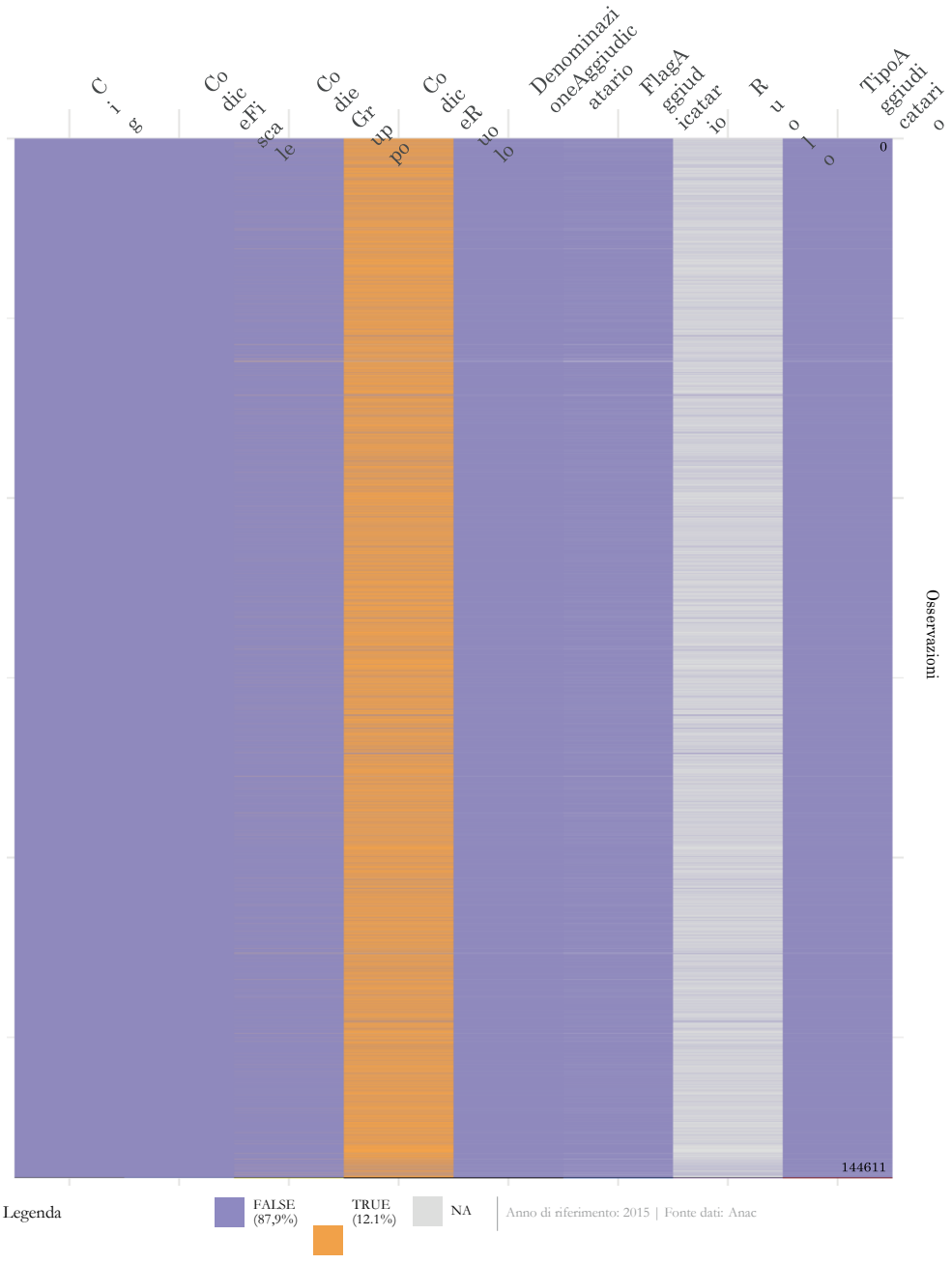

Figura 4.3: *Distribuzione dei dati mancanti nel dataset "ds_aggiudicatari"*

Questo e i successivi risultati, rappresentano degli strumenti utili per prendere dimestichezza con la mole di informazione a disposizione. Si è deciso di condurre queste analisi preliminari esclusivamente sui dati relativi all'anno 2015. Il codice prodotto però, disponibile a chiunque ne fosse interessato in formato open source, permette di operare anche su dati di input diversi da quelli selezionati in questo caso, come quelli riferiti agli anni 2016 e 2017, oppure a dati riferiti ad anni successivi al 2017, purché caratterizzati dalla stessa struttura.

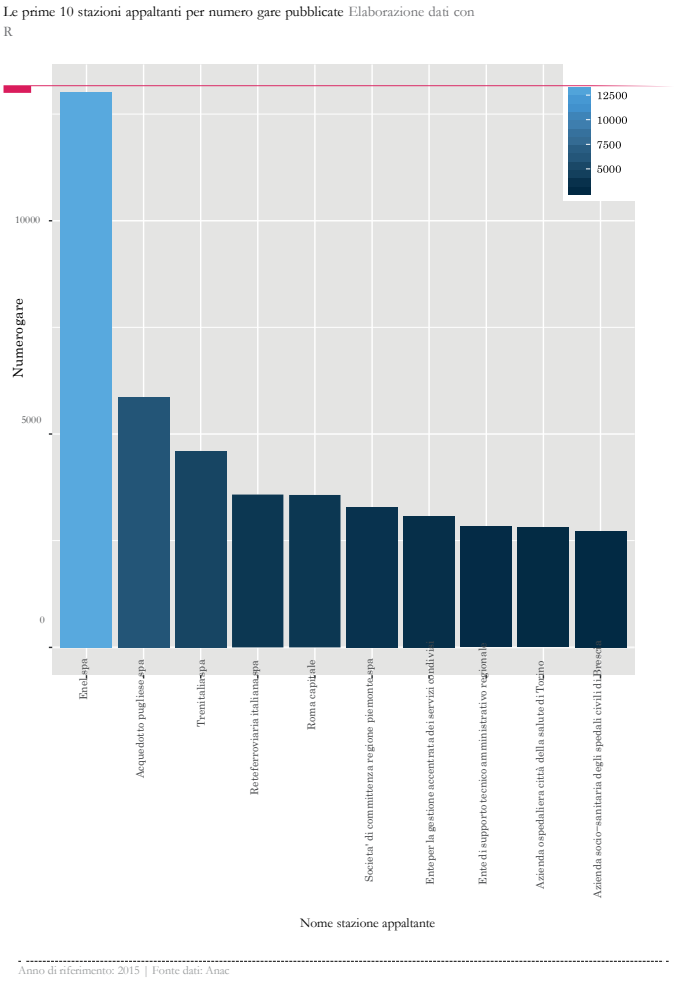

Figura 4.4: *Le prime dieci stazioni appaltanti*

Di seguito è riportata un'analisi della distribuzione delle stazioni appaltanti per province, sulla base del numero di gare pubblicate in un determinato anno. Il risultato è riportato in figura 4.5. La città di Roma risulta la città con il maggior numero di gare pubblicate nell'anno 2015.

Nella figura 4.6 invece, è riportata la suddivisione delle tipologie di scelta del contraente utilizzate dalle stazioni appaltanti. Le tipologie di scelta del contraente, sono delle categorie che determinano vincoli e modalità con cui una stazione appaltante può individuare l'aggiudicatario.

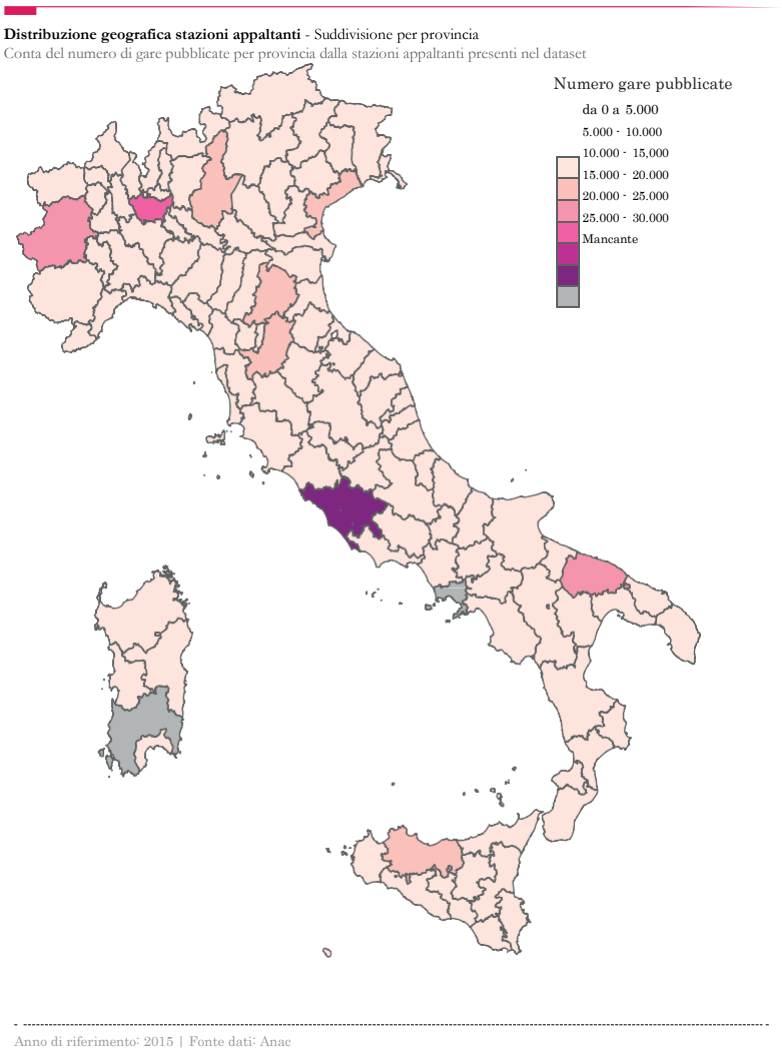

Figura 4.5: *Distribuzione stazioni appaltanti sulla base del numero di gare pubblicate*

Attraverso il grafico riportato, è facile osservare come rispetto ad una lunga lista di tipologie, sono solo poche quelle che realmente incidono sul totale. Le principali tipologie, in questo caso, sono:

- l'*affidamento in economia - affidamento diretto*;
- l'*affidamento diretto in adesione ad accordo quadro/convenzione*;
- la *procedura aperta*;
- l'*affidamento in economia - cottimo fiduciario*;
- la *procedura negoziata senza previa pubblicazione*;

- la *procedura negoziata senza previa indizione di gara.*

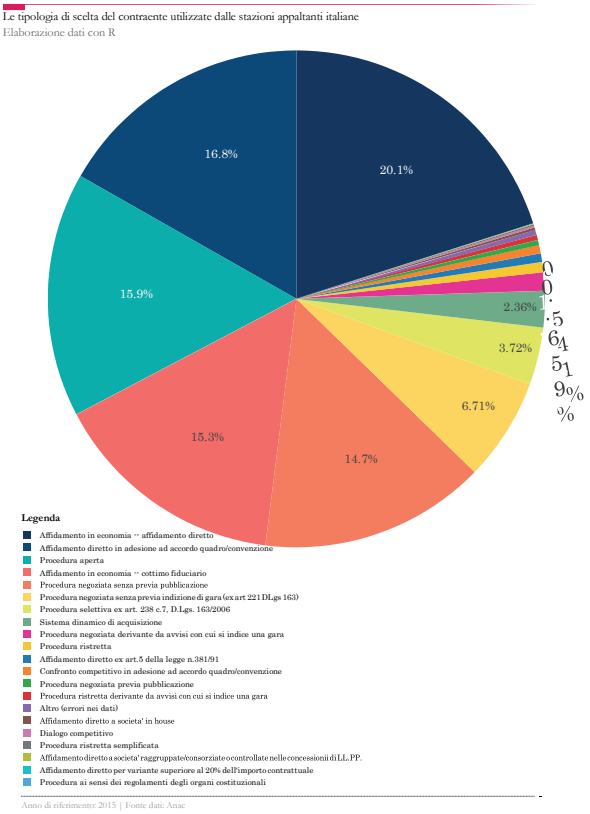

Figura 4.6: *Distribuzione delle tipologie di scelta del contraente*

A completamento delle visualizzazioni grafiche appena discusse, si è proceduto all'esplorazione statistica di alcuni attributi rilevanti. Le tabelle che seguono riguardano gli attributi inerenti agli importi delle gare, in diversi momenti del processo di aggiudicazione. Si riportano alcuni indici statistici degli attributi *"Importo complessivo della gara"* e *"Importo di aggiudicazione"*.

Per quanto riguarda l'attributo *"Importo complessivo della gara"*, i risultati sono riportati nella tabella che segue[3]:

Minimo	Primo quartile	Mediana	Media	Terzo quartile	Massimo
0,1 e	420 e	6.000 e	156.500e	19.000 e	344.200.000 e

Per quanto riguarda invece l'attributo *"Importo di aggiudicazione"*, i risultati sono riportati nella tabella che segue:

[3]I valori di minimo e di massimo delle tabelle riportate sembrano non essere veritieri, e pertanto si ritiene che l'errore possa essere ricondotto a degli errori di inserimento del dato all'interno della banca dati.

Minimo	Primo quartile	Mediana	Media	Terzo quartile	Massimo
-1 e	41.950e	87.300e	671.379e	210.000e	6.388.017.663 e

Queste analisi preliminari hanno lo scopo di mostrare in che modo l'informazione racchiusa nei dati può essere valorizzata e indirizzata ad arricchire i risultati che ci si pone di ottenere. L'utilizzo di visualizzazioni grafiche, di tabelle riassuntive o il calcolo di alcuni indici statistici in riferimento a determinati attributi, da un lato rappresentano degli strumenti utili da consegnare alle istituzioni competenti, affinché possano meglio indirizzare le loro politiche. Dall'altro aiutano e completano le analisi sugli indicatori di corruzione a cui è dedicato il prossimo capitolo, fornendo dei valori "di riferimento" a cui potersi ricondurre. In generale, l'analisi esplorativa di uno o più *dataset*, è sempre utile ed importante per muoversi verso analisi più avanzate.

4.2. Analisi degli indicatori di corruzione - In questa sezione viene illustrata un'analisi critica degli indicatori di corruzione proposti nel paragrafo 2.2.3. Nella prosecuzione della trattazione verranno usati i termini "indicatore" o "indice" in maniera interscambiabile. I paragrafi successivi sono caratterizzati da uno schema comune: in una prima fase viene riportata la formulazione dell'indicatore *i*-esimo così come proposto da ANAC, mentre successivamente viene riportata un'analisi dell'indicatore volta ad individuarne i punti di forza e i punti di debolezza. Nei casi in cui è stata individuata l'opportunità di migliorare l'efficacia dell'indicatore in esame, questa è stata argomentata nell'apposito paragrafo e implementata di conseguenza in R. Inoltre, per ogni indice studiato, è stato descritto il modo con cui è stato implementato e la sua relativa "*funzione punteggio*" utile al calcolo del *rischio di corruzione*, una funzione che normalizzerà i valori degli indici in un intervallo continuo [0,1]. Infine, per alcuni indicatori studiati, sono state riportate alcune analisi di contorno.

4.2.1 Indicatore I_{oepv}

Il primo indicatore studiato riguarda le procedure che utilizzano il criterio dell'*offerta più vantaggiosa* (OEPV). L'indicatore si propone di misurare il numero dei bandi della *i*-esima stazione appaltante al tempo *t*, che utilizzano il criterio dell'offerta economicamente più vantaggiosa ($NTPOEPV_{i,t}$) in rapporto al numero totale delle procedure di appalto utilizzate dalla *i*-esima stazione appaltante al tempo *t* ($NTP_{i,t}$).

