DATA QUALITY

TRASPARENZA e DATI APERTI per prevenire la corruzione:

COME SI COMPORTANO I COMUNI ITALIANI?

UN'ANALISI SISTEMATICA SULLA QUALITÀ DEI DATI RELATIVI AI CONTRATTI DI ACQUISTO DEI COMUNI ITALIANI

synapta

ContrattiPubblici.org

DATA QUALITY

Introduzione

L'ANALISI DEI DATI

Le pubbliche amministrazioni italiane spendono ogni anno circa 150 miliardi di Euro per l'acquisto di servizi e forniture e per l'esecuzione di lavori[1]. Le PA hanno l'obbligo di pubblicare online i dati che riguardano questi affidamenti.

1 - Fonte: Autorità Nazionale Anticorruzione, Relazione annuale 2018, secondo la quale *"nel 2018 il valore complessivo degli appalti di importo pari o superiore a 40.000 euro per entrambi i settori ordinari e speciali si è attestato attorno ai 139,5 miliardi di euro"*.
A tale ammontare vanno aggiunti circa 11,25 miliardi di euro, che sono il valore stimato da *ContrattiPubblici.org* per gli affidamenti sotto la soglia dei 40.000 euro.

01

Il valore assoluto dei contratti pubblici stipulati ogni anno giustifica da solo l'interesse verso questi dati; guardarli in rapporto alla ricchezza prodotta dal Paese nel suo complesso rafforza questa impressione: si tratta dell'8,5% del PIL Italiano[2] (o circa il 5,5% se si considerano "solo" i circa 100 miliardi relativi agli acquisti intermedi di beni e servizi). Ma oltre al significativo impatto diretto sull'economia, il procurement pubblico è strategico per gli effetti che può avere sull'innovazione (anche tramite la trasformazione digitale della stessa PA), nonché sulla dotazione infrastrutturale (materiale ed immateriale) del Paese. Infine, l'interesse pubblico verso i dati relativi a contratti e bandi è anche legato alla prevenzione della corruzione e di altre cattive pratiche amministrative.[3]

Non a caso, l'obbligo di pubblicare i dati relativi ai contratti pubblici è stato normato in modo molto preciso dalla Legge Anticorruzione (legge n. 190/2012), ai sensi della quale le stazioni appaltanti (l'obbligo si estende, ad esempio, oltre alle PA in senso stretto, anche a controllate e partecipate, limitatamente alla loro attività di pubblico interesse) sono tenute a pubblicare nei propri siti web istituzionali una serie di dati, che include: la struttura proponente; l'oggetto del contratto; l'elenco degli operatori invitati a presentare offerte; l'aggiudicatario; l'importo di aggiudicazione; i tempi di completamento dell'opera, servizio o fornitura; l'importo delle somme liquidate.
La pubblicazione di questi dati è prevista entro il 31 gennaio di ogni anno e l'adempimento viene vigilato dall'Autorità Nazionale Anticorruzione (ANAC), che ha previsto specifici standard tecnici per la pubblicazione e segnala i soggetti inadempienti alla Corte dei Conti.
I dati pubblicati ai sensi della Legge Anticorruzione sono ricordati esplicitamente nel Decreto Trasparenza (Decreto legislativo 33/2013), che regola il diritto di accesso civico e gli obblighi di pubblicità, trasparenza e diffusione di informazioni da parte delle pubbliche amministrazioni. In particolare, questi dati fanno parte de "i documenti, le informazioni e i dati oggetto di pubblicazione obbligatoria" ai sensi della normativa vigente e sono dunque "Dati Aperti" (o Open Data) (Art. 7), il cui riutilizzo è libero ai sensi della normativa sull'informazione detenuta dal settore pubblico "senza ulteriori restrizioni diverse dall'obbligo di citare la fonte e di rispettarne l'integrità". Si tratta quindi di dati su cui cittadini, associazioni ed imprese possono fare ulteriori analisi e costruire servizi a valore aggiunto.

Questa ricerca ha esplorato la qualità dei Dati Aperti dei contratti pubblicati dai 7.914[4] **Comuni italiani**, dal 2012 al 2018. Il nostro studio ha rilevato **una generale buona qualità** senza sistematiche differenze geografiche, ma con una quota minoritaria ma ancora rilevante di singole pubbliche amministrazioni che pubblicano dati la cui bassa qualità rende sostanzialmente impossibile il monitoraggio civico.
Abbiamo poi concentrato l'analisi sui diversi aspetti di uno degli indicatori più interessanti, quello relativo alla completezza dei dati, e infine proposto un focus sulle 12 maggiori città per popolazione.

2 - Secondo il comunicato ISTAT del 1 Marzo 2019, "*nel 2018 il Pil ai prezzi di mercato è stato pari a 1.753.949 milioni di euro correnti*".
3 - Si veda anche Misurare la corruzione oggi. Obiettivi, metodi, esperienze , a cura di M. Gnaldi e B. Ponti, open access su *https://ojs.francoangeli.it/_omp/index.php/oa/catalog/book/310*.
4 - Dato aggiornato Istat a ottobre 2019.

ContrattiPubblici.org

Principale progetto dell'azienda Synapta S.r.l,[5] *ContrattiPubblici.org* è un database online che raccoglie tutti i dati dei contratti pubblici resi disponibili online dalle pubbliche amministrazioni (PA in seguito) italiane dal 2013.

Ad ottobre 2019 sul portale sono presenti **19 milioni** di contratti di più di **20 mila PA** con più di un **milione di fornitori**. I contratti riguardano qualunque settore - dalla fornitura di farmaci a quella di cancelleria, dai servizi di pulizia alle grandi opere - per qualunque importo - da pochi euro a miliardi -, di qualunque zona d'Italia, di qualunque tipo di PA - dalla Regione al piccolo Comune, dalla scuola alla ASL, dall'Università al tribunale (anche se questo report si concentra sui Comuni). Sono presenti anche i contratti di molte SpA ed alcune fondazioni a controllo pubblico. Per la provenienza dei dati si rimanda all'omonima sezione in *Appendice*.