Nel misurare questo primo indice, si è scelto di studiare esclusivamente le gare marcate come "aggiudicate" all'interno del *dataset*, nell'intento di ricavare una partizione di gare in cui gli attributi sia-

$$I_{oepv} = \frac{NT\,POEPV_{i,t}}{NT\,P_{i,t}}$$

no valorizzati in misura maggiore, in virtù del fatto che la gara (aggiudicata), ha raggiunto tutte le fasi finali di questo primo *iter*, popolando ad esempio i campi "criterio di aggiudicazione", "importo di aggiudicazione", "data di aggiudicazione definitiva" e così via.

Per cominciare, è stato effettuato un controllo del campo "criterio di aggiudicazione", che secondo quanto previsto dall'indice può essere valorizzato con il valore *"offerta economicamente più vantaggiosa"*, oppure con il valore *"prezzo più basso"*. Dalle analisi è emerso che nonostante gli appalti aggiudicati nell'anno considerato[4], ammontino a 125.011, il 32% del totale (39.735) non ha valorizzato il campo "criterio di aggiudicazione", che essendo marcato come "NULL", costringe ed eliminare le relative righe, pena l'ottenimento di risultati distorti; ne deriva così un dominio ridimensionato. Il valore delle gare prive di un criterio di aggiudicazione, ammonta a 25.340.910.813e (pari al 25% del totale annuo, nell'anno considerato), una cifra piuttosto significativa, che verrà però esclusa dalle analisi, non potendo stabilire a priori il criterio con cui sia stato aggiudicato questo importo. Gli appalti considerati invece, rappresentano un valore di 74.439.212.811e (pari al 75% del totale annuo, nell'anno considerato).

In riferimento al calcolo dell'indicatore I_{oepv}, si è ritenuto opportuno apportare alcune modifiche, alla luce di alcune osservazioni preliminari esposte di seguito.

Anzitutto, il semplice calcolo dell'indicatore così come proposto, non fornisce sufficienti indicazioni sulle modalità di valutazione dei risultati ottenuti. Si immagini di calcolare questo indicatore su un numero arbitrario di stazioni appaltanti e di estrarre infine il risultato dell'indicatore di una di queste stazioni appaltanti. Appurato il suo valore, (pari ad esempio al valore 16), non vi è modo di stabilire se quel valore sia un valore buono per la stazione appaltante considerata o un valore negativo, né se è da considerarsi attendibile in riferimento agli altri valori misurati. Inoltre, il valore così calcolato, potrebbe condurre il lettore a delle conclusioni avventate. A tal proposito si immagini che una determinata stazione appaltante abbia contratto 10 diverse gare, e di queste, 8 siano state aggiudicate tramite il criterio dell'"*offerta economicamente più vantaggiosa*", mentre le restanti 2 attraverso il criterio del "*prezzo più basso*". L'indicatore, calcolato su questa specifica istanza, ammonterebbe in percentuale all'80% (dal rapporto 8/10). Un valore così prossimo al totale sembrerebbe esprimere una tendenza elevata ad aggiudicare degli appalti tramite il criterio dell'*"offerta economicamente più vantaggiosa"*. Se l'intento è quello di costruire degli indicatori che ci segnalino le gare (o equivalentemente le stazioni appaltanti) più esposte al rischio di corruzione, il risultato riportato nell'esempio collocherebbe quella gara all'interno delle gare da monitorare.

[4] 2015

L'indicatore senz'altro calcola il rapporto che esprime, ma in quel rapporto non vi è traccia del valore economico delle gare considerate. Se le 2 gare dell'esempio aggiudicate attraverso il criterio del *"prezzo più basso"*, possedessero un importo di aggiudicazione superiore alla somma degli importi di aggiudicazione delle gare contratte tramite il criterio dell'*"offerta economicamente più vantaggiosa"*, l'indicatore così come proposto non sarebbe in grado di tenere conto di questo. Pertanto si è ritenuto opportuno riscrivere l'indicatore I_{oepv} nella seguente formula:

$$I_{oepv2} = \frac{IT\ POEPV_{i,t}}{IT\ P_{i,t}}$$

dove $IT\ POEPV_{i,t}$ rappresenta l'*importo* delle procedure aggiudicate dalla *i*-esima stazione appaltante al tempo *t* tramite il criterio dell'*"offerta economicamente più vantaggiosa"*, mentre $IT\ P_{i,t}$, rappresenta l'*importo* totale delle procedure contratte dalla *i*-esima stazione appaltante al tempo *t*. La nuova formulazione proposta, tiene conto degli importi totali dei due criteri di aggiudicazione considerati. Il rapporto viene quindi pesato in funzione del valore economico delle gare. Nelle analisi condotte è stato utilizzato solamente l'indicatore sugli importi, escludendo quello espresso da ANAC. Una volta calcolato il valore dell'indicatore per tutte le gare bandite dalle stazioni appaltanti considerate, sono stati effettuati alcuni calcoli statistici.

La media dell'indicatore ammonta a 0,28. Questo significa che in media, nell'anno considerato, le stazioni appaltanti hanno destinato il 28% del proprio denaro a gare aggiudicate tramite il criterio dell'*"offerta economicamente più vantaggiosa"*. La deviazione standard invece dell'indicatore, ammonta a 0,39, denotando una significativa variabilità nei dati analizzati.

Dopo aver calcolato l'indicatore, si è proceduto alla costruzione di una "funzione punteggio", in grado di assegnare un punteggio reale compreso nell'intervallo *0* e *1* a ciascuna stazione appaltante, sulla base del valore dell'indicatore. La "funzione punteggio" verrà calcolata per tutti gli indicatori presentati e discussi in questa tesi. Essa avrà il compito di trasformare il valore dell'indicatore in un nuovo valore, sul quale si andrà poi a costruire un *ranking* delle stazioni appaltanti.

La "funzione punteggio" costruita in relazione all'indicatore *i* è così definita:

- la *prima parte* controlla se i valori dell'indicatore *i* sono maggiori o uguali ( ) della media dell'indicatore *i*.

```
if_else(risultato_indice_1$indice_01>=media_indice_01,
```

- se il controllo ha esito postivo, allora il valore dell'indicatore viene riscalato nel *range* [0, 1], attingendo al vettore "valori_sospetti", che contiene tutti e soli i valori dell'indicatore *i* con valore maggiore o uguale ( ) alla media.

```
round(rescale(risultato_indice_1$indice_01,to=c(0,1),
from=range(valori_sospetti,na.rm=TRUE,finite=TRUE)),2),
```

la *seconda parte* invece, assegna il valore 0 quando la guardia del primo if_else fallisce.

```
if_else(risultato_indice_1$indice_01<media_indice_01,0
```

Il calcolo della media, diventa una misura per pesare l'assegnamento del punteggio a ciascuna stazione appaltante. Essendo la media pari allo 0,28, ogni valore prossimo alla media è da considerarsi in linea con quanto, nell'anno considerato, le stazioni appaltanti hanno speso denaro tramite l'offerta economicamente più vantaggiosa. Pertanto si è deciso di mappare al valore 0 tutte le stazioni appaltanti con un valore dell'indicatore inferiore alla media, mentre si è scelto di mappare nell'intervallo continuo [0, 1] tutte le stazioni appaltanti con un valore dell'indicatore maggiore o uguale ( ) alla media. L'operazione di *rescaling* dei valori nell'intervallo [0, 1], ha posizionato lo 0 in corrispondenza del valore medio, e l'1 in corrispondenza del valore massimo misurato dagli indicatori.
La stampa dell'istogramma delle frequenze della colonna "punteggio", dell'indicatore
I_{oepv2} è riportata nella figura 4.7.
La predominanza del valore 1 è evidente. Un punteggio pari a 1 corrisponde a tutte quelle stazioni appaltanti che hanno contratto esclusivamente gare tramite il criterio dell'*"offerta economicamente più vantaggiosa"*. Non avendo svolto alcuna gara tramite il criterio del *"prezzo più basso"*, l'importo totale riferito a quest'ultimo criterio è nullo su tutte le righe. Si segnala inoltre, che la gran parte delle stazioni appaltanti che hanno contratto un punteggio pari a 1, hanno svolto, in media, un numero di gare pari a 2.
L'istogramma di figura 4.8 invece rappresenta tutti i punteggi con valore minore di 1.
Tutti i calcoli ed i punteggi finali sono stati memorizzati in una tabella intestata con le opportune variabili. La tabella viene popolata in automatico da una serie di algoritmi che attingono i dati dalle tabelle ufficiali descritte nel capitolo 3. Il calcolo del "punteggio" viene effettuato come ultima operazione, per poi essere aggiunto alla tabella di sintesi. Fa seguito la descrizione della tabella.

- **CF stazione appaltante**: contiene il codice fiscale della stazione appaltante;
- **N. gare totale**: riporta il numero di gare totali della stazione appaltante;
- **Importo totale**: riporta il totale aggiudicato dalla stazione appaltante;

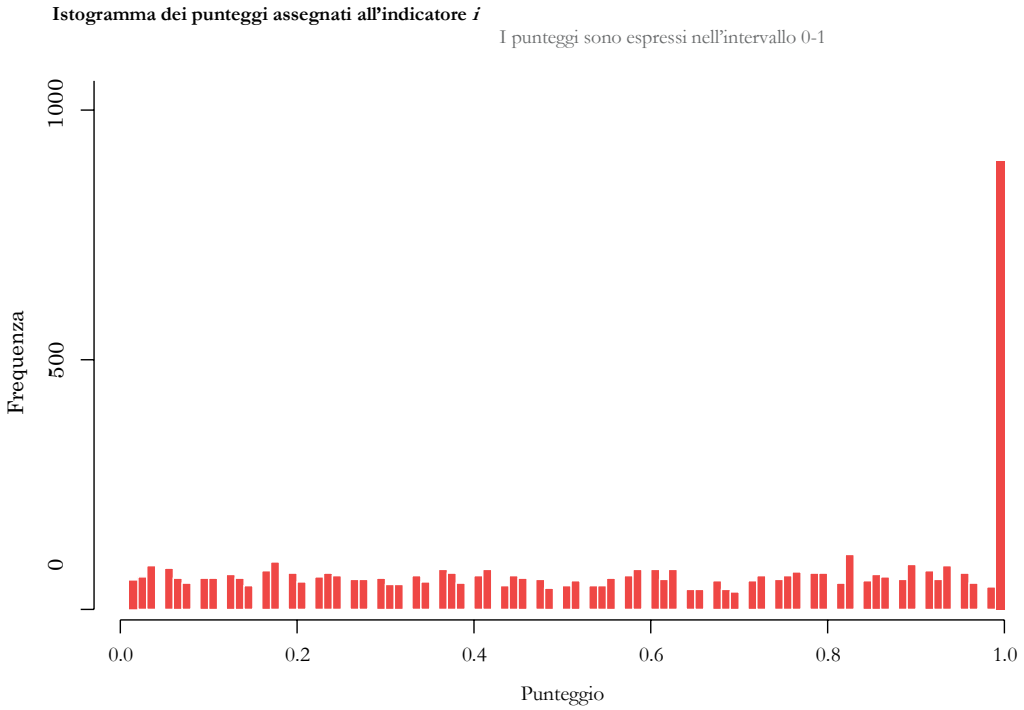

Figura 4.7: *Istogramma dei punteggi indicatore* I_{oepv2}

- **N. gare OEPV**: riporta il numero di gare contratte con il criterio OEPV;
- e**gare OEPV**: riporta il totale aggiudicato delle gare contratte con il criterio OEPV;
- **N. gare PPB**: il numero di gare contratte con il criterio PPB;
- e**gare PPB**: riporta il totale aggiudicato delle gare contratte con il criterio PPV;
- **Indice**: riporta il valore dell'indicatore I_{oepv2}, della stazione appaltante;
- **Punteggio**: riporta il valore del punteggio calcolato sull'indicatore I_{oepv2} per ciascuna stazione appaltante.

[5]Offerta economicamente più vantaggiosa.
[6]Prezzo più basso.

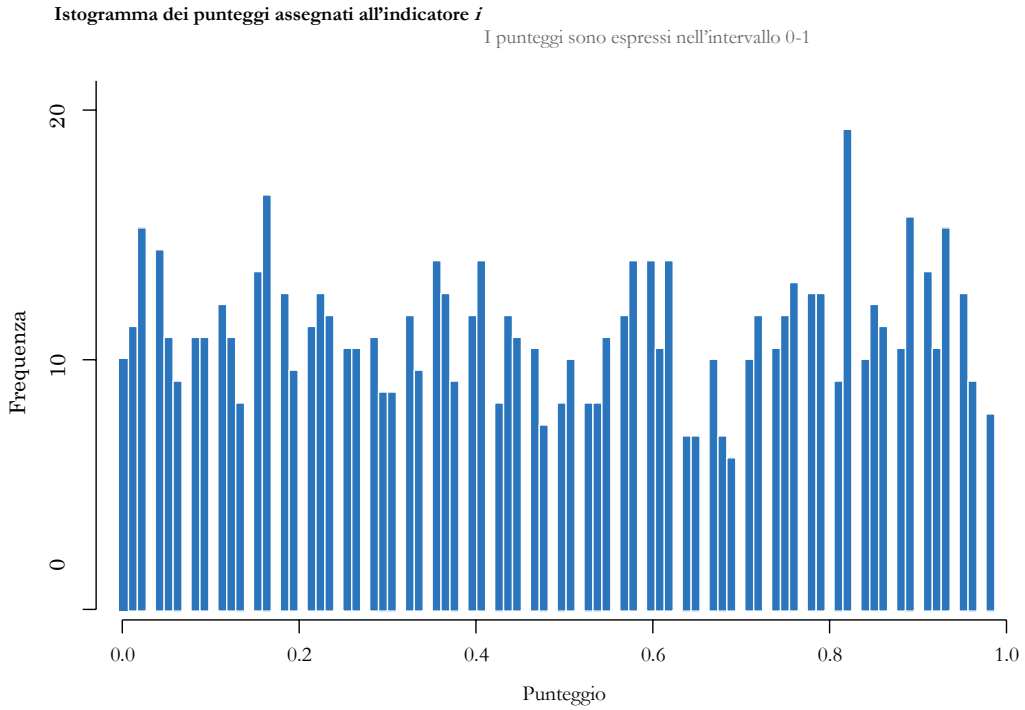

Figura 4.8: *Istogramma dei punteggi indicatore* I_{oepv2}

4.2.2 *Indicatore* I_{npna}

Attraverso il secondo indicatore, è stato possibile studiare la distribuzione delle tipologie di scelta del contraente delle stazioni appaltanti. Le diverse modalità di scelta del contraente, servono a regolamentare le fasi di aggiudicazione di un appalto, stabilendone le modalità e definendo alcuni vincoli. Comprendere quali di queste tipologie risultano più diffuse, diventa interessante per gli obbiettivi perseguiti da queste analisi. Il Codice dei contratti pubblici, prevede che una stazione appaltante debba individuare una delle tipologie di scelta del contraente per selezionare l'operatore economico al quale verrà affidato l'incarico sancito dal bando. Data la lista delle tipologie utilizzate, è possibile suddividerle in due macro-categorie distinte, denominate "affidamenti diretti" e "procedure competitive". La famiglia degli "affidamenti diretti", è composta da procedure in cui non vi è competizione per l'aggiudicazione di un determinato appalto. Viceversa, la famiglia delle "procedure competitive" prevede la fase competitiva già discussa nel capitolo 1. All'interno di queste due macro-categorie è possibile costruire un ulteriore frazionamento, così come illustrato nella figura 4.9.

L'indicatore I_{npna} intende affrontare l'analisi degli appalti dal punto di vista delle tipologie di scelta del contraente, valutandone la loro distribuzione, e in particolar modo il rapporto degli appalti aggiudicati attraverso una tipologia di scelta del contraente priva di competizione, sul totale. L'indicatore *ii* è espresso dalla formula

Suddivisione delle principali tipologie di scelta del contraente
Lo schema divide le tipologie di scelta del contraente in "procedure competitive" e "affidamenti diretti"

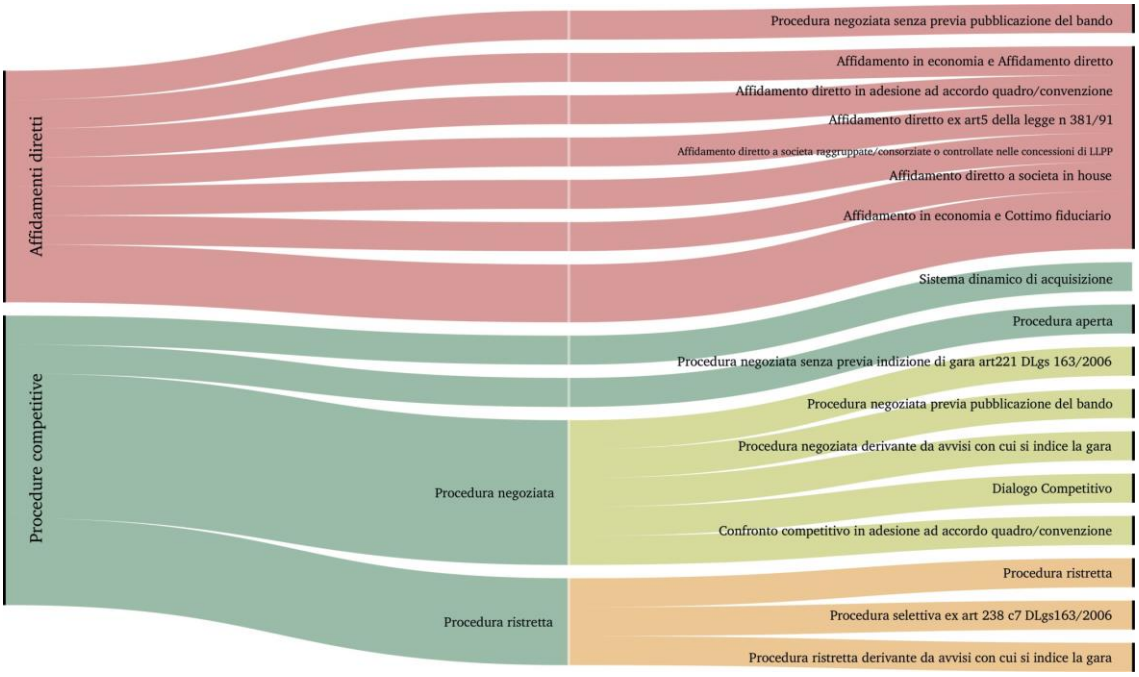

Figura 4.9: *Suddivisione delle tipologia di scelta del contraente*

$$I_{npna} = \frac{NTPNA_{i,t}}{NTP_{i,t}}$$

dove il termine $NTPNA_{i,t}$ è il numero delle procedure di appalto non aperte o ristrette utilizzate dalla *i*-esima stazione appaltante al tempo *t*, mentre $NTP_{i,t}$ è il numero totale delle procedure di appalto utilizzate dalla *i*-esima stazione appaltante al tempo *t*.

Nell'implementare il calcolo dell'indicatore con R, si è proceduto a costruire le due partizioni di figura 4.9, in modo tale da poter assegnare a ciascuna gara esaminata, un'etichetta corrispondente alla macro-categoria di appartenenza, a seconda della tipologia di scelta del contraente indicata. Si ritiene importante sottolineare che la suddivisione proposta in figura 4.9 è una proposta di suddivisione delle tipologie di scelta del contraente costruita assieme ad ANAC. Tuttavia potrebbe essere modificata, sulla base della "rigidità" dei vincoli che si intende imporre all'analisi. L'insieme delle procedure ristrette, ad esempio, nella figura 4.9 è stato collocato tra le procedure competitive. Le procedure ristrette sono procedure di scelta del contraente in cui solo gli operatori economici prescelti da una stazione appaltante possono partecipare al bando di gara. Gli appalti aggiudicati secondo questo tipo di tipologia dunque, da un lato hanno una componente competitiva, dall'altro sono chiusi rispetto

agli operatori economici che possono effettivamente partecipare alla gara. Per alcune tipologie di scelta del contraente dunque vi sono diverse possibili collocazioni, che dipendono dagli aspetti che si intende modellare e dai vincoli imposti sulla classificazione adottata.

L'indicatore I_{npna} proposto da ANAC, è stato rivisitato rispetto alla sua formulazione originale, poiché nell'implementarlo, si è osservato che i risultati finali risultavano fuorvianti in almeno due casi:

- nel primo caso, data una stazione appaltante, il semplice calcolo del rapporto tra le procedure d'appalto "non aperte" su quelle totali, restituiva 1 (valore massimo) in tutti quei casi in cui le stazioni appaltanti considerate avevano pubblicato esclusivamente gare "non aperte", tuttavia, dalle analisi statistiche condotte sui dati, è emerso che le stazioni appaltanti con un punteggio pari a 1, in media hanno contratto un numero di gare molto piccolo, contenuto nell'intervallo [1,3];

- nel secondo caso, si è osservato che l'indicatore restituisce lo stesso valore a prescindere dall'ordine di grandezza considerato. Si immagini a tal proposito due diverse stazioni appaltanti Sa_1 e Sa_2. Sa_1 ha pubblicato 15 gare, di cui 10 secondo una tipologia di scelta del contraente "non aperta". Sa_2 ha pubblicato 1500 gare, di cui 1000 secondo una tipologia di scelta del contraente "non aperta". In entrambi i casi, l'indicatore I_{npna} restituirebbe il valore 0,66, derivato dal rapporto 10/15 o dal rapporto analogo 1000/1500.

Alla luce dei due casi descritti, si ritiene che l'assegnamento di un punteggio pari a 1 (punteggio massimo), ad una stazione appaltante che ha pubblicato un numero di gare totali piccolo, risulti eccessivamente penalizzante per quella stazione appaltante. Si assume che la pubblicazione di pochissime gare (diciamo un numero compreso tra 1 e 3, derivato dai dati a disposizione), non possa essere sufficientemente significativo della tendenza di una stazione appaltante ad utilizzare le tipologie di scelta del contraente "non aperte". Viceversa, una stazione appaltante che totalizza una significativa percentuale di gare "non aperte" su un numero elevato di gare totali, evidenzia un comportamento recidivo da segnalare ed eventualmente da monitorare. Di conseguenza, si ritiene che la stazione appaltante Sa_1 dell'esempio proposto nel secondo caso, non possa avere un indicatore di corruzione pari alla stazione appaltante Sa_2, poiché in Sa_2, la tendenza a contrarre gare attraverso una tipologia di scelta del contraente "non aperta" risulta maggiore rispetto a Sa_1.

I ragionamenti appena descritti hanno condotto alla rivisitazione dell'indicatore I_{npna}, ritenendo utile pesarlo sul numero di gare "non aperte" aggiudicate da ciascuna stazione appaltante. La nuova formulazione dell'indicatore pertanto diventa la seguente:

$$I_{npna2} = \frac{NTPNA_{i,t}}{NTP_{it}} \times NTPNA_{i,t}$$

dove il termine $NTPNA_{i,t}$ è il numero delle procedure di appalto non aperte utilizzate dalla *i*-esima stazione appaltante al tempo *t*, mentre $NTP_{i,t}$ è il numero totale delle procedure di appalto utilizzate dalla *i*-esima stazione appaltante al tempo *t*.

Dopo aver implementato questo nuovo indicatore in R, si è proceduto al calcolo del suo valore per ciascuna stazione appaltante presente nei dati. Si è nuovamente partiti da un *dataset* contenente 125.011 appalti aggiudicati, da cui sono stati tolti tutti gli appalti in cui non è stata valorizzata la tipologia di scelta del contraente, ottenendo così un *dataset* contenente 93.831 appalti.

La "funzione punteggio" di rischio corruzione costruita per questo indicatore si occupa di assegnare un punteggio a ciascuna stazione appaltante, attraverso un'operazione di *rescaling* del valore dell'indicatore I_{npna2} in un intervallo continuo compreso tra [0, 1].

La media del punteggio così ottenuto è pari a 0.005, suggerendo una forte concentrazione dei dati attorno al valore 0. La deviazione standard della colonna dei punteggi è pari a 0,03, a conferma dell'affermazione precedente. Il grafico riportato in figura 4.10 illustra quanto espresso dagli indici statistici appena menzionati.

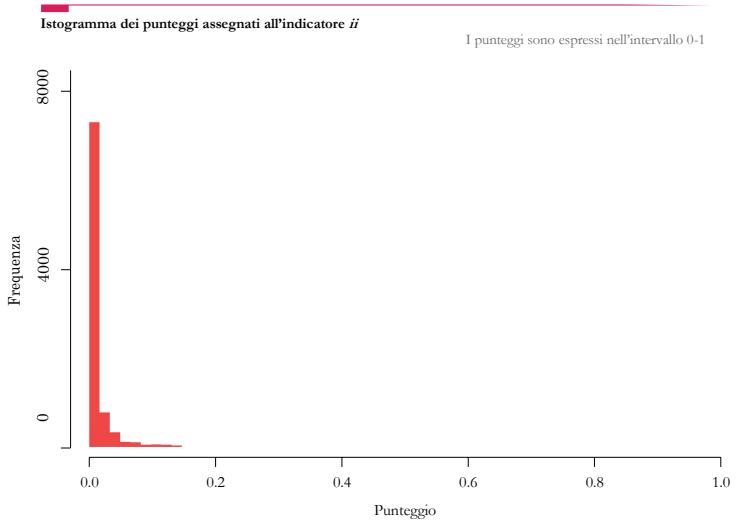

Figura 4.10: *Istogramma dei punteggi indicatore* I_{npna2}

Per osservare la concentrazione dei dati, occorre effettuare uno "*zoom*" sulla parte sinistra dell'asse *x*, come riportato in figura 4.11.

Anche in questo caso è stata costruita una tabella con le seguenti colonne:

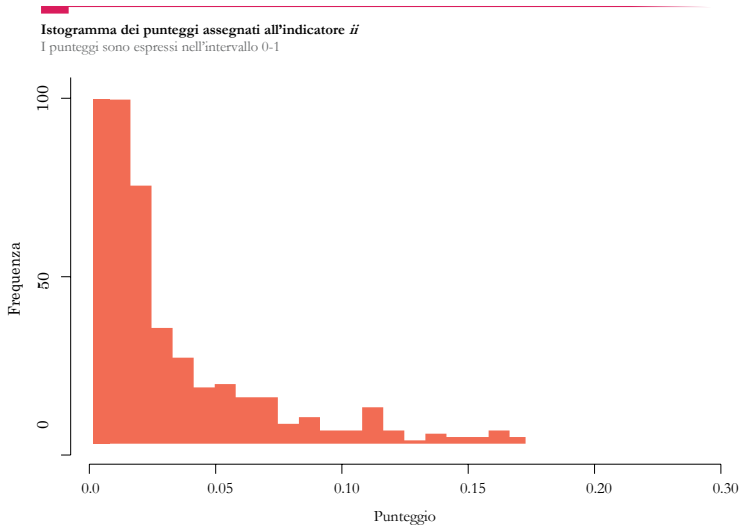

Figura 4.11: *Istogramma dei punteggi indicatore* I_{npna2}

- **CF stazione appaltante**: contiene il codice fiscale della stazione appaltante;
- **N. gare totale**: riporta il numero di gare totali della stazione appaltante;
- **Importo totale**: riporta il totale aggiudicato dalla stazione appaltante;
- **N. gare chiuse**: riporta il numero di gare che possiedono una tipologia di scelta del contraente etichettata come "non aperta";
- e**gare chiuse**: riporta il totale aggiudicato delle gare "non aperte";
- **N. gare aperte**: riporta il numero di gare che possiedono una tipologia di scelta del contraente etichettata come "competitiva" (aperta);
- e**gare aperte**: riporta il totale aggiudicato delle gare "aperte";
- **Indice**: riporta il valore dell'indicatore I_{npna2}, della stazione appaltante;
- **Punteggio**: riporta il valore del punteggio calcolato sull'indicatore I_{npna2} per ciascuna stazione appaltante.

4.2.3 *Indicatore* I_{vpna}

Il terzo indicatore studiato, viene espresso attraverso la seguente formula:

$$I_{vpna} = \frac{VTPNA_{i,t}}{VTP_{i,t}}$$

dove il termine $VTPNA_{i,t}$ è il valore totale delle procedure non aperte attivate dalla *i*-esima stazione appaltante al tempo *t*, mentre $VTP_{i,t}$ è il valore totale delle procedure attivate dalla *i*-esima stazione appaltante al tempo *t*. Questo indicatore è strettamente collegato con il precedente, a differenza che in questo caso l'attenzione è focalizzata sugli importi delle procedure aperte e non sul loro numero. In questo caso, l'indicatore esprime un rapporto che non ha bisogno di alcuna modifica o integrazione. Il valore economico delle gare pubblicate secondo una tipologia di scelta del contraente "non aperta", è un dato da studiare assieme al loro numero, per avere una panoramica completa di ciascuna stazione appaltante. Anche in questo terzo caso, è stata implementata una "funzione punteggio", che si occupa di normalizzare i valori dell'indicatore *iii* all'interno dell'intervallo [0, 1].

La media degli importi delle gare aggiudicate con un criterio di scelta del contraente "non aperto" ammonta a 3.625.794e, mentre l'importo di aggiudicazione totale massimo ammonta a 5.562.011.491e ed è riferito all'Azienda Ospedaliera di Perugia. Di seguito sono riportati due grafici relativi agli appalti "non aperti". Il primo mette in correlazione l'importo totale aggiudicato con il numero totale di questi appalti (4.12). Il grafico 4.13 invece, allargando la zona dove i punti sono maggiormente concentrati, mette in risalto la fitta concentrazione di punti in un limitato intervallo di importi, ad eccezione di alcuni pochi punti (*outliers*) che si discostano dal *cluster* centrale.