> **Su ContrattiPubblici.org sono presenti 19 milioni di contratti di più di 20 mila PA con più di un milione di fornitori**

5 - Nata come spin off del centro di ricerca Nexa del Politecnico di Torino nel 2016, Synapta si è occupata fino ad oggi di analisi dati con tecnologie open source e innovative nel settore.

DATA QUALITY

Metodo

Abbiamo associato un indice di qualità ai contratti:
più errori ci sono e più è basso l'indice.

02

Cosa significa che un *dato è di qualità*? Con dati *puliti* o *di qualità* ci si riferisce a dati che sono privi di errori, standardizzati in un formato unico per ogni campo in modo tale che si possano confrontare fra loro e che siano coerenti rispetto al loro significato e al resto dei dati. Dati puliti permettono di effettuare analisi affidabili e quanto più attinenti alla realtà. Viceversa, in presenza di dati di bassa qualità, è complesso - e a volte addirittura fuorviante - fare leva su questi ultimi per prendere decisioni.

Per conoscere quali fossero le condizioni qualitative dei dati pubblicati dai Comuni italiani abbiamo usato come riferimento lo standard ISO/IEC 25024[6] "*Measurement of data quality: provides measures including associated measurement methods and quality measure elements for the quality characteristics in the data quality model.*" per ricavarne un set di misure di data quality per ogni tipologia di dato presente nel tracciato di un contratto.

Sulla base dello standard ISO abbiamo identificato le seguenti categorie di misurazione:

- **Completezza**: un dataset è completo se riporta, per ogni entità, tutti gli attributi di cui è richiesta la compilazione (es., tutte le celle di una tabella sono state compilate)
- **Coerenza**: un dato è detto coerente se ha attributi che non hanno contraddizioni e che sono coerenti rispetto agli altri dati nel loro contesto d'uso

> **Dati puliti permettono di effettuare analisi affidabili.**

- **Precisione**: un dato è preciso quando ha attributi esatti o se è nel giusto formato, la precisione riguarda perciò l'espressione del dato

Queste categorie sono state scelte considerando quelle più interessanti, ma anche ciò che poteva essere calcolato automaticamente, dovendo compiere l'analisi su milioni di contratti.

I campi del tracciato di un contratto considerati in questa ricerca[7] sono:

- **CIG**: codice alfanumerico identificativo del contratto (Codice Identificativo Gara)
- **Importo**: importo pattuito per il servizio fra la PA e l'ente fornitore
- **Importo liquidato**: somme effettivamente erogate dalla PA al fornitore
- **Data inizio**: data di inizio per il servizio, lavoro o fornitura del contratto
- **Data fine**: data stimata di fine/completamento del servizio, lavoro o fornitura
- **Oggetto**: descrizione testuale del servizio, lavoro o fornitura oggetto del contratto
- **Scelta contraente**: tipologia di procedura utilizzata per la scelta del contraente

6 - *http://www.iso25000.it/styled-4/* e *https://www.iso.org/standard/35749.html*

7 - Sono esclusi gli identificativi della PA stessa (codice fiscale o partita iva) e i partecipanti, rispettivamente perché per i primi è necessario che gli identificativi siano ben compilati per poter identificare che il contratto appartiene proprio ad un Comune e perché per i secondi approfondire la qualità della compilazione dei dati riguardanti i partecipanti richiederebbe uno studio a parte.

Per calcolare un indice di qualità per ogni singolo contratto abbiamo verificato su tutto il tracciato la presenza dei possibili errori per ogni categoria di misurazione, dove applicabili (si consulti la Tabella 1 in *Rilevazioni* per trovare le categorie applicabili ai campi).

Ad esempio, abbiamo considerato se i campi erano compilati o meno per la completezza; se la data di inizio di un contratto era precedente a quella di fine per la coerenza; se gli importi avevano il giusto numero di cifre dopo la virgola per la precisione. Per un elenco esaustivo di tutte le casistiche consultare l'*Allegato A*.

Abbiamo dunque valutato ad uno ad uno se erano presenti gli errori identificati e poi abbiamo associato un punteggio ad ogni categoria: più errori sono presenti in quella categoria di misurazione, più il punteggio della categoria stessa si abbassa. Prendendo ad esempio la completezza, più sono i campi che mancano e più è basso il "punteggio" relativo alla completezza, ovvero l'**Indice di completezza**.
Abbiamo infine riassunto con una media pesata[8] tutti i punteggi delle singole categorie, ottenendo quindi un unico valore rappresentativo della qualità complessiva del contratto, l'**Indice di qualità** del contratto: 1 è il massimo ed è associato ad un contratto perfettamente compilato (dal punto di vista degli indicatori che abbiamo misurato in modo automatico), mentre 0 è il minimo; empiricamente, possiamo affermare che già per valori attorno allo 0.5 la comprensibilità del contratto risulta sostanzialmente compromessa.

Ad ogni modo, un punteggio alto è una condizione necessaria ma non sufficiente per una buona qualità a tutto tondo dei dati relativi ad un contratto. Alcuni casi di incoerenza o di dati formalmente corretti ma bizzarri, come può essere la fornitura di carta igienica da parte di un'azienda informatica, sono errori ovvi per un essere umano ma complessi da individuare automaticamente in modo affidabile. Per fare di più, si potrebbe ricorrere ad algoritmi di *machine learning*[9] o prevedere la pubblicazione obbligatoria come dati aperti di altre informazioni, come la categoria merceologica (CPV)[10] o il codice Ateco[11] dei fornitori insieme al contratto. Gli indicatori riassunti in questo report ci paiono comunque sufficienti ad individuare numerosi casi di cattiva pubblicazione dei dati sulla trasparenza amministrativa e per questo abbiamo deciso di rendere pubbliche queste misurazioni.

8 - Ad ogni categoria di misurazione è stato associato un peso scelto sull'importanza del campo nel contesto dei contratti pubblici e sull'incidenza dell'errore nell'abbassamento della comprensione del contratto stesso. Tale peso ha elementi di arbitrarietà di cui siamo ben consci, ma è indispensabile a riassumere in un unico numero la qualità complessiva di ogni contratto.
9 - Il machine learning è lo studio scientifico degli algoritmi e dei modelli statistici che i computer usano per eseguire specifici task senza ricevere istruzioni esplicite, basandosi invece su pattern e inferenza.
10 - Common Procurement Vocabulary, classificazione unica per gli appalti pubblici.
11 - ATtività ECOnomiche, classificazione delle attività economiche.