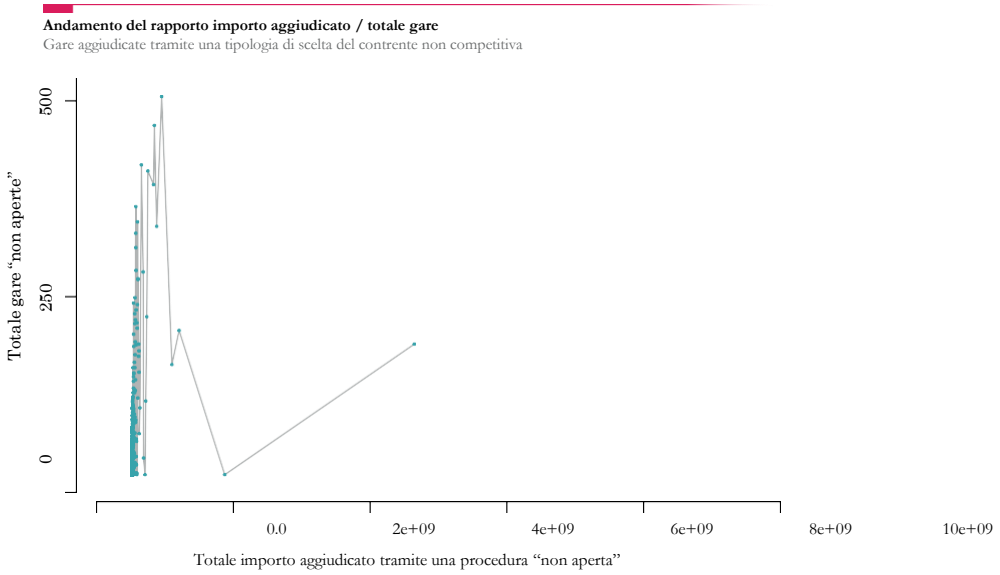

Figura 4.12: *Importi aggiudicati e numero appalti "non aperti"*

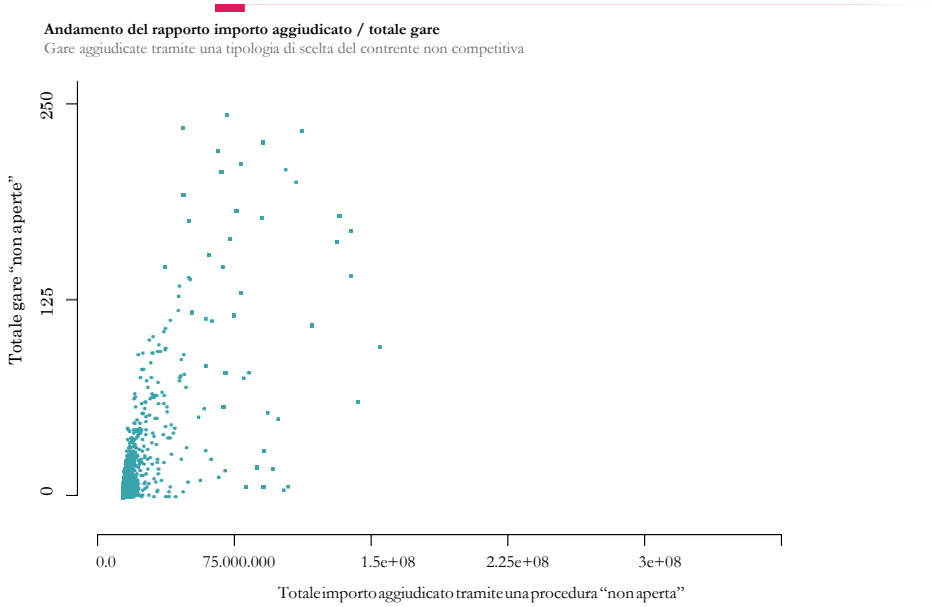

Figura 4.13: *Dettaglio: importi aggiudicati e numero appalti "non aperti"*

Infine è stato costruito l'istogramma di figura 4.14, in cui è riportata la frequenza dei punteggi assegnati a ciascuna stazione appaltante considerata. Nel grafico sono stati omessi il valore 0 e il valore 1, per poter visualizzare meglio l'intero intervallo.

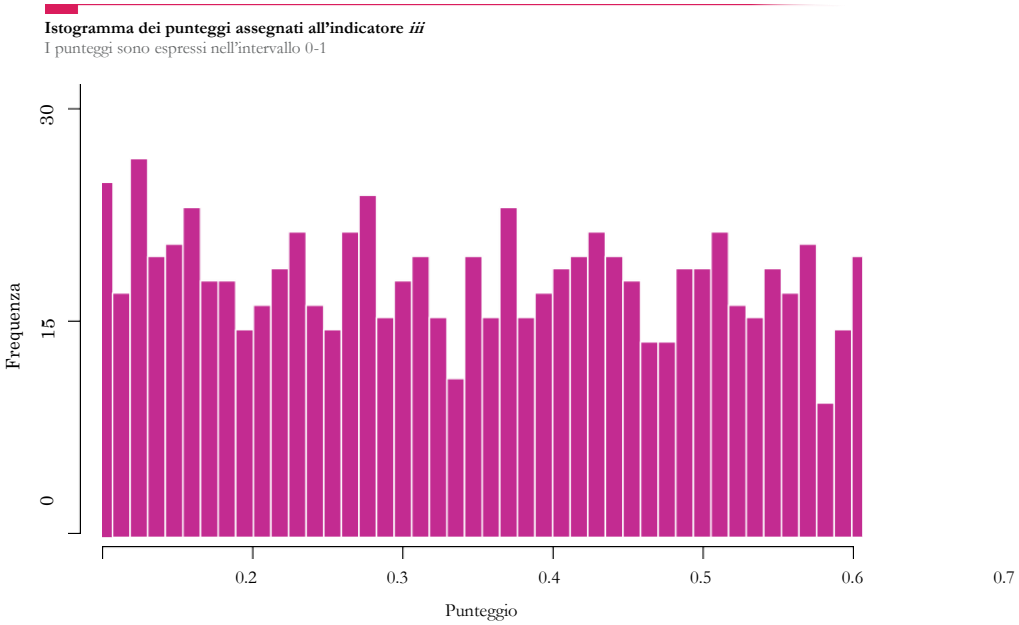

Figura 4.14: *Istogramma dei punteggi indicatore I_{vpna}*

L'analisi dell'indicatore è stata completata con la costruzione della tabella di sintesi di tutti i calcoli effettuati, di cui si riporta l'intestazione.

- **CF stazione appaltante**: contiene il codice fiscale della stazione appaltante;

- **N. gare totale**: riporta il numero di gare totali della stazione appaltante;
- **Importo totale**: riporta il totale aggiudicato dalla stazione appaltante;
- **N. gare chiuse**: riporta il numero di gare che possiedono una tipologia di scelta del contraente etichettata come "non aperta";
- e**gare chiuse**: riporta il totale aggiudicato delle gare "non aperte";
- **N. gare aperte**: riporta il numero di gare che possiedono una tipologia di scelta del contraente etichettata come "competitiva" (aperta);
- e**gare aperte**: riporta il totale aggiudicato delle gare "aperte";
- **Indice**: riporta il valore dell'indicatore I_{vpna}, della stazione appaltante;
- **Punteggio**: riporta il valore del punteggio calcolato sull'indicatore I_{vpna} per ciascuna stazione appaltante.

4.2.4 *Indicatore* I_{uo}

Il quarto indicatore, conta i bandi per il quale è stata presentata una sola offerta e che di conseguenza hanno coinvolto un solo partecipante. L'attributo relativo a questo aspetto dell'appalto è denominato "*numero imprese offerenti*". L'indicatore, espresso dalla formula che segue, si limita ad un semplice conteggio.

$$I_{uo} = \frac{N\,TA1_{i,t}}{N\,TA_{i,t}}$$

$N\,TA1_{i,t}$ è il numero delle procedure aggiudicate della *i*-esima stazione appaltante al tempo *t* con un numero dei partecipanti uguale ad uno. $N\,TA_{i,t}$ è il numero totale delle procedure di appalto aggiudicate dalla *i*-esima stazione appaltante al tempo *t*.

Per ottenere dei risultati significativi, sono state escluse le righe dove l'attributo "*numero imprese offerenti*" non era valorizzato. La "funzione punteggio" di rischio corruzione, mappa i valori calcolati dall'indicatore nell'intervallo [0, 1]. Non si è ritenuto necessario effettuare alcuna pesatura dell'indicatore. Di 125.011 appalti aggiudicati, 118.390 non hanno un valore nullo sull'attributo "numero imprese offerenti". Di questi, 33.668 possiedono una sola impresa offerente, pari al 28,44% del totale. Inoltre, su un importo aggiudicato totale pari a 99.025.959.372e, l'importo aggiudicato totale delle gare con un solo partecipante ammonta al 28,6% (28.318.825.305e). Nel calcolare questo quarto indicatore, è stato misurato anche il valore minimo, medio e massimo del ribasso delle gare con un solo partecipante. Il valore minimo del ribasso presente nei dati è pari a 0. Il valore medio è pari a 21,31e, mentre il valore massimo è pari a 1.366,19e.

Complessivamente, la presenza di gare con un solo partecipante, sul totale delle gare considerate, sembra riguardare all'incirca 1/3 dei dati considerati (sia per numero gare, sia per importo).

La tabella riassuntiva dei calcoli effettuati è composta dalle seguenti colonne:

- **CF stazione appaltante**: contiene il codice fiscale della stazione appaltante;
- **N. gare totale**: riporta il numero di gare totali della stazione appaltante;
- **Importo totale**: riporta il totale aggiudicato dalla stazione appaltante;
- **N. gare un partecipante**: riporta il numero di gare aggiudicate ad una sola impresa offerente;
- e**gare un partecipante**: riporta il totale aggiudicato delle gare con una sola impresa offerente;
- e**ribasso gare con un partecipante**: riporta il totale del ribasso delle gare con una sola impresa offerente;
- **Indice**: riporta il valore dell'indicatore I_{uo}, della stazione appaltante;
- **Punteggio**: riporta il valore del punteggio calcolato sull'indicatore I_{uo} per ciascuna stazione appaltante.

4.2.5 *Indicatore* I_{tmpo}

L'indicatore I_{tmpo}, definito dalla formula

$$I_{tmpo} = \frac{\sum_{k=1}^{NTA_{i,t}} (DSPO_{i,k} - DPB_{i,k})}{NTA_{i,t}}$$

misura il tempo compreso tra la *data di pubblicazione del bando* e la sua *data di scadenza*, dove $DSPO_{i,k}$ è la data di scadenza di presentazione delle offerte per l'amministrazione i-esima e per l'affidamento k-esimo, $DPB_{i,k}$ è la data di pubblicazione del bando per l'amministrazione i-esima e per l'affidamento k-esimo. Infine, $NTA_{i,t}$ è il numero totale delle procedure di appalto aggiudicate dalla i-esima stazione appaltante al tempo t.

Attraverso gli indicatori precedenti, è stato possibile misurare quantità e importo degli appalti, in riferimento a specifici vincoli (la modalità di aggiudicazione, la tipologia di scelta del contraente, il numero di partecipanti). Con questo quinto indicatore invece, il *focus* si sposta sui tempi degli appalti. Le variabili riguardanti il tempo, presenti nei *dataset*, oltre alla *data di pubblicazione* del bando e la *data di scadenza*, riguardano la *data di aggiudicazione definitiva*, la *data di stipula del contratto* , la *data di inizio definitiva*,

la *data del termine contrattuale* e la *data di effettiva ultimazione*. Purtroppo, tutti questi campi risultano in larga parte vuoti, rientrando nelle colonne arancioni della figura 4.2 d'inizio capitolo.

La tabella riassuntiva dei calcoli effettuati è composta dalle seguenti colonne:

- **CF stazione appaltante**: contiene il codice fiscale della stazione appaltante;
- **N. gare totale**: riporta il numero di gare totali della stazione appaltante;
- **Giorni totali di differenza**: riporta la somma dei giorni di differenza tra la data di scadenza di un'offerta e la data di pubblicazione di ciascun appalto, pubblicato dalla stazione appaltante considerata;
- **Indice**: riporta il valore dell'indicatore I_{tmpo}, della stazione appaltante;
- **Punteggio**: riporta il valore del punteggio calcolato sull'indicatore I_{tmpo} per ciascuna stazione appaltante.

4.2.6 *Indicatore* $I_{tmpo}2$

L'indicatore $I_{tmpo}?$, non rientra tra gli indicatori riportati nel capitolo 2.2.3. Esso nasce dall'esigenza di intercettare alcuni aspetti che il precedente indicatore non considera. L'indicatore si concentra nuovamente sul calcolo di un intervallo temporale, ma in questo caso viene misurata la distanza tra la *data di aggiudicazione* di un bando e la sua *data di pubblicazione*, pesando l'indicatore sul valore degli importi aggiudicati in proporzione al numero di giorni calcolati. L'indicatore può essere espresso come segue:

$$I_{tmpo^{\star}} = \frac{\left(\sum_{k=1}^{NTA_{it}} (DAO_{ik} - DPB_{ik})\right) \times VTP_{it}}{NTA_{i,t} + \left(\sum_{k=1}^{NTA_{it}} (DAO_{ik} - DPB_{ik})\right)}$$

dove $DAO_{i,k}$ è la data di aggiudicazione definitiva delle offerte per l'amministrazione i- esima e per l'affidamento k-esimo, $DPB_{i,k}$ è la data di pubblicazione del bando per l'amministrazione i-esima e per l'affidamento k-esimo. $VTP_{i,t}$ rappresenta invece il valore totale aggiudicato delle procedure pubblicate dalla i-esima stazione appaltante, mentre $NTA_{i,t}$ è il numero totale delle procedure di appalto aggiudicate dalla i-esima stazione appaltante al tempo t.

L'idea alla base di questo indicatore, è quella di analizzare i tempi che intercorrono tra la pubblicazione di un bando e la data della sua effettiva aggiudicazione, pesando questo rapporto sulla base di alcune assunzioni argomentate di seguito.

Si immagini la situazione descritta dall'immagine 4.15 (*Situazione 1*), in cui due diverse stazioni appaltanti Sa_1 e Sa_2 totalizzano lo stesso numero di giorni[7] e lo stesso numero di gare totali. L'indicatore *v* descritto in precedenza, restituirebbe per entrambe le stazioni appaltanti lo stesso valore, pari a 0,23 (derivato dal rapporto 10/43). Questo valore non è proporzionato sull'importo totale che le due stazioni appaltanti hanno aggiudicato nel corso dell'anno. Nel caso raffigurato in 4.15, la stazione appaltante Sa_1 ha aggiudicato appalti per un totale di 1.250e, mentre la stazione appaltante Sa_2, attraverso lo stesso numero di gare, ha aggiudicato appalti per un totale di 17.250e. Da qui nasce l'idea di pesare il calcolo del tempo tra la data di aggiudicazione del bando e la sua data di pubblicazione attraverso l'importo aggiudicato.

Inoltre, si è voluto modellare un'ulteriore situazione potenzialmente fuorviante, espressa dalla figura 4.16. A parità di denaro aggiudicato, diventa importante il "fattore tempo". Nell'immagine 4.16, la stazione appaltante Sa_1 ha aggiudicato 15.000e in un solo giorno, mentre la stazione appaltante Sa_2 ha aggiudicato la medesima cifra in un tempo superiore, pari a 32 giorni. Si assume che l'aggiudicazione di una certa cifra in un tempo molto stretto, sia in generale una situazione da monitorare con più attenzione rispetto all'aggiudicazione della stessa cifra in un tempo nettamente superiore. Per questo, il numero di giorni diventa un ulteriore peso per l'indicatore in esame. L'aggiunta del numero di giorni al denominatore dell'indicatore $I_{impo?}$, controlla in maniera proporzionale la crescita del risultato finale.

Nome staz. appaltante	Numero totale di giorni	Numero totale gare	Importo totale aggiudicato
Stazione appaltante 1	10	43	1.250€
Stazione appaltante 2	10	43	17.250€

Figura 4.15: *Situazione 1*

Nome staz. appaltante	Numero totale di giorni	Numero totale gare	Importo totale aggiudicato
Stazione appaltante 1	1	43	15.000€
Stazione appaltante 2	32	43	15.000€

Figura 4.16: *Situazione 2*

La tabella riassuntiva dei calcoli effettuati è composta dalle seguenti colonne:

- **CF stazione appaltante**: contiene il codice fiscale della stazione appaltante;

[7] calcolato come somma delle differenze tra la data di aggiudicazione del bando e la sua data di pubblicazione per ciascun appalto pubblicato

- **N. gare totale**: riporta il numero di gare totali della stazione appaltante;
- **Giorni totali di differenza**: riporta la somma dei giorni di differenza tra la data di aggiudicazione definitiva di un'offerta e la data di pubblicazione di ciascun appalto, pubblicato dalla stazione appaltante considerata;
- **Importo totale**: riporta il valore totale degli importi aggiudicati dalla stazione appaltante considerata;
- **Indice**: riporta il valore dell'indicatore $I_{tmpo?}$, della stazione appaltante;
- **Punteggio**: riporta il valore del punteggio calcolato sull'indicatore $I_{tmpo?}$ per ciascuna stazione appaltante.

Tutto il codice prodotto per costruire le tabelle descritte in questo paragrafo è stato raccolto in un apposito *repository* GitHub. Alcune parti del codice sono riportate anche tra i documenti in allegato. All'interno del *repository*, è possibile visualizzare il codice sviluppato per la costruzione degli indicatori di corruzione, e quello relativo alla pulizia dei dati e alla loro analisi preliminare. Il *repository* è disponibile all'indirizzo: https://github.com/mamatteo/analisi-corruzione-appalti-pubblici.