DATA QUALITY

Rilevazioni

La qualità dei dati rilevata è spesso buona. Per questo abbiamo potuto utilizzarli nella costruzione di servizi come *ContrattiPubblici.org*.

Là dove la qualità si abbassa, è possibile che la pubblicazione sia stata considerata una formalità da sbrigare rapidamente. Anche per questo, pensiamo che mostrare il potenziale del riutilizzo sia utile ad ottenere dati di qualità crescente

03

Indice generale

La distribuzione della qualità dei contratti dei Comuni italiani è concentrata su valori alti dell'Indice di qualità (Grafico 2) - ovvero quelli maggiori di 0.95 - con una lunga coda verso valori più bassi - dove un contratto con indice sotto 0.6 è quasi completamente vuoto, e sotto 0.5 è già sostanzialmente incomprensibile. Ciò significa che generalmente i contratti sono pubblicati bene, ma anche che esistono numerose eccezioni a questa tendenza generale.
Indicativamente, nel Grafico 1 si vede come più del 60% dei contratti è di qualità alta, circa il 30% è di qualità intermedia (ovvero presenta errori, ma è ancora leggibile) mentre la percentuale rimanente è nella zona bassa di valori dell'Indice di qualità, dove i contratti sono incomprensibili (e la loro pubblicazione è sostanzialmente inutile o addirittura fuorviante).

GRAFICO 1. **FASCE DELL'INDICE DI QUALITÀ PER CONTRATTO**

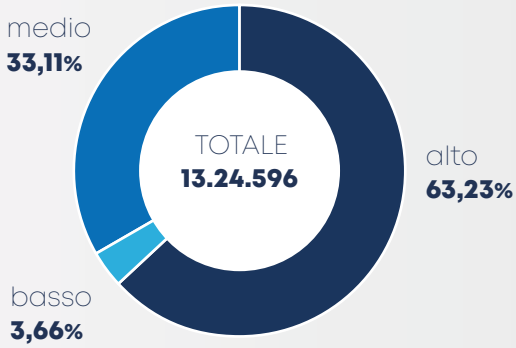
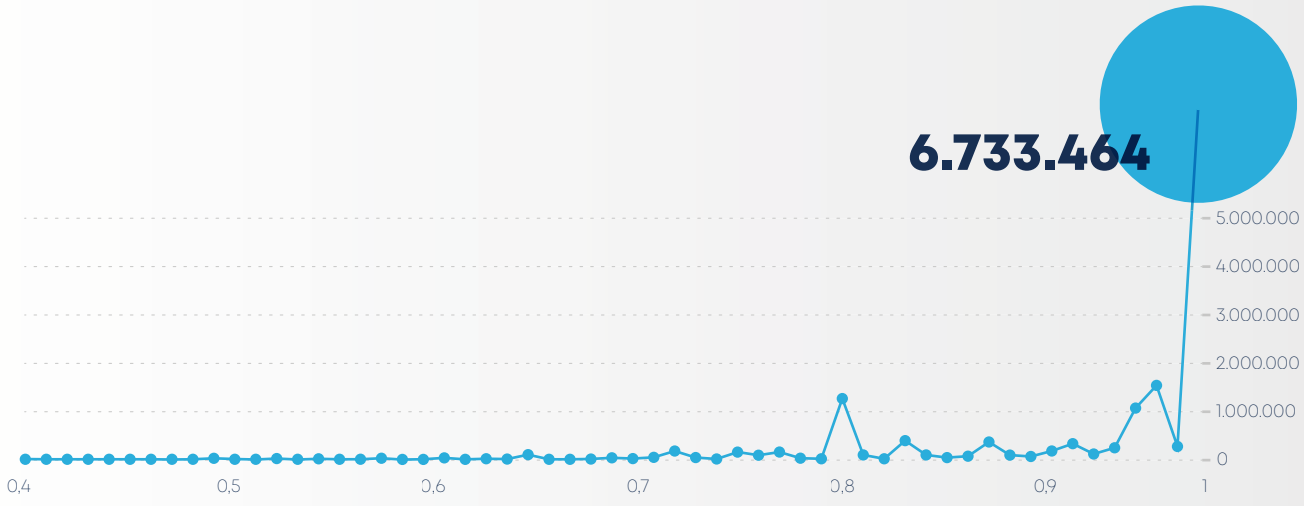


GRAFICO 2. **INDICE DI QUALITÀ DI TUTTI I CONTRATTI DAL 2012 AL 2018**

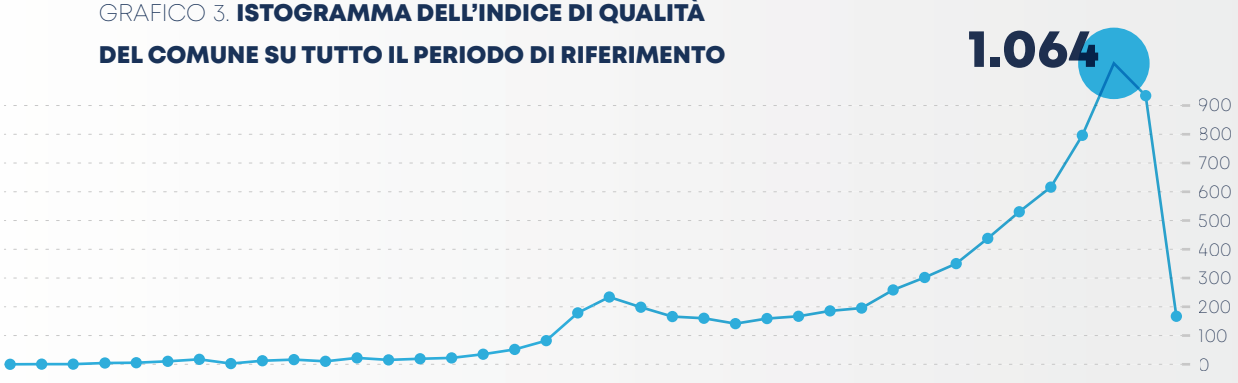


> **Abbiamo poi concentrato l'analisi sulla completezza dei dati e infine un focus sulle 12 principali città per popolazione.**

Siamo convinti che, là dove la qualità si abbassa, si sia sottovalutata l'importanza della divulgazione di questi dati e con essa anche le richieste tecniche e la precisione indispensabile per la loro compilazione. Un'ipotesi è che la pubblicazione sia percepita come una pura formalità dalle PA: quando il dato non ha uno scopo, che sia di analisi o consultazione, si perde la necessità di mantenerlo, aggiornarlo o pulirlo. Anzi, si può arrivare a pensare che perdere meno tempo possibile per sbrigare questa formalità sia la scelta migliore. È un circolo vizioso: una cattiva qualità rende i dati meno utili, perché la loro analisi non porta a risultati affidabili ed "azionabili" per prendere decisioni corrette; se i dati non sono utili, non è giustificato investire per migliorarli, quindi la qualità peggiora ulteriormente. Questo, probabilmente, accade anche perché il formato scelto e il decentramento di pubblicazione imposti dalla legge non permettono un raffronto ed un'analisi immediata dei dati. Molte delle amministrazioni inadempienti, dunque, probabilmente non si rendono neppure conto di pubblicare dati di scarsa qualità.

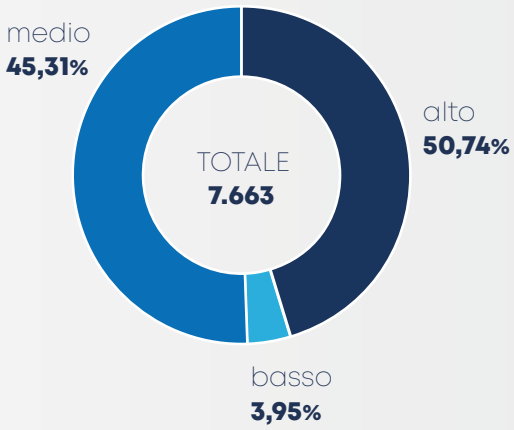
Di contro, non ci sentiamo di criticare la scelta di procedere ad una pubblicazione strutturata e decentrata di questi dati, poiché si tratta pur sempre di uno dei maggiori casi di successo dell'Open Data in Italia, in termini di informazioni disponibili per l'accesso ed il riutilizzo di chiunque, ad esempio per effettuare l'analisi che state leggendo! Oltre a questo, dati di cattiva qualità non dovrebbero essere presenti se fossero state seguite le chiare linee guida indicate da ANAC per la compilazione dei campi.[12]

GRAFICO 3. **ISTOGRAMMA DELL'INDICE DI QUALITÀ DEL COMUNE SU TUTTO IL PERIODO DI RIFERIMENTO**



12 - Linee guida ANAC per la compilazione dei contratti *https://www.google.com/url?q=https://www.anticorruzione.it/portal/rest/jcr/repository/collaboration/Digital%2520Assets/anacdocs/Servizi/ServiziOnline/AdempimentoLegge190/Specifiche%2520Tecniche%2520Legge%2520190%2520v1.2_finale.pdf&sa=D&ust=1574350161998000&usg=AFQjCNFAF-zejbJg_-2CKjZoGQDgCPYefQ*

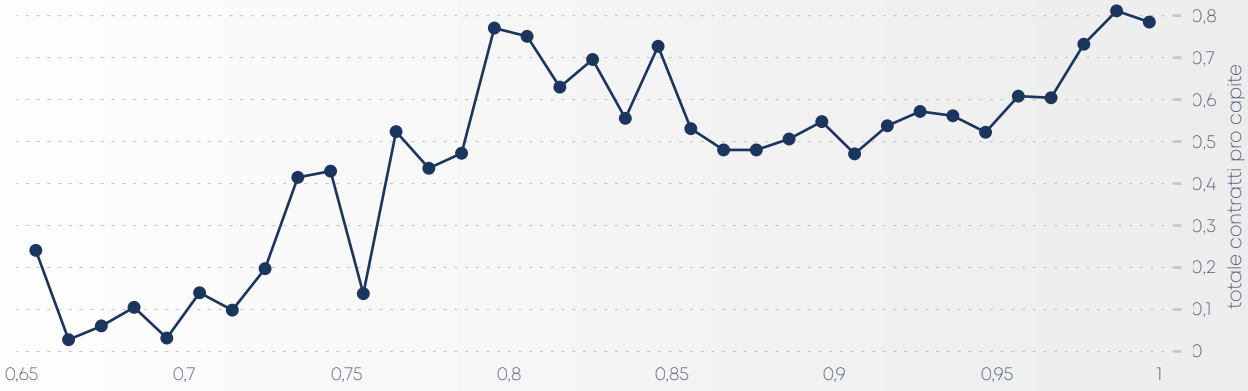
GRAFICO 4. **FASCE DELL'INDICE DI QUALITÀ DEL COMUNE**



Se si guarda lo stesso dato, ma aggregando per Comune (Grafico 3), definendo così un **Indice di qualità** del Comune si riproduce all'incirca lo stesso andamento che si ha per la qualità dei contratti: molte amministrazioni hanno trovato il giusto metodo per pubblicare senza fare quasi errori, mentre alcune hanno ancora difficoltà. È probabile che queste PA producano dati con poco supporto informatico sulla validazione dei campi o, più in generale, con metodologie poco affidabili, dando adito a più errori. Dopo aver normalizzato il numero di contratti sulla popolazione del Comune, per rendere questi ultimi confrontabili fra loro (numero di contratti pro capite), si è istogrammato questo valore rispetto all'Indice di qualità del Comune nel Grafico 5: si nota che quando la qualità è alta o medio/alta, il numero di contratti pro capite si mantiene abbastanza costante; quando però la qualità è bassa, allora anche il numero di contratti pro capite si abbassa, evidenziando una **correlazione fra una peggiore qualità e una tendenza a pubblicare di meno**. In altre parole, sembra che gli enti che pubblicano male lo facciano anche poco (in termini di contratti per abitante).