Classifiche e confronti finali - L'analisi dei sei indicatori descritti nei capitoli precedenti ha permesso la costruzione di altrettante tabelle. Ciascuna tabella raccoglie i valori degli attributi importanti di ciascun indicatore, fornendo un'utile raccolta di informazioni. In questo senso, ciascuna delle sei tabelle costruite può diventare un *dataset* su cui effettuare ulteriori analisi. Spesso è stato utile studiare la distribuzione degli indici (o dei punteggi), al fine comprendere meglio la distribuzione del fenomeno misurato. Inoltre, grazie alle informazioni raccolte, è stato possibile calibrare meglio il calcolo dei vari indici. Ciascuna tabella è stata corredata dalla colonna "Punteggio_i_k", che riporta per ciascuna stazione appaltante presente nei *dataset*, il valore del punteggio relativo all'indicatore i_k. Una volta costruite tutte le "colonne punteggio", queste sono state combinate assieme in un' unica tabella finale, formata nella prima colonna dall'elenco delle stazioni appaltanti considerate nelle analisi (elencate tramite il loro codice fiscale), mentre nelle sei colonne successive sono state riportate le sei "colonne punteggio" secondo l'ordine con cui sono stati costruiti e calcolati i sei indicatori. L'ultima colonna di ciascuna tabella è formata dalla somma per riga di ciascun punteggio, restituendo così l'indicatore aggregato finale. Il modello costruito è stato ottenuto in maniera incrementale, svolgendo iterativamente i seguenti passaggi:

1. Costruzione della partizione di dati utile al calcolo dell'*Indicatore* i_k;

2. Calcolo dell'*Indicatore* i_k eventualmente ridefinito;
3. Costruzione della *funzione punteggio* e assegnamento dei punteggi.

I tre passaggi sopra descritti sono stati ripetuti per tutti i k indicatori discussi in questa tesi.

Nella formula che segue è riportata la formalizzazione del modello realizzato, a cui è stato dato il nome di *Corruption Indicator Score* (*CIS*).

$$Corruption\ Indicator\ Score = \sum_{i=1}^{k} \varphi_i$$

$$\text{dove } \varphi_i = punteggio_{ind_k}, \text{ con } k = 6.$$

Il *Corruption Indicator Score* presentato rappresenta un indicatore sintetico anziché una combinazione lineare. Senza dubbio l'utilizzo di una combinazione lineare avrebbe avuto un carattere più generale e proprio per questo sarebbe stata preferibile. Il punto è che non si dispone di indicazioni circa il criterio con cui scegliere i pesi con cui calibrare e valutare la combinazione. Un'ipotesi sarebbe potuta essere quella di pesare in base alla varianza, ma la scelta sarebbe stata del tutto arbitraria.

La tabella ottenuta dalle trasformazioni elencate nella pagina precedente è stata poi sottoposta ad alcune semplici operazioni utili a visualizzare meglio i dati. Innazitutto è stata ordinata in ordine decrescente la colonna *Corruption Indicator Score*, in modo da poter visualizzare facilmente le stazioni appaltanti con l'indice *Cis* maggiore. Inoltre, sono stati colorati i valori riportati in tabella in tre modalità distinte: è stato assegnato il colore rosso a tutti quei punteggi il cui valore supera il valore di 0.5. È stato dato il colore verde a tutti quei punteggi il cui valore è inferiore a 0.5. I punteggi con valore uguale a 0.5 sono stati colorati di blu. In questa maniera, scorrendo la tabella, diventa più facile fare delle analisi "a colpo d'occhio", sia per riga sia per colonna, sull'andamento positivo (in verde) o negativo (in rosso), delle varie stazioni appaltanti in riferimento ai sei indicatori considerati. Scorrendo la tabella, formata da oltre 7000 righe, è interessante notare come variano i colori per colonna, studiando come il cambiamento delle zone di colore inizialmente uniforme verso zone di colorazioni differenti e viceversa. Si noti a tal proposito come in figura 4.18, la colonna *p1* completamente rossa in figura 4.17 diventi quasi completamente verde all'abbassarsi del *Corruption Indicator Score*, così come accade, ad esempio, per la colonna *p4*. L'analisi delle correlazioni tra i diversi indicatori con il punteggio aggregato espresso dal *Corruption Indicator Score*, ha mostrato come ci sia una correlazione positiva (pari al 50%), tra la colonna dei punteggi *p1* e il *Corruption Indicator Score*, e tra la colonna dei punteggi *p4* e il *Corruption Indicator Score*. I risultati completi sulle correlazioni, verranno riportati nel paragrafo 4.4.

Quel che è emerge dai calcoli, infine, è che il *Corruption Indicator Score* più alto raggiunto dalle stazioni appaltanti considerate è pari al valore 3.60, in una scala che va da 0 (*punteggio minimo*) a 6 (*punteggio massimo*).

Tabella dei risultati finali - anno 2015

Prime 12 righe

CFStazapp	p1	p2	p3	p4	p5	p6	Corruption_Indicator_Score
02101050546	0.99	0.13	0.99	0.59	0.03	0.87	3.60
02675880583	1.00	0.00	1.00	1.00	0.24	0.00	3.24
01571730470	1.00	0.00	1.00	1.00	0.10	0.00	3.10
91002520293	1.00	0.00	1.00	1.00	0.10	0.00	3.10
97420690584	1.00	0.01	0.80	0.26	0.02	1.00	3.09
97513110151	1.00	0.00	1.00	1.00	0.09	0.00	3.09
03462000161	1.00	0.00	1.00	1.00	0.06	0.00	3.06
86003470019	1.00	0.00	1.00	1.00	0.06	0.00	3.06
00221280688	1.00	0.00	1.00	1.00	0.05	0.00	3.05
01009620152	1.00	0.00	1.00	1.00	0.05	0.00	3.05
02747480123	1.00	0.00	1.00	1.00	0.05	0.00	3.05
00111210779	1.00	0.00	1.00	1.00	0.04	0.00	3.04

Figura 4.17: *Tabella finale anno 2015*

Tabella dei risultati finali - anno 2015

Righe casuali della tabella

CFStazapp					p1	p2	p3	Corruption_Indicator_Score	p6
93430140728	0.83	0.00	0.41	0.00	0.06	0.00		1.30	
00063640684	0.00	0.01	1.00	0.25	0.03	0.00		1.29	
00123220428	0.14	0.00	0.97	0.17	0.01	0.00		1.29	
00128850229	0.00	0.00	1.00	0.25	0.04	0.00		1.29	
00234140234	0.27	0.00	1.00	0.00	0.02	0.00		1.29	
00242500395	0.00	0.00	1.00	0.27	0.02	0.00		1.29	
00282190891	0.00	0.00	0.76	0.50	0.03	0.00		1.29	
00307600635	0.25	0.01	1.00	0.00	0.03	0.00		1.29	
00316390228	0.00	0.00	1.00	0.25	0.04	0.00		1.29	
00324100163	0.00	0.00	1.00	0.25	0.04	0.00		1.29	
00339040388	0.00	0.01	1.00	0.25	0.03	0.00		1.29	
00405030586	0.58	0.02	0.43	0.24	0.02	0.00		1.29	
00441100351	0.34	0.00	0.67	0.25	0.03	0.00		1.29	
00483290045	0.00	0.00	1.00	0.00	0.29	0.00		1.29	

Figura 4.18: *Tabella finale anno 2015 - estratto*

Di seguito si riporta invece la classifica delle prime dodici stazioni appaltanti riferite all'anno 2016. Le differenze dei risultati rispetto alla tabella riferita all'anno 2015 cambiano significativamente lungo le colonne dei punteggi, mentre rimangono abbastanza simili lungo la colonna del *Corruption Indicator Score*. Dal confronto tra le due tabelle, emerge che la stazione appaltante con codice fiscale "03462000161" appare in entrambi i risultati di sintesi.

Tabella dei risultati finali - anno 2016
Prime 12 righe

CFStazapp	p1	p2	p3	p4	p5	p6	Corruption_Indicator_Score
01602780387	1.00	0.00	1.00	1.00	0.13	0.00	3.13
84000360036	1.00	0.00	1.00	1.00	0.12	0.00	3.12
00117340539	1.00	0.00	1.00	1.00	0.11	0.00	3.11
82000910511	1.00	0.00	1.00	1.00	0.11	0.00	3.11
00201490166	1.00	0.00	1.00	1.00	0.09	0.00	3.09
00231300526	1.00	0.00	1.00	1.00	0.08	0.00	3.08
03462000161	1.00	0.00	1.00	1.00	0.08	0.00	3.08
00139940407	1.00	0.00	1.00	1.00	0.07	0.00	3.07
05728560961	1.00	0.00	1.00	1.00	0.07	0.00	3.07
80007520499	1.00	0.00	1.00	1.00	0.05	0.00	3.05

Figura 4.19: *Tabella finale anno 2016*

Infine, si riporta un estratto della tabella relativa ai risultati calcolati sull'ultimo anno disponibile, il 2017.

Tabella dei risultati finali - anno 2017
Prime 12 righe

CFStazapp	p1	p2	p3	p4	p5	p6	Corruption_Indicator_Score
93106990422	1.00	0.00	1.00	1.00	0.23	0.00	3.23
00288640162	1.00	0.00	1.00	1.00	0.11	0.00	3.11
80015560214	1.00	0.00	1.00	1.00	0.10	0.00	3.10
00201490166	1.00	0.00	1.00	1.00	0.09	0.00	3.09
01993250164	1.00	0.00	1.00	1.00	0.09	0.00	3.09
80003070051	1.00	0.00	1.00	1.00	0.09	0.00	3.09
00374850220	1.00	0.00	1.00	1.00	0.07	0.00	3.07
00571510130	1.00	0.00	1.00	1.00	0.07	0.00	3.07
02239670215	1.00	0.00	1.00	1.00	0.07	0.00	3.07
83003010275	1.00	0.00	1.00	1.00	0.07	0.00	3.07
00300450129	1.00	0.00	1.00	1.00	0.06	0.00	3.06
01865850018	1.00	0.00	1.00	1.00	0.06	0.00	3.06

Figura 4.20: *Tabella finale anno 2017*

La distribuzione del *Corruption Indicator Score* nei tre diversi anni considerati è rappresentata dai grafici 4.21, 4.22 e 4.23.

Distribuzione del *Corruption Indicator Score* **- anno 2015**
Il *Corruption Indicator Score* esprime un punteggio da 0 a 6

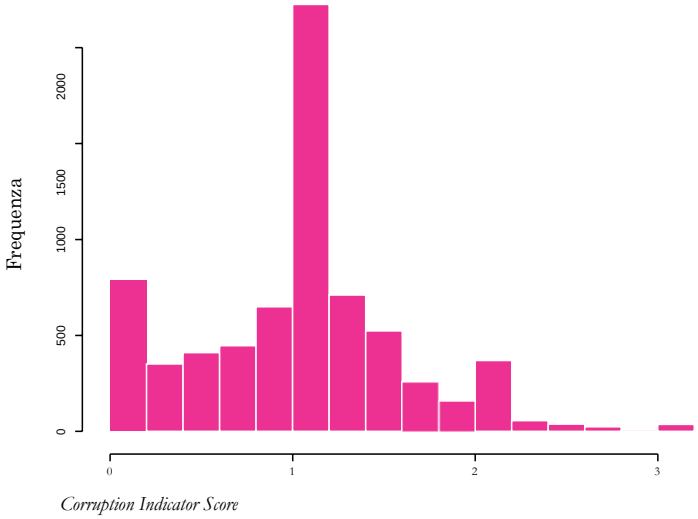

Figura 4.21: *Distribuzione Corruption Indicator Score anno 2015*

Distribuzione del *Corruption Indicator Score* **- anno 2016**
Il *Corruption Indicator Score* esprime un punteggio da 0 a 6

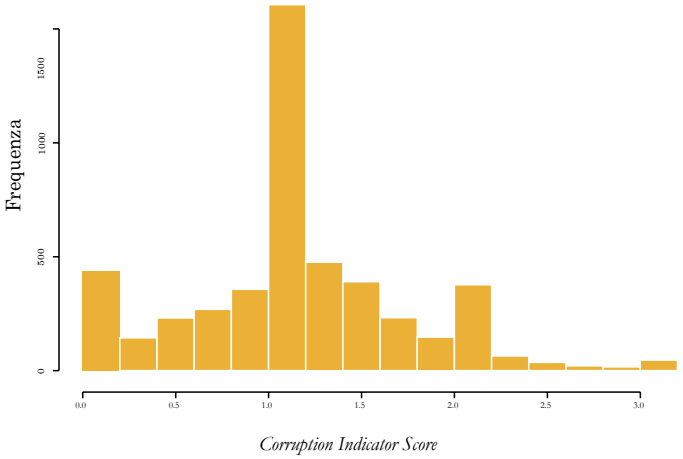

Figura 4.22: *Distribuzione Corruption Indicator Score anno 2016*

Distribuzione del *Corruption Indicator Score* - **anno 2017**
Il *Corruption Indicator Score* esprime un punteggio da 0 a 6

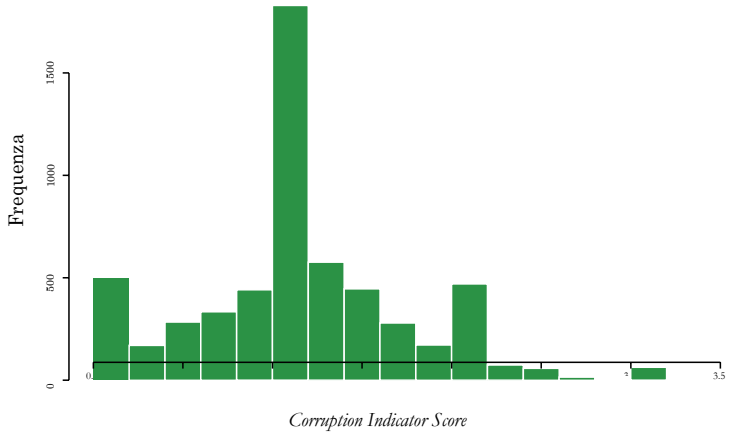

Figura 4.23: *Distribuzione Corruption Indicator Score anno 2017*

Gli istogrammi riportati nella pagina precedente, mostrano come in tutti gli anni considerati, vi è una predominanza dei punteggi di poco superiori al valore 1. Nel valutare i risultati ottenuti, si è osservato che una generica stazione appaltante potrebbe ottenere un punteggio di rischio di corruzione molto alto (diciamo prossimo a 1) in un solo indicatore, e, contemporaneamente, ottenere il valore 0 in tutti i cinque indicatori rimanenti. In questo frangente, il *Corruption Indicator Score* restituirebbe il valore 1, ottenuto dalla somma dei punteggi dei sei indicatori della stazione appaltante considerata. In generale, si è assunto che una stazione appaltante debba essere marcata come "ad alto rischio di possibile corruzione" anche se totalizza un punteggio alto in un unico indicatore. In altre parole, il *Cis* fin qui utilizzato non fa una somma pesata dei sei indicatori. Si immagini ad esempio il caso in cui una stazione appaltante "generi" corruzione attraverso una sfilza di affidamenti diretti. In questo caso, l'indicatore che misura il rapporto tra affidamenti diretti e gare totali, restituirebbe un valore molto alto. Gli altri cinque indicatori potrebbero invece rimanere prossimi allo zero, ma non per questo si è ritenuto utile attenuare il punteggio finale.

Tuttavia, per intercettare alcuni casi anomali e per provare a pesare il *Cis* finora utilizzato nelle analisi, si è pensato di affiancare a quello a cui d'ora in avanti ci riferiremo come *Corruption Indicator Score* - tradizionale, un ulteriore punteggio, che non fa altro che normalizzare il *Corruption Indicator Score* - tradizionale, in maniera tale da intercettare gli *outliers* e fornire delle misure finali pesate.

Nella figura 4.24 sono riportate due situazioni diverse. Nella metà alta della figura è riportato l'andamento di una stazione appaltante che ha totalizzato il valore 1 in un solo indicatore. Nella metà bassa della figura invece, è riportata una stazione appaltante, che ha totalizzato dei valori superiori allo

Esempio di *trend* sui sei indicatori
Dai punteggi originali ai nuovi punteggi

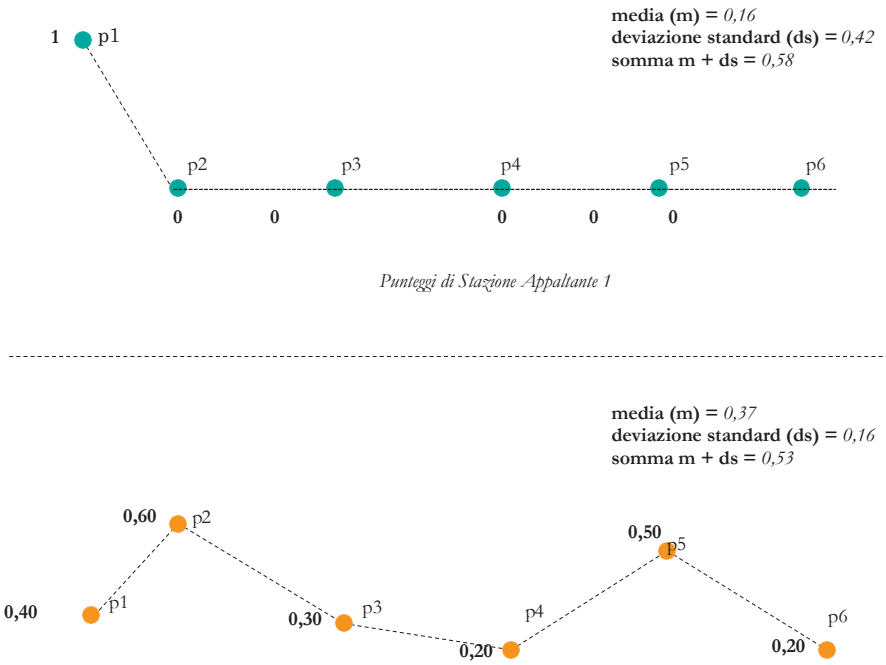

Figura 4.24: *Trend degli indicatori*

zero in tutti gli indicatori. Nell'elaborare il *Corruption Indicator Score* per queste due stazioni appalanti, si osserva che nel primo caso si otterrebbe un valore pari a 1, mentre nel secondo caso il *Corruption Indicator Score* ammonterebbe al valore 2,20, classificando la *stazione appaltante 2* in una posizione superiore rispetto alla *stazione appaltante 1*. In altre parole, *la stazione appaltante 2* totalizzerebbe un punteggio di rischio di corruzione maggiore rispetto alla *stazione appaltante 1*.

Come anticipato nei precedenti capoversi, si è pensato di fornire un nuovo indicatore, strettamente imparentato con il *Corruption Indicator Score*, a cui è stato dato il nome di *Corruption Indicator Score Outliers (Cis outliers)*. Questo nuovo indicatore assegnerà il valore 0 ai valori prossimi alla somma della media più la deviazione standard del *Corruption Indicator Score*, e il valore 1 ai valori massimi. In figura 4.25 è stata messa in evidenza la soglia che individua i valori a cui verrà assegnato un punteggio tra 0 e 1 (tutti e soli i valori sopra la soglia), e i valori a cui verrà assegnato un punteggio pari a 0 (tutti i valori sotto la soglia).

Il nuovo indicatore *Corruption Indicator Score Outliers (Cis outliers)* è stato implementato in R e i risultati

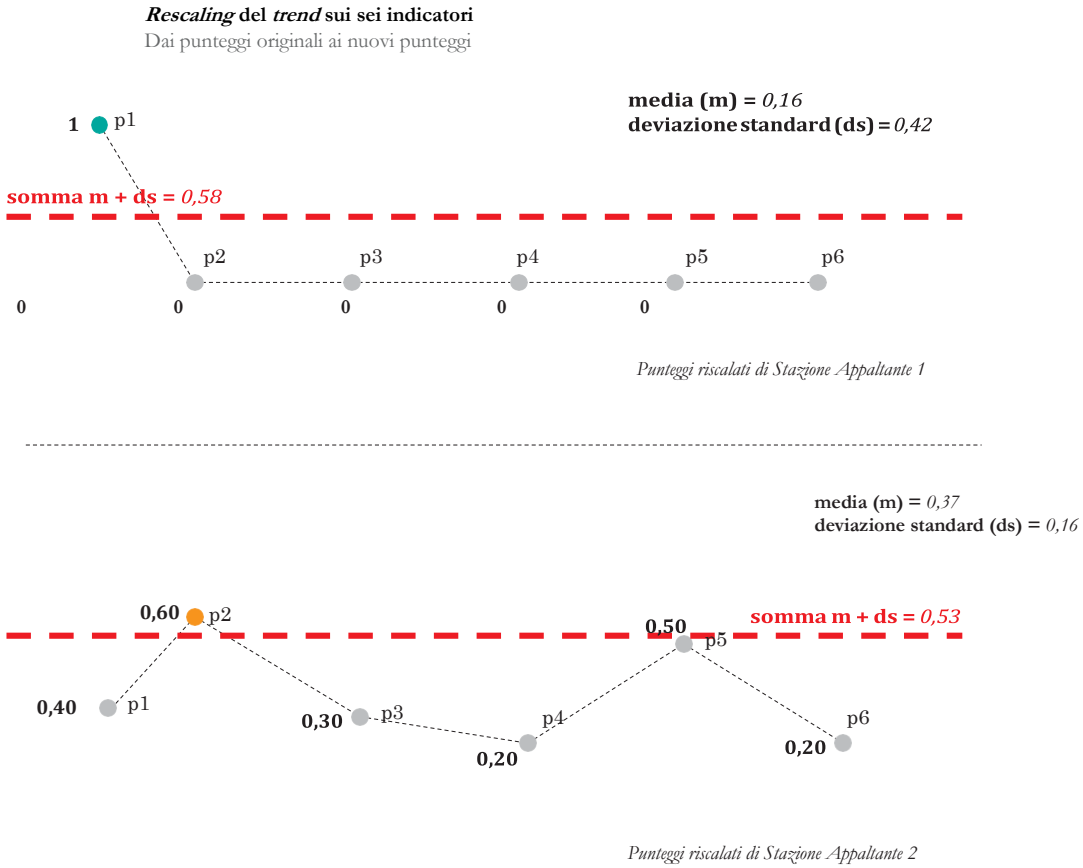

Figura 4.25: *Trend degli indicatori*

ad esso collegati sono stati aggregati nella tabella di sintesi, come è possibile osservare nella figura 4.26.

Come ultima operazione sui dati, è stata fatta la somma di tutti i *Corruption Indicator Score* ottenuti nelle tabelle di ciascun anno considerato, nell'idea che il valore così ottenuto possa rappresentare un ulteriore indicatore, questa volta riferito all'intero Paese. Dai dati emergono i seguenti risultati:

- nell'anno 2015, la somma di tutti i *Corruption Indicator Score* ammonta a 7.254;
- nell'anno 2016, la somma di tutti i *Corruption Indicator Score* ammonta a 5.545;
- nell'anno 2017, la somma di tutti i *Corruption Indicator Score* ammonta a 6.627. Occorre tenere a mente che da un anno all'altro il numero di stazioni appaltanti considerate ha subito delle naturali variazioni. Non è detto che una determinata stazione appaltante abbia necessariamente bandito delle gare in ciascuno dei tre anni considerati. Per questo

motivo, il confronto tra i tre totali riportati in elenco è possibile solamente tenendo a mente la considerazione esposta poc'anzi. Dall'anno 2015 all'anno 2016 "l'indicatore Paese" ha subito una significativa riduzione, ma va considerato che il numero di stazioni appaltanti presenti nell'anno 2015 ammonta a 7044, mentre nell'anno successivo si riduce a 4847.

Tabella dei risultati finali per l'anno 2015 con il nuovo indicatore *"Cis outliers"*
In tabella sono riportate solo alcune delle stazioni appaltanti presenti nei dati

CFStazapp	p_i1_15	p_i1_15_out	p_i2_15	p_i2_15_out	p_i3_15	p_i3_15_out	p_i4_15	p_i4_15_out	p_i5_15	p_i5_15_out	p_i6_15	p_i6_15_out	Corruption_Indicator_Score	Cis_Outliers
02101050546	0.99	0.97	0.14	0.10	0.99	0.0	0.59	0.15	0.03	0.00	0.87	0.87	3.61	2.09
02675880583	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.25	0.18	0.00	0.00	3.25	2.68
01571730470	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.11	0.03	0.00	0.00	3.11	2.53
91002520293	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.10	0.02	0.00	0.00	3.10	2.52
97420690584	1.00	1.00	0.01	0.00	0.80	0.0	0.26	0.00	0.02	0.00	1.00	1.00	3.09	2.00
03462000161	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.06	0.00	0.00	0.00	3.06	2.50
86003470019	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.06	0.00	0.00	0.00	3.06	2.50
00221280688	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.05	0.00	0.00	0.00	3.05	2.50
01009620152	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.05	0.00	0.00	0.00	3.05	2.50
02747480123	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.05	0.00	0.00	0.00	3.05	2.50
80007520499	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.04	0.00	0.00	0.00	3.04	2.50
80012730224	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.04	0.00	0.00	0.00	3.04	2.50
95000360727	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.04	0.00	0.00	0.00	3.04	2.50
CFAVCP-0000D96	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.04	0.00	0.00	0.00	3.04	2.50
00111210779	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.03	0.00	0.00	0.00	3.03	2.50
00188370456	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.03	0.00	0.00	0.00	3.03	2.50
00301260048	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.03	0.00	0.00	0.00	3.03	2.50
00309450120	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.03	0.00	0.00	0.00	3.03	2.50
01644090134	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.03	0.00	0.00	0.00	3.03	2.50
02425670961	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.03	0.00	0.00	0.00	3.03	2.50
80243510585	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.03	0.00	0.00	0.00	3.03	2.50
94031080222	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.03	0.00	0.00	0.00	3.03	2.50
98003770785	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.03	0.00	0.00	0.00	3.03	2.50
00065080707	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.02	0.00	0.00	0.00	3.02	2.50
00190310052	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.02	0.00	0.00	0.00	3.02	2.50
00335800470	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.02	0.00	0.00	0.00	3.02	2.50
01199840115	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.02	0.00	0.00	0.00	3.02	2.50
04175700964	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.02	0.00	0.00	0.00	3.02	2.50
05562231216	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.02	0.00	0.00	0.00	3.02	2.50
80010010561	1.00	1.00	0.00	0.00	1.00	0.5	1.00	1.00	0.02	0.00	0.00	0.00	3.02	2.50

Figura 4.26: *Trend degli indicatori*

Dall'anno 2016 all'anno 2017 invece, "l'indicatore Paese" torna ad aumentare, ma aumentano anche il numero di stazioni appaltanti analizzate (5725).
Per provare a mitigare l'effetto del diverso numero di stazioni appaltanti, si è provveduto a costruire una nuova tabella, in grado di contenere esclusivamente le stazioni appaltanti che compaiono in tutti e tre gli anni considerati. Questa nuova tabella, non solo permetterà di ottenere dei nuovi "indicatori Paese" realmente confrontabili, ma rappresenterà allo stesso tempo un utile strumento di sintesi per visualizzare l'evoluzione di una determinata stazione appaltante anno dopo anno, in riferimento ai sei indicatori calcolati. La tabella è riportata in figura 4.27.
La tabella di figura 4.27, si rivela interessante sotto diversi aspetti. Innanzitutto può essere letta "a *slot"*, ovvero è possibile conoscere il dettaglio di ciascun punteggio di ciascuna stazione appaltante per l'anno 2015, 2016 e 2017. Dopodiché può essere letta dalla prima colonna all'ultima, per studiare l'evoluzione di una determinata stazione appaltante in riferimento ai sei indicatori calcolati, anno dopo anno (è questo il caso evidenziato a titolo d'esempio con il colore grigio). Infine, sulla tabella è possibile visualizzare il *Corruption Indicator Score* ottenuto da ciascuna stazione appaltante nell'anno considerato, per poterlo poi facilmente confrontare con l'anno successivo. Su questi dati, è stata ricalcolata la somma delle tre colonne indicanti il *Corruption Indicator Score*. Dai risultati emerge che:

- nell'anno 2015, la somma di tutti i *Corruption Indicator Score* ammonta a 1.030;
- nell'anno 2016, la somma di tutti i *Corruption Indicator Score* ammonta a 1.108;
- nell'anno 2017, la somma di tutti i *Corruption Indicator Score* ammonta a 1.133.

In questo nuovo scenario, si osserva che le variazioni su base annua dei totali sono molto più lievi del caso precedente, ed evidenziano un leggero aumento anno dopo anno. Di conseguenza questo aumento esprime un aumento del potenziale rischio di corruzione nel Paese.

4.4. Correlazioni tra punteggi - A conclusione delle analisi è stata calcolata la matrice di correlazione tra i sei punteggi ottenuti, per tutti e tre gli anni considerati. I risultati riferiti all'anno 2015 sono riportati nella figura 4.28 dove sono messe in evidenza tutte le correlazioni possibili. Seguono poi le figure 4.29 e 4.30 che riportano le correlazioni per gli anni 2016 e 2017. Quel che emerge, è che non vi è alcuna combinazione di indicatori che genera una forte correlazione positiva. Risultati analoghi si ottengono calcolando la matrice di correlazione dei sei indicatori riferiti agli anni 2016 e 2017. Quello che ci aspetterebbe, è una forte correlazione positiva tra gli indicatori utilizzati. In caso di correlazione forte e positiva infatti, ciascun indicatore rappresenterebbe un *proxy* della variabile corruzione, rappresentando quest'ultima in via indiretta.