Con una visione temporale sui dati, la qualità complessiva in tutta Italia è leggermente scesa dal 2014 al 2018[12]: 0,95→0,93. Questo dato potrebbe essere influenzato dai campi data e importo liquidato che a volte vengono compilati ex-post, abbassando di conseguenza il ranking dell'anno corrente, delineando una sostanziale stabilità nel tempo.

GRAFICO 5. **ISTOGRAMMA DELL'INDICE DI QUALITÀ DEL COMUNE RISPETTO AL NUMERO DI CONTRATTI PRO CAPITE**



13 - Si sono scelti questi due anni perché sebbene la legge che obbliga le PA a pubblicare i dati in formato open sia del 2012, sono stati necessari un paio d'anni prima che si iniziasse a pubblicare in modo sistematico. Mentre si è preferito fare riferimento all'anno passato perché i dati del 2019, essendo in corso, sono incompleti.

Aggregando i contratti del 2018 per regione (Grafico 6) - **Indice di qualità regionale** - non si nota una divisione per zone, ma cali e picchi di qualità sparsi. Ciò detto, il Piemonte è all'ultimo posto di questa classifica, mentre l'Emilia Romagna e le Marche mostrano la qualità più alta. Mentre è interessante notare come sia in netto miglioramento parte del Sud e la Valle d'Aosta, e in peggioramento il Lazio, il Veneto e la Puglia.

L'ordine sparso con cui la qualità fluttua nelle varie zone d'italia porta ad escludere analisi semplicistiche: non ci sono divisioni Nord/Sud, né è chiaro se i piccoli centri abbiano più problemi a pubblicare i dati rispetto alle grandi amministrazioni. Infatti, la qualità dei dati è rimasta relativamente stabile nel periodo dal 2014 al 2018 per tutte le zone d'italia (Grafico 7) e non c'è una parte d'Italia che abbia un andamento decisamente peggiore o migliore rispetto alle altre; anche dividendo per fasce di popolazioni dei Comuni (Grafico 8), non si notano segmentazioni particolarmente accentuate (si noti l'asse verticale dove la variazione complessiva coinvolge un piccolo intervallo di valori).

Disaggregando l'indice complessivo di qualità nelle sue componenti si possono valutare le varie categorie di misurazione su ogni campo, come riportato nella Tabella 1.

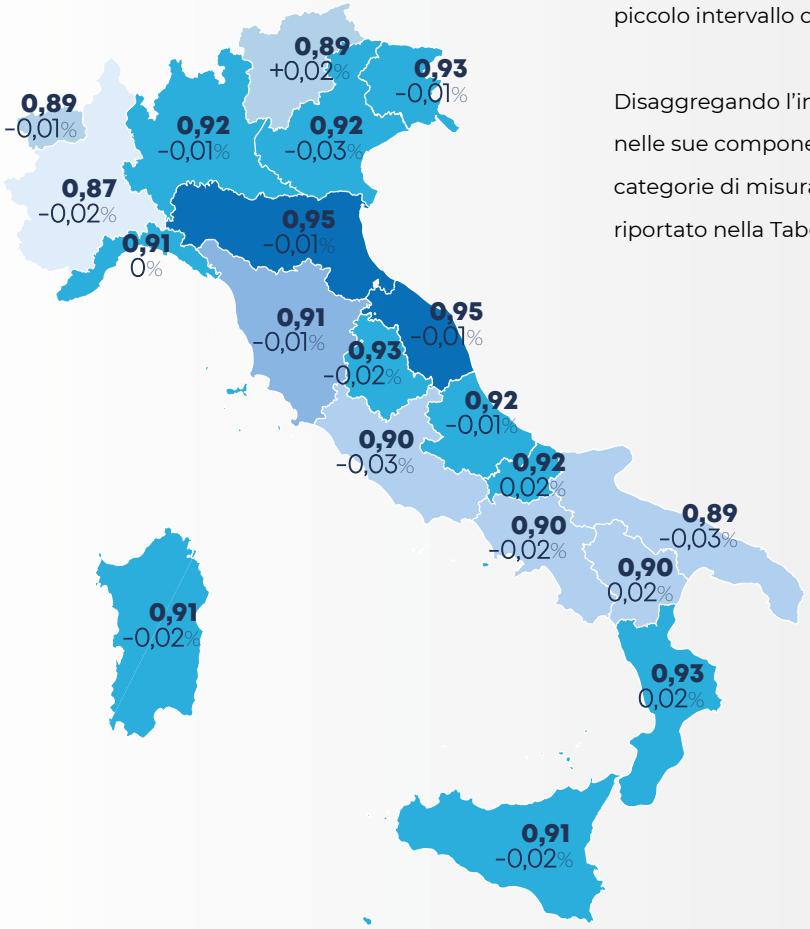


GRAFICO 6. **INDICE DI QUALITÀ REGIONALE NEL 2018, CON VARIAZIONE DAL 2014**

Campo	Completezza	Coerenza	Precisione
Importo	93.68%	98.27%	88.28%
Importo liquidato	85.77%	92.64%	86.14%
Data inizio	85.26%	93.21%	-
Data fine	77.30%	98.85%	-
CIG	92.47%	-	-
Oggetto	99.12%	-	-
Scelta contraente	92.05%	-	-
Identificativo della PA (p.iva o cf)	96.47%	-	-

TABELLA 1

Per ogni campo di un contratto è riportato se il criterio di misura non è applicabile (segnato con "-"), altrimenti è presente la percentuale con la quale il criterio di misura stesso è stato soddisfatto sul totale dei contratti. Per un raffronto con i valori nazionali di tutti gli enti pubblici italiani, consultare l'*Allegato B*.