Tabella dei risultati finali per gli anni 2015, 2016 e 2017

In tabella sono riportate le stazioni appaltanti che figurano in tutti gli anni considerati

CFStazapp	p1_15	p2_15	p3_15	p4_15	p5_15	p6_15	Cis_2015	p1_16	p2_16	p3_16	p4_16	p5_16	p6_16	Cis_2016	p1_17	p2_17	p3_17	p4_17	p5_17	p6_17	Cis_2017
00033120437	0.00	0.00	0.00	0.67	0.05	0.00	0.72	0.00	0.00	0.04	0.00	0.05	0.00	0.09	0.51	0.00	0.00	0.00	0.08	0.00	0.59
00034670943	1.00	0.00	0.00	0.50	0.03	0.00	1.53	0.00	0.00	0.05	0.00	0.05	0.00	0.10	0.70	0.00	0.82	0.50	0.06	0.00	2.08
00040490773	0.00	0.00	0.00	0.12	0.05	0.00	0.17	0.00	0.00	0.07	0.10	0.04	0.00	0.21	0.00	0.00	0.00	0.50	0.04	0.00	0.54
00040720070	0.52	0.00	0.00	0.67	0.00	0.00	1.19	0.82	0.00	0.79	0.50	0.01	0.00	2.12	0.93	0.00	0.00	0.40	0.02	0.00	1.35
00041720079	1.00	0.00	0.00	0.80	0.00	0.00	1.80	0.68	0.00	0.00	0.80	0.05	0.00	1.53	0.77	0.00	0.00	0.70	0.01	0.00	1.48
00050540327	0.00	0.00	0.14	0.12	0.06	0.00	0.32	0.00	0.00	0.39	0.21	0.14	0.00	0.74	0.30	0.01	0.48	0.08	0.12	0.00	0.99
00050800523	0.44	0.01	0.34	0.26	0.02	0.00	1.07	0.52	0.02	0.89	0.14	0.03	0.00	1.60	0.46	0.01	0.45	0.18	0.04	0.00	1.14
00051390318	0.00	0.01	1.00	0.00	0.02	0.00	1.03	0.00	0.00	1.00	0.00	0.06	0.00	1.06	0.00	0.01	0.89	0.00	0.03	0.00	0.93
00052090958	0.44	0.00	0.26	0.08	0.04	0.00	0.82	0.00	0.00	0.34	0.00	0.10	0.00	0.44	0.00	0.00	0.75	0.00	0.01	0.00	0.76
00053070918	0.00	0.00	0.19	0.31	0.04	0.00	0.54	0.00	0.01	0.97	0.00	0.04	0.00	1.02	0.00	0.00	0.07	0.00	0.05	0.00	0.12
00053520326	0.00	0.01	0.93	0.00	0.06	0.00	1.00	0.70	0.01	0.14	0.06	0.08	0.00	0.99	0.43	0.02	0.35	0.10	0.07	0.00	0.97
00055590327	0.00	0.00	0.09	0.64	0.04	0.00	0.77	0.00	0.00	0.89	0.67	0.05	0.00	1.61	0.34	0.00	0.52	0.50	0.09	0.00	1.45
00060850906	0.00	0.00	0.84	0.38	0.02	0.02	1.26	1.00	0.00	0.00	0.00	0.06	0.00	1.06	0.00	0.00	1.00	0.50	0.03	0.00	1.53
00061400073	0.00	0.00	1.00	0.00	0.01	0.00	1.01	0.00	0.00	0.00	0.25	0.03	0.00	0.28	0.00	0.00	1.00	1.00	0.00	0.00	2.00
00061800678	0.00	0.00	1.00	0.00	0.03	0.00	1.03	1.00	0.00	0.00	0.00	0.08	0.00	1.08	0.00	0.00	1.00	0.67	0.05	0.00	1.72
00061820742	0.00	0.00	0.00	0.00	0.05	0.00	0.05	0.00	0.00	0.00	0.00	0.06	0.00	0.06	0.00	0.00	0.00	0.00	0.05	0.00	0.05
00062890686	0.00	0.01	0.92	0.12	0.03	0.00	1.08	0.00	0.00	0.14	0.00	0.05	0.00	0.19	0.00	0.00	1.00	0.00	0.03	0.00	1.03
00063640684	0.00	0.01	1.00	0.25	0.03	0.00	1.29	0.00	0.00	1.00	0.50	0.00	0.00	1.50	0.00	0.00	1.00	0.75	0.09	0.00	1.84
00064240310	0.48	0.01	1.00	0.25	0.01	0.00	1.75	0.00	0.00	1.00	0.00	0.04	0.00	1.04	0.00	0.00	1.00	0.00	0.06	0.00	1.06
00064560709	0.06	0.01	0.73	0.25	0.03	0.00	1.08	0.00	0.00	1.00	0.00	0.02	0.00	1.02	0.46	0.00	0.68	0.00	0.03	0.00	1.17
00064780281	0.00	0.01	0.46	0.05	0.03	0.00	0.55	0.00	0.04	0.64	0.00	0.04	0.00	0.72	0.00	0.00	0.11	0.21	0.05	0.00	0.37
00067590703	0.00	0.00	0.00	0.00	0.04	0.00	0.04	0.00	0.00	0.14	0.00	0.08	0.00	0.22	0.00	0.00	1.00	0.00	0.03	0.00	1.03
00070680707	0.00	0.00	1.00	1.00	0.05	0.00	2.05	0.00	0.00	1.00	0.00	0.05	0.00	1.05	0.00	0.00	1.00	0.00	0.06	0.00	1.06
00071460935	0.00	0.00	1.00	0.00	0.03	0.00	1.03	0.00	0.01	0.89	0.07	0.03	0.00	1.00	0.00	0.03	0.96	0.00	0.05	0.00	1.04
00071560700	0.00	0.00	0.21	0.00	0.05	0.00	0.26	0.00	0.00	0.57	0.00	0.03	0.00	0.60	0.89	0.00	0.10	0.10	0.08	0.00	1.17
00071740955	0.00	0.00	1.00	0.00	0.03	0.00	1.03	0.00	0.00	1.00	0.00	0.03	0.00	1.03	0.00	0.00	1.00	0.00	0.32	0.00	1.32
00071770952	0.00	0.00	0.87	0.67	0.02	0.00	1.56	0.00	0.00	0.00	0.00	0.08	0.00	0.08	0.00	0.00	1.00	0.00	0.03	0.00	1.03
00072360050	0.00	0.01	0.83	0.35	0.03	0.00	1.22	0.44	0.01	0.23	0.40	0.04	0.00	1.12	0.54	0.01	0.41	0.29	0.04	0.00	1.29
00073740953	0.00	0.00	1.00	0.00	0.03	0.00	1.03	0.00	0.00	1.00	0.00	0.02	0.00	1.02	0.00	0.00	1.00	0.00	0.04	0.00	1.04
00073840894	0.00	0.00	0.00	0.00	0.05	0.00	0.05	0.00	0.01	1.00	0.00	0.01	0.00	1.02	0.00	0.00	0.57	0.00	0.04	0.00	0.61

Figura 4.27: *Tabella anni 2015, 2016 e 2017*

I risultati ottenuti però, suggeriscono una rivisitazione degli indicatori e magari un loro raffinamento, allo scopo di ottenere delle correlazioni maggiori. D'altra parte, anche il calcolo del coefficiente ↵di Cronbach sembrerebbe suggerire una rivisitazione degli indicatori: il coefficiente, misurato sui tre anni, ammonta a 0,51 (anno 2015), -0,26 (anno 2016) e -0,24 (anno 2017). I valori negativi dell'indice di Cronbach sono ascrivibili alla presenza di correlazioni molto deboli tra i vari indicatori (talvolta anche negative).

Le matrici di correlazione che seguono, riportano una situazione controintuitiva rispetto a quanto si aspetterebbe il lettore, giunto a questo punto. Intuitivamente infatti, gli indicatori si dovrebbero "muovere assieme", riportando una forte correlazione tra di essi. Questo non avviene in nessun caso e in nessuno dei tre anni considerati. Al netto di qualche riflessione e osservazione già fatta in precedenza, si ritiene che vi sia comunque del buono nei risultati ottenuti e degli utili punti di partenza per ulteriori riflessioni. Come motivazione della "conclusione controintuitiva" ottenuta, si è ipotizzato che gli indicatori misurano aspetti diversi dello stesso fenomeno, ma che ciò che determina quello che intendiamo per "stesso fenomeno" (la corruzione), si configura, nella realtà dei fatti, come una molteplicità di sotto-fenomeni differenti, che necessitano pertanto di indicatori di misurazione differenti, *non* necessariamente correlati positivamente fra loro.

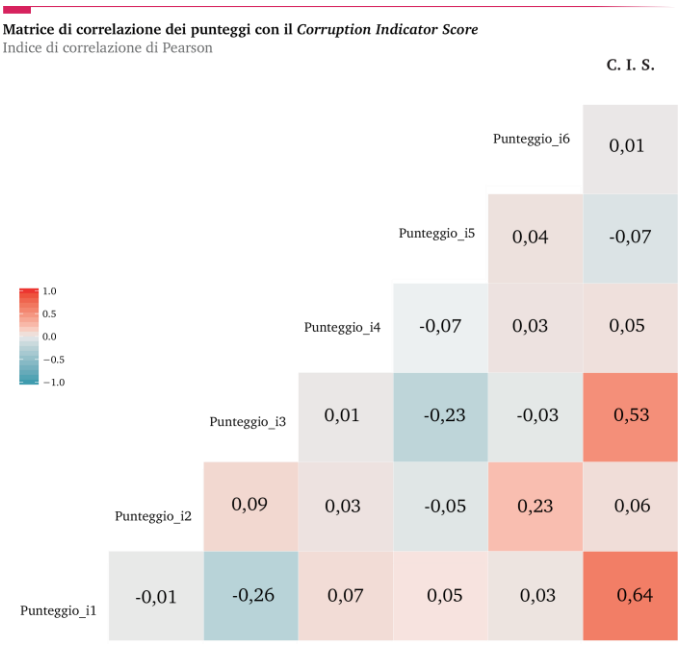

Figura 4.28: *Matrice delle correlazioni dei punteggi riferiti all'anno 2015.*
Le correlazioni appaiono deboli in larga misura. In alcuni casi risultano negative, anche se per valori molto ridotti. I valori di correlazione tra il *Punteggio_i1* e il *C.I.S.* e tra il *Punteggio_i3* e il *C.I.S.* risultano maggiori rispetto a tutti gli altri valori calcolati; tuttavia, in generale, la matrice esprime dei valori di correlazione piuttosto bassi, impedendo così di intercettare dei forti legami di correlazione tra gli indicatori esaminati.

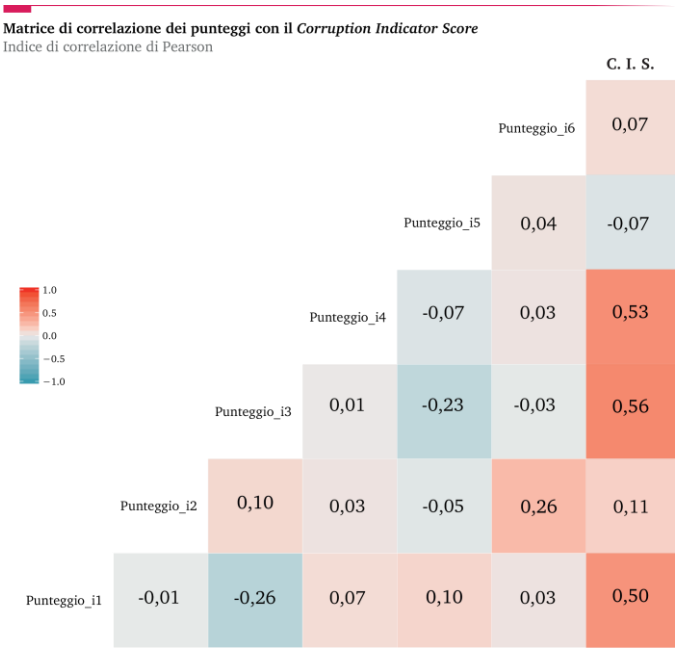

Figura 4.29: *Matrice delle correlazioni dei punteggi riferiti all'anno 2016*.
Le correlazioni appaiono deboli in larga misura. In alcuni casi risultano negative, anche se per valori molto ridotti. I valori di correlazione tra il *Punteggio_i1*, il *Punteggio_i3* e il l *Punteggio_i4* con il *C.I.S.* risultano maggiori rispetto a tutti gli altri valori calcolati; tuttavia, in generale, la matrice esprime dei valori di correlazione piuttosto bassi, impedendo così di intercettare dei forti legami di correlazione tra gli indicatori esaminati.

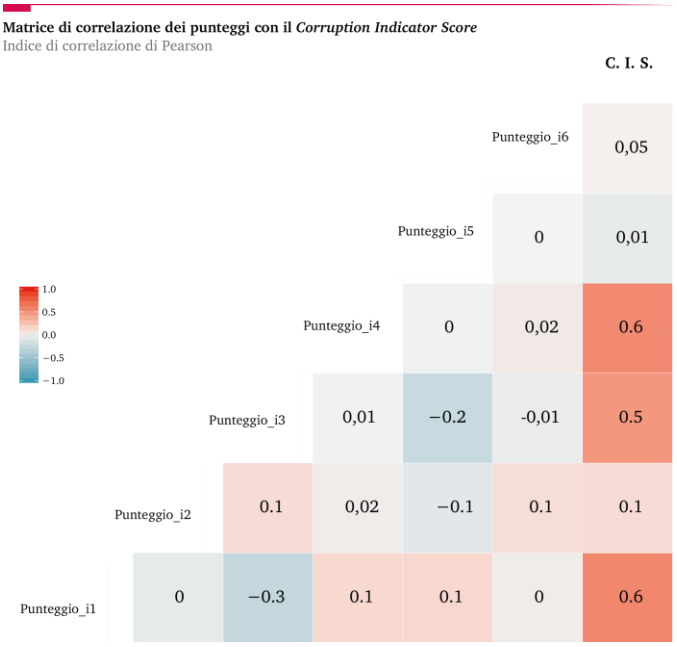

Figura 4.30: *Matrice delle correlazioni dei punteggi riferiti all'anno 2017*.
Le correlazioni appaiono deboli in larga misura. In alcuni casi risultano negative, anche se per valori molto ridotti. I valori di correlazione tra il *Punteggio_i1*, il *Punteggio_i3* e il l *Punteggio_i4* con il *C.I.S.* risultano maggiori rispetto a tutti gli altri valori calcolati; tuttavia, in generale, la matrice esprime dei valori di correlazione piuttosto bassi, impedendo così di intercettare dei forti legami di correlazione tra gli indicatori esaminati.

5. CONCLUSIONI

5.1. Il lavoro svolto e l'approccio adottato - A conclusione di questo lavoro, si ritiene opportuno riassumere brevemente i passaggi chiave che hanno condotto alla stesura di questo testo, e ancor prima alla ricerca e all'elaborazione di un percorso logico che potesse condurre soprattutto il lettore non addetto ai lavori, alla scoperta e alla conoscenza del tema della corruzione e di come questa possa essere misurata attraverso l'analisi dei dati.

Questa tesi nasce dalla volontà di riflettere su due tematiche diverse eppure connesse: da un lato la corruzione, dall'altro gli appalti pubblici.

Per quanto riguarda la corruzione, è stato necessario comprendere a fondo il significato di una parola che presenta molteplici significati, che si declinano in maniera diversa a seconda del contesto studiato e sulla base degli interessi di chi se ne occupa. Quello che emerge approfondendo il tema, è un complesso ed articolato sistema di definizioni da cui, diversamente da come si potrebbe pensare, non possiamo prescindere. Proprio per questo, i tentativi di ricondurre le diverse declinazioni ad un'unica definizione universalmente riconosciuta apparirebbe in ultima analisi uno sforzo vano, che ridurrebbe e semplificherebbe troppi aspetti che spesso risultano fondamentali. Assumendo che questa riflessione sia condivisa dal lettore, è bene ricordare la principale difficoltà che ne consegue: il tentativo di formalizzare algoritmicamente il concetto di corruzione diventa una sfida complessa.

In relazione al tema degli appalti invece, la formalizzazione è stata più semplice. Si è rivelato molto utile descrivere i primi *step* di un appalto pubblico attraverso la costruzione e l'implementazione del modello entità-relazione e attraverso l'ontologia presentata. Poiché gli appalti pubblici evolvono secondo una serie di passaggi standard definiti dalla normativa, la loro "trattazione algoritmica" è risultata più facile.