La **completezza** è la categoria di misurazione che si può sempre applicare a tutto il tracciato di un contratto e che ha la maggior variabilità: ogni campo ha una sua peculiarità. Vi abbiamo dedicato un approfondimento ad-hoc nel capitolo successivo. Parlando invece di coerenza, nei casi in cui essa viene a mancare, si mina la credibilità di tutto il contratto: se ad esempio l'importo liquidato è 2 volte maggiore dell'importo concordato[15] è lecito mettere in dubbio che si sia prestata la giusta attenzione nella compilazione di tutti gli altri campi. La **precisione** è stata trascurata circa una volta su 10: fortunatamente è un tipo di errore che nel caso degli importi non compromette la corretta interpretazione dei dati nella maggior parte dei casi da parte degli esseri umani, ma può essere problematico in un ambiente informatico automatizzato (e potrebbe essere l'indice di una trascuratezza generale).

GRAFICO 7. **INDICE DI QUALITÀ DELLE ZONE D'ITALIA NEGLI ANNI 2012-2018**

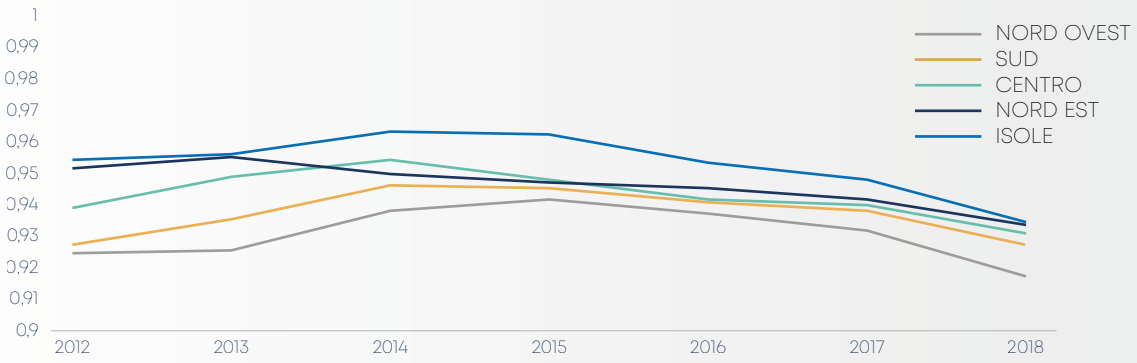
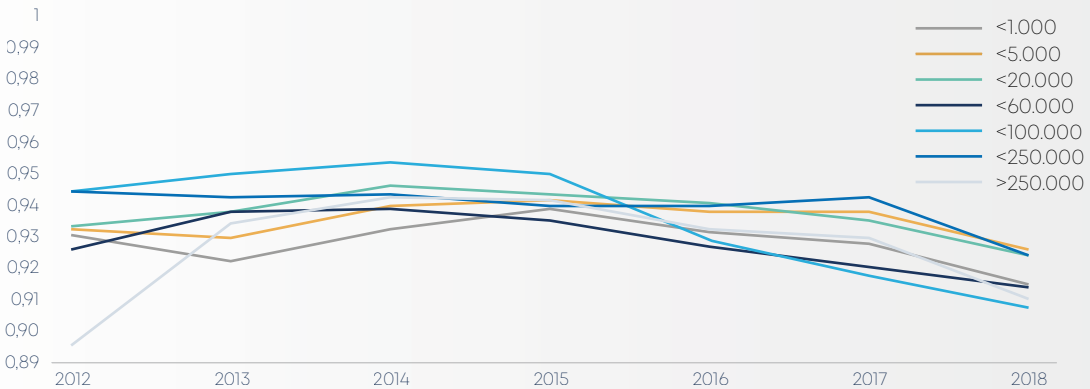


GRAFICO 8. **INDICE DI QUALITÀ DEI COMUNI PER FASCE DI POPOLAZIONE NEGLI ANNI 2012-2018**



14 - Casistica presente per l'1,08% dei contratti dei Comuni.

Completezza

> **La mancata compilazione dei campi inficia gravemente la possibilità di comprendere un contratto nel suo complesso**

Il **CIG** è fra i campi più compilati e questo è rassicurante perché un codice identificativo univoco è fra i pilastri portanti di un censimento affidabile. Esistono delle procedure specifiche esenti dalla richiesta del CIG, ciò può falsare parte delle statistiche e rendere impossibile la disambiguazione rispetto ad eventuali aggiornamenti[15]. Oltre a questo non si può distinguere chi non ha compilato il CIG perché era debitamente esente, da chi ha fatto un'omissione colpevole, se non con laboriose e complesse verifiche caso per caso di adesione del contratto a certi parametri.[16] In generale, tuttavia, soprattutto negli anni più recenti, i casi in cui un contratto può essere affidato senza un CIG valido dovrebbero essere assolutamente residuali.

Le date sono le informazioni meno presenti ed in particolare la **data di fine** è quella più assente in assoluto. Questo accade perché la maggior parte delle volte la PA deve pubblicare i dati di contratti che non sono ancora conclusi, o non ancora definiti in ogni parte, e la data di fine è l'informazione che rimane più spesso incerta. Tuttavia una data di fine presunta dovrebbe essere sempre nota e dunque l'eventuale incertezza sulla data di fine effettiva non scusa la mancata pubblicazione. Purtroppo, evidentemente, capita che la data di fine presunta non venga indicata, ma lasciata semplicemente in bianco. Ancor peggio, la data di fine lasciata inizialmente in bianco spesso non viene compilata neppure a consuntivo, sicché nel Grafico 10 si vede come gli anni più recenti siano leggermente più carenti sotto questo aspetto (coerentemente con l'idea che la data di fine a volte venga aggiunta in seguito), ma anche come in generale questo campo continui a restare meno compilato degli altri anche per il passato.

15 - In linea generale non si può affermare con certezza se due tracciati si riferiscano allo stesso contratto o no solo perché presentano gli stessi valori per gli stessi campi, nel caso in cui non abbiano un codice identificativo univoco. Con milioni di contratti questa non è una casistica né rara né trascurabile.