A valle di queste premesse, si ritiene importante mettere in evidenza un primo risultato raggiunto, forse piccolo, ma fondamentale. La "traduzione" dei diversi concetti discorsivi affrontati in questa tesi (corruzione, appalto, affidamento, bando di gara, …) nel linguaggio dell'informatica (tramite

l'utilizzo di apposite strutture dati, di librerie di codice e funzioni *ad hoc*, di formalizzazioni concettuali, ...) è stato il primo passaggio utile per poter lavorare in maniera efficace sui dati.
L'idea di mettere i dati al centro del ragionamento esposto in questa tesi, si ritiene essere per certi aspetti un approccio innovativo rispetto alla maggior parte dei contributi presenti in letteratura. Il dato diventa non più un elemento di corredo rispetto alla trattazione, non più un noioso insieme di numeri relegati tra gli allegati finali dei report, ma il faro che illumina l'esplorazione, la guida ai nostri ragionamenti e alle nostre indagini, la misura con cui occorre confrontarsi ripetutamente man mano che si avanza nella trattazione e che si operano delle scelte. Questa è stata la novità che si è tentato di presentare in questo lavoro, nell'idea che attraverso i dati possiamo comprendere fenomeni complessi come quelli studiati in questo testo. Di questa opinione sono anche gli autori del libro "Creare valore con i Big Data[1]", quando scrivono: *"stiamo vivendo (. . .) una fase storica in cui la complessità emerge come sfida e richiede la definizione di nuovi paradigmi. D'altra parte, la crescente disponibilità di dati digitali, prodotti dalle macchine e dalle azioni umane, offre una rappresentazione viva dell'evoluzione della realtà, permettendo di esaminare i fenomeni nelle loro dinamiche e di comprendere le leggi che li regolano."*
Questa "rappresentazione viva dell'evoluzione della realtà" deve essere collocata sempre di più tra le prorità dei *decision maker* a tutti i livelli istituzionali.

5.2. Le criticità emerse - Ogni *data scientist* è consapevole del fatto che i dati del mondo reale non sono mai perfetti come quelli utilizzati a scopi didattici. Il dato è il prodotto di un processo spesso ancora frutto del lavoro umano; proprio per questo è facile che esso sia sporco, incoerente o incompleto. Anche i dati analizzati in questa tesi si sono rivelati affetti da alcuni problemi, che hanno generato alcune criticità, riportate di seguito.

1. criticità connesse ai dati ricevuti;
2. criticità connesse agli indicatori di corruzione misurati.

In riferimento alle *criticità connesse ai dati ricevuti*, si ritiene opportuno segnalare le difficoltà riscontrate in diversi momenti delle analisi, dovute a dati sporchi o incompleti. Spesso ci si è imbattuti in importi di aggiudicazione negativi, o in importi esageratamente grandi. Oppure sono state riscontrare delle anomalie sugli attributi relativi alle date, che hanno fatto emergere situazioni anomale, come ad esempio delle gare aggiudicate 800 anni dopo o diversi anni prima della data della loro pubblicazione, solo per citare alcuni tra i casi più eclatanti.

[1]Creare valore con i Big Data - Gli strumenti, i processi, le applicazioni pratiche. L. Camiciotti, C. Racca - Edizioni LSWR

Al netto dei casi d'esempio citati e di tante altre problematiche legate alla pulizia dei dati, problematiche che come detto in apertura, fanno parte in maniera naturale di qualsiasi *dataset*, sembrerebbe che a livello informatico non vi siano ancora sufficienti meccanismi di controllo sui dati immessi dalle diverse stazioni appaltanti, in modo tale che questi possano essere raccolti alla massima qualità possibile.

Per quanto invece riguarda le *criticità connesse agli indicatori di corruzione misurati*, è utile partire dalle matrici di correlazione riportate nel capitolo precedente. Nonostante la variabile *corruzione* sia una variabile non direttamente osservabile all'interno dei dati disponibili, ci si sarebbe aspettato che essa fosse positivamente correlata ai fenomeni indagati nelle analisi, e quindi ai valori degli indicatori che misurano questi fenomeni. Quello che emerge dalle diverse matrici di correlazione, è un insieme di valori il più delle volte opposti rispetto a quelli attesi. In uno scenario ottimale, la matrice delle correlazioni dovrebbe apparire valorizzata con valori positivi prossimi al valore 1 (forte correlazione positiva) in gran parte delle celle rappresentate. Questo non avviene nel caso in esame, e intuitivamente suggerisce una nuova valutazione degli indici considerati, al fine di migliorare i risultati finali e sperare in una serie di correlazioni più forti.

Tuttavia, come si è suggerito in conclusione del capitolo precedente, potrebbe rivelarsi utile ammettere anche un'altra strada. Intendere la corruzione come un fenomeno complesso, articolato, diverso per impatti, forme, contesti e soggetti coinvolti, spingerebbe all'adozione di una serie di indicatori differenti e specializzati sui diversi aspetti via via considerati, da cui non aspettarsi, necessariamente, una forte correlazione positiva. Questo è il caso che si ritiene di aver ottenuto in questa tesi. Sarebbe certamente interessante proseguire con ulteriori analisi al fine di validare la bontà di questa seconda opzione. Un'indagine "per tematiche" dei dati processati, potrebbe rivelarsi un utile spunto per proseguire il lavoro su questa seconda strada. Si potrebbe ad esempio decidere di studiare il tema delle gare stipulate con una forma di scelta del contraente che non prevede alcun aspetto competitivo, e per questo particolare tematica, costruire di conseguenza una serie di indicatori *ad hoc*, nell'intento di mettere in campo misurazioni diverse all'interno però dello stesso perimetro. Immaginando di avere realizzato un *set* di indicatori tematici, questi restituirebbero un rischio di potenziale corruzione in riferimento alla sola tematica studiata (ad esempio la scelta di procedure di gara non competitive) e non un insieme di valutazioni su aspetti differenti fra loro.

Indicazioni conclusive - In conclusione di questo testo, si ritiene utile condividere con il lettore alcune osservazioni, molte delle quali imparate proprio realizzando questo lavoro. Ci si augura che il lettore con un background scientifico e informatico abbia trovato giovamento dalle introduzioni teoriche collocate all'inizio questa tesi. Comprendere il dominio di una qualsiasi analisi dati è di assoluta importanza per condurre delle analisi sensate.

Se ai tecnici è consigliato studiare ed approfondire, in questo caso, il mondo giuridico ed amministrativo, a quest'ultimo ci si sente di suggerire l'individuazione di sempre più profili tecnici all'interno delle istituzioni adibite al contrasto e alla prevenzione della corruzione. La multidisciplinarietà di questo lavoro sottolinea l'importanza della collaborazione e della contaminazione di diverse competenze, al fine di superare con efficacia le sfide complesse che ci si prospettano.

Una seconda indicazione riguarda la formalizzazione del *Corruption Indicator Score* riportata di seguito:

$$Corruption\ Indicator\ Score = \sum_{i=1}^{k} \varphi_i$$

dove $\varphi_i = punteggio_{ind_k}$, con $k = 6$.

La formula rappresenta un tentativo di formalizzare un indice di corruzione oggettiva, in grado di superare gran parte dei limiti degli indicatori *perception based*. La formula inoltre, è nata sulla base dei dati relativi agli appalti pubblici italiani, e questo conduce ad almeno due considerazioni di merito. Innanzitutto il modello proposto dal *Corruption Indicator Score*, seppur imperfetto, si candida ad essere una prima rudimentale formalizzazione oggettiva della corruzione basata sui dati degli appalti pubblici italiani. Si ritiene che il modello costruito rispecchi due caratteristiche importanti: la semplicità con cui viene rappresentato, non essendo altro che una semplice somma, e la facile possibilità di estensione, attraverso nuovi ed ulteriori indicatori, che potrebbero essere aggiunti a quelli già presenti. In secondo luogo, si ritiene importante rendere più trasparenti i meccanismi di raccolta dei dati, in maniera tale che possano essere conosciuti anche dai non addetti ai lavori e magari migliorati attraverso dei sistemi di *feedback*. All'interno di questa tesi si è ampiamente discusso dell'importanza di avere dati di qualità, dell'esigenza di dotarsi di standard nazionali ed internazionali, del dovere che la pubblica amministrazione e le istituzioni in generale hanno di rendere facilmente accessibili i loro patrimoni informativi, e dell'opportunità, in sintesi, di considerare i dati il cuore dell'azione politica e istituzionale del nostro Paese.

Ci attendono delle sfide estremamente complesse dal punto di vista sociale e quindi istituzionale e politico. Spiegare la complessità non è semplice, ma i dati possono essere nostri alleati. D'altra parte, come sosteneva l'economista Ronald Coase, *if you torture the data long enough, it will confess to anything.*

ABSTRACT

Questo *working paper* rappresenta un estratto della Tesi di Laurea Magistrale in Informatica redatta dall'autore riportato in copertina nell'anno accademico 2019 - 2020. La tesi è stata discussa presso il Dipartimento di Scienze Matematiche, Informatiche e Fisiche dell'Università degli Studi di Udine in data 13.03.2020. Il relatore della tesi è stato l'On. Prof. Paolo Coppola.

Il presente *working paper* si pone l'obiettivo di condurre un'analisi della Banca Dati Nazionale dei Contratti Pubblici, nell'intento di costruire, attraverso l'analisi, degli indicatori di rischio di corruzione potenziale basati sui dati. La Banca Dati Nazionale dei Contratti Pubblici rappresenta uno dei patrimoni informativi più significativi del nostro Paese. L'accesso a questi dati, seppur contingentato ad un solo triennio, ha permesso di testare alcune tecniche di *data analysis* con cui si sono ricavati importanti risultati.

A partire dall'analisi effettuata, è stato costruito un nuovo indice di corruzione denominato *Corruption Indicator Score*, che a differenza dei tanti indicatori già presenti in letteratura supera i limiti degli indicatori *perception based*, poiché in grado di elaborare non più le opinioni e le percezioni di un campione di intervistati, ma i dati ufficiali delle amministrazioni pubbliche italiane.

Il documento riporta alcune modifiche e integrazioni rispetto al testo originale, nate a valle del confronto e dei preziosi suggerimenti ricevuti da parte dei componenti del Comitato scientifico della Collana ANAC.

BIBLIOGRAFIA

[1] Staffan ANDERSSON e Paul M HEYWOOD. The politics of perception: use and abuse of transparency international's approach to measuring corruption. *Political studies*, 57(4):746–767, 2009.

[2] Jens ANDVIG. Alternative perspectives. *A comment on corruption and governance in B. Lomborg (ed.) Global Crises, Global Solutions, Cambridge*, pp. 345–355, 2004.

[3] Autorità Nazionale Anticorruzione. *Pubblicazione quadrimestrale - Secondo quadrimestre*. Anac, 2019.

[4] Andrea ASPERTI e Agata CIABATTONI. *Logica a informatica*. McGraw-Hill, 1997.

[5] Paolo ATZENI, Stefano CERI, Stefano PARABOSCHI, e Riccardo TORLONE. *Basi di dati: modelli e linguaggi di interrogazione (seconda edizione)*. McGraw-Hill, 2006.

[6] Tim BERNERS-LEE, James HENDLER, e Ora LASSILA. The semantic web. *Scientific american*, 284(5):34–43, 2001.

[7] Paola BONACCI. *Il riordino normativo nel settore degli appalti di opere pubbliche*. Camera dei Deputati, 1995.

[8] Konstantinos BOVALIS, Vassilios PERISTERAS, Margarida ABECASIS, Raul-Mario Abril- JIMENEZ, Miguel Alvarez RODRIGUEZ, Corinne GATTEGNO, Athanasios KARALOPOULOS, Ioannis SAGIAS, Szabolcs SZEKACS, e Suzanne WIGARD. Promoting interoperability in europe's e-government. *Computer*, 47(10):25–33, 2014.

[9] Mark BOVENS, Robert E GOODIN, e Thomas SCHILLEMANS. *The Oxford handbook public accountability*. Oxford University Press, 2014.

[10] Raffaele CANTONE e Francesco CARINGELLA. *La corruzione spuzza: tutti gli effetti sulla nostra vita quotidiana della malattia che rischia di uccidere l'Italia.* Mondadori, 2017.

[11] Raffaele CANTONE e Enrico CARLONI. *Corruzione e anticorruzione: dieci lezioni*. Feltrinelli Editore, 2018.

[12] Marco CELENTANI e Juan-José GANUZA. Corruption and competition in procurement. *European Economic Review*, 46(7):1273–1303, 2002.

[13] Peter Pin-Shan CHEN. The entity-relationship model—toward a unified view of data. *ACM transactions on database systems (TODS)*, 1(1):9–36, 1976.

[14] Edgar F CODD. *The relational model for database management: version 2*. Addison- Wesley Longman Publishing Co., Inc., 1990.

[15] Edwin DE JONGE e Mark VAN DER LOO. *An introduction to data cleaning with R*. Statistics Netherlands Heerlen, 2013.

[16] William DE MARIA. Measurements and markets: deconstructing the corruption perception index. *International Journal of Public Sector Management*, 2008.

[17] Donatella DELLA PORTA e Alberto VANNUCCI. Forme di controllo e corruzione politica in italia. 1997.

[18] Fabio DI CRISTINA. La corruzione negli appalti pubblici. *Rivista trimestrale di diritto pubblico*, 1:177–226, 2012.

[19] Elisa D'ALTERIO. Luci e ombre del sistema degli acquisti delle pubbliche amministrazioni. *Gli acquisti delle amministrazioni pubbliche nella Repubblica federale, a cura di L. Fiorentino, Bologna, Il Mulino*, pp. 19–51, 2011.

[20] Ramez A ELMASRI e Shamkant B NAVATHE. *Sistemi di basi di dati. Fondamenti*. Pearson Italia Spa, 2004.

[21] Mihály FAZEKAS e István János TÓTH. New ways to measure institutionalised grand corruption in public procurement. *U4 Brief: October*, 9:U4, 2014.

[22] I FILIPPETTI. Osservatorio appalti pubblici e legalità. *Urbanistica e appalti*, 10:1203– 1206, 2008.

[23] Nadia FIORINO e Emma GALLI. *La corruzione in Italia: un'analisi economica*. Il mulino, 2013.

[24] Autorità garante della concorrenza e del mercato. *Appalti pubblici e concorrenza*. AGCM, 1992.

[25] Miriam A GOLDEN e Lucio PICCI. Proposal for a new measure of corruption, illustrated with italian data. *Economics & Politics*, 17(1):37–75, 2005.

[26] Stephen KNACK. *Measuring corruption in Eastern Europe and Central Asia: a critique of the cross-country indicators*. The World Bank, 2006.

[27] Maurizio LISCIANDRA e Emanuele MILLEMACI. The economic effect of corruption in italy: a regional panel analysis. *Regional Studies*, 51(9):1387–1398, 2017.

[28] G MELE. La dimensione economica e il funzionamento del mercato degli appalti pub- blici. *Relazione presentata al Convegno Confindustria, Concorrenza come bene pubblico, Vicenza*, 2006.

[29] Francesco MERLONI e Luciano VANDELLI. *La corruzione amministrativa: cause, prevenzione e rimedi*. Passigli editori, 2010.

[30] GNALDI Michela e PONTI Benedetto. Misurare la corruzione oggi. obiettivi, metodi, esperienze, 2018.

[31] Riccardo MILANI, Francesco CALDERONI, Carlotta CARBONE, e Martina ROTONDI. L'impatto di corruzione e mafia sugli appalti pubblici: un'esplorazione empirica. 2018.

[32] K PEARSON. Notes on regression and inheritance in the case of two parents proceedings of the royal society of london, 58, 240-242, 1895.

[33] Jeremy POPE. Ti source book 2000: Confronting corruption: The elements of a national integrity system. 2000.

[34] Tina SØREIDE. *Corruption in public procurement. Causes, consequences and cures*. Chr. Michelsen Intitute, 2002.

[35] Tina SØREIDE. *Is it wrong to rank? A critical assessment of corruption indices*. Chr. Michelsen Institute, 2006.

[36] Irene TINAGLI. *La Grande Ignoranza*. Rizzoli, 2019.

[37] Alberto VANNUCCI. *Atlante della corruzione*. Associazione Gruppo Abele Onlus-Edizioni Gruppo Abele, 2017.

[38] Hadley WICKHAM e altri. Tidy data. *Journal of Statistical Software*, 59(10):1–23, 2014.

[39] Hadley WICKHAM e Garrett GROLEMUND. *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.", 2016.

[40] G Udny YULE. Why do we sometimes get nonsense-correlations between time-series?– a study in sampling and the nature of time-series. *Journal of the royal statistical society*, 89(1):1–63, 1926.