16 - Normativa sulla tracciabilità, punto A12 *https://www.anticorruzione.it/portal/rest/jcr/repository/collaboration/Digital%20Assets/PDF/FAQ.Tracci abilita.pdf*

L'**importo liquidato** invece viene aggiornato in modo un po' più sistematico, perché si può vedere come sia il dato più mancante negli anni più recenti. Questo è dovuto al fatto che non si percepisce la differenza fra il "dato zero" e il "dato mancante": in genere la stazione appaltante non ha ancora erogato alcun importo alla pubblicazione del contratto, dunque invece che indicare che l'importo liquidato è uguale a zero, il campo non viene compilato. Questa sottile differenza fa sì che gli importi liquidati zero siano indistinguibili dagli importi liquidati che sarebbero non zero ma non compilati, creando incertezza nell'interpretazione.

GRAFICO 9. **FASCE DELL'INDICE DI COMPLETEZZA DEI COMUNI**

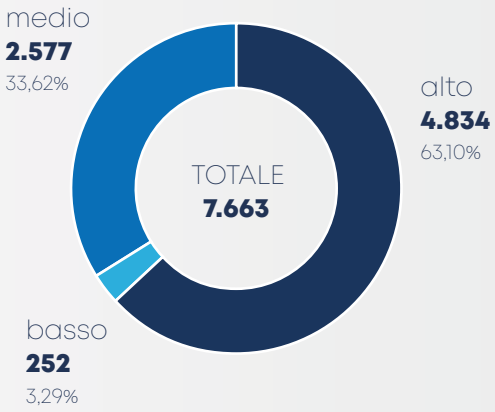


GRAFICO 10. **ANDAMENTO NEGLI ANNI DELLA PERCENTUALE DI COMPLETEZZA DEI SINGOLI CAMPI**

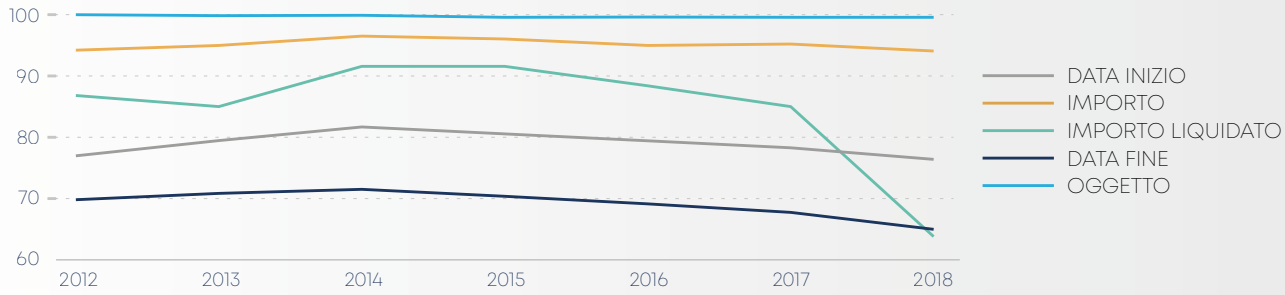
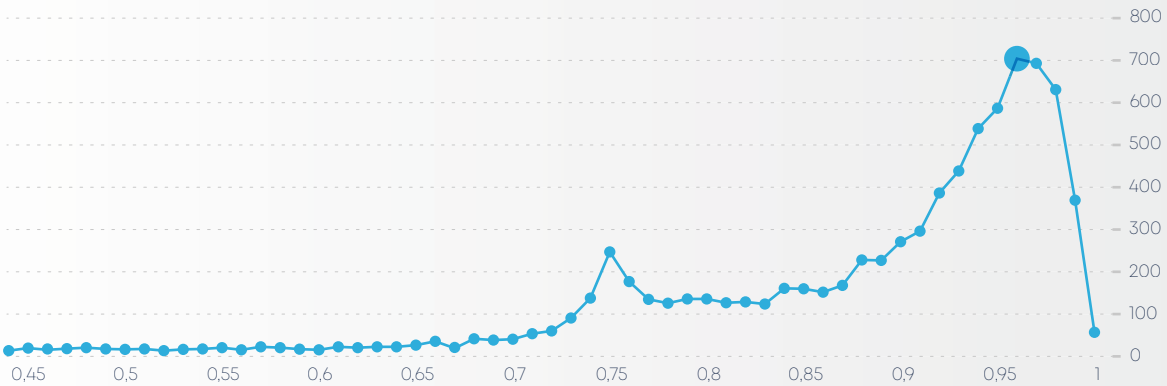


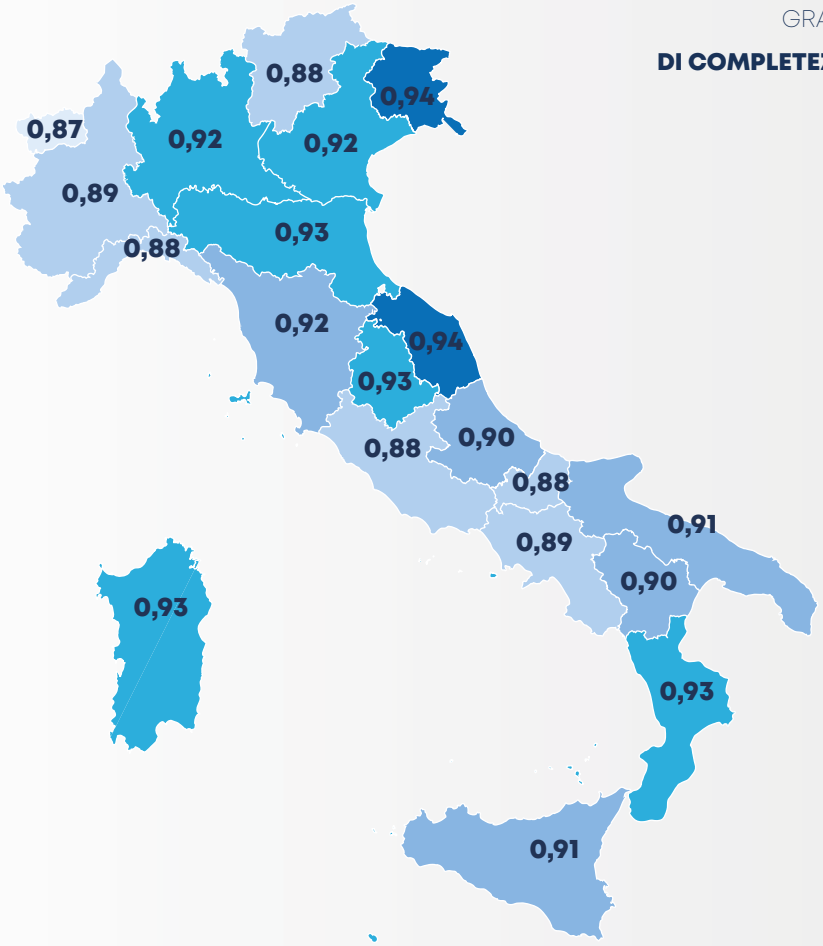
GRAFICO 11. **ISTOGRAMMA DELL'INDICE DI COMPLETEZZA DEI COMUNI**



L'**oggetto** è il dato più presente e questo aiuta molto nella disambiguazione dei contratti, anche in mancanza del CIG. Ad ogni modo la presenza di un testo descrittivo non è sufficiente ad assicurarne l'utilità effettiva. Sono infatti molti i casi in cui questi non sono significativi: un oggetto generico (come "Acquisto" o "Polizza") non permette un reale riconoscimento dell'acquisto effettivamente avvenuto, né la distinzione da contratti dallo stesso titolo. In media l'oggetto di un contratto è composto da 5 parole, in genere sufficienti a renderlo comprensibile. Lo studio degli oggetti è un tema molto ricco che richiede tecniche di analisi semantica automatica del linguaggio naturale, che potrebbero essere tema di approfondimenti futuri.

I diversi campi sono pesati rispetto alla loro importanza nella comprensione del contratto nell'Indice di completezza, che riassume in un valore quanto è completo il contratto. Si può vedere come la media di questo indice nelle regioni (Grafico 12) abbia una forchetta di valori che rimane comunque alta fra 0.87 e 0.94: non c'è nessuna regione, presa complessivamente, che sia gravemente indietro su questo parametro. Ad ogni modo esplorando il valore dell'indice medio di completezza prendendo singolarmente i Comuni, si può vedere dal Grafico 9 e 11 che non sono pochi i Comuni (252) che in media pubblicano contratti semi vuoti (fascia bassa), a cui mancano la maggior parte delle informazioni più importanti.

GRAFICO 12. **INDICE DI COMPLETEZZA REGIONALE**



Le 12 cittá maggiori

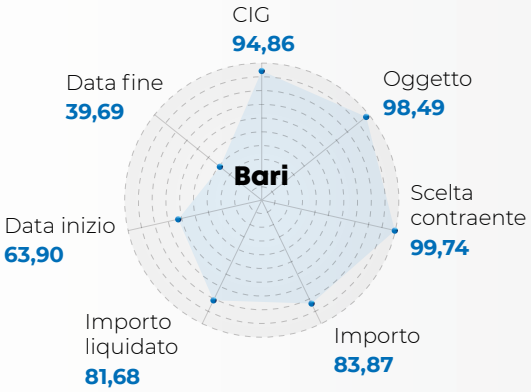
Per entrare ulteriormente nel dettaglio della completezza, e per mostrare quanto possa essere variegato il quadro a riguardo, abbiamo scelto a titolo di esempio le 12 maggiori città italiane per popolazione.

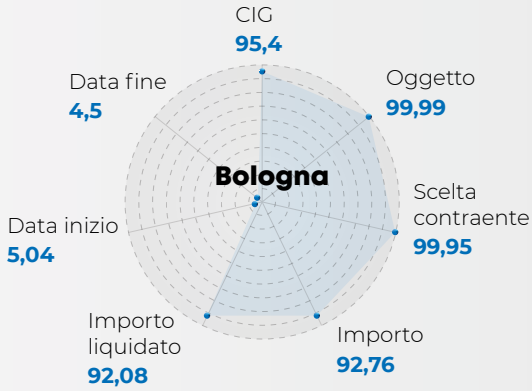
Analizzando il numero medio di contratti pro capite si può notare una blanda correlazione fra il numero di contratti stipulati e il numero di abitanti di un Comune. All'estremo inferiore si notano le anomalie del Comune di Palermo e del Comune di Catania e in parte anche il Comune di Torino in quanto valori troppo bassi: se in alcuni casi opportune strategie di aggregazione o l'utilizzo delle società in-house potrebbero giustificare una certa varianza, è anche possibile che alcuni enti non abbiano pubblicato tutti i loro dati. Viceversa il Comune di Verona e il Comune di Firenze hanno un valore pro capite molto superiore alla media: per entrambi si nota dal portale *ContrattiPubblici.org* che nel periodo di questo studio vengono stipulati molti contratti riguardanti lavori di edilizia pubblica. In ogni caso è complesso stimare la copertura dei contratti pubblicati rispetto al totale di quelli effettivamente stipulati, poiché mancano dati ufficiali di sintesi sul numero di CIG (e dunque contratti) effettivamente utilizzati. Anche gli incroci coi dati di bilancio o delle uscite di cassa degli enti sono tutt'altro che banali, ma potrebbero fornire prospettive interessanti e rappresentare l'oggetto di futuri approfondimenti.

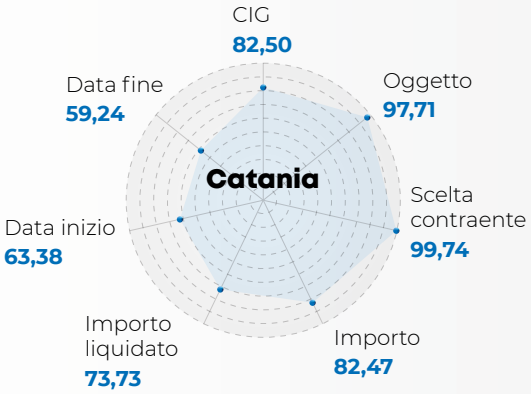
Considerando la completezza complessiva dei contratti la forchetta di variazione è piuttosto grande, dovuta in genere a campi singoli che vengono sistematicamente omessi. Infatti, guardando il dettaglio dei grafici dei singoli Comuni si nota che il Comune di Palermo non pubblica le date e gli importi e il Comune di Bologna non pubblica le date, ed è per questo che occupano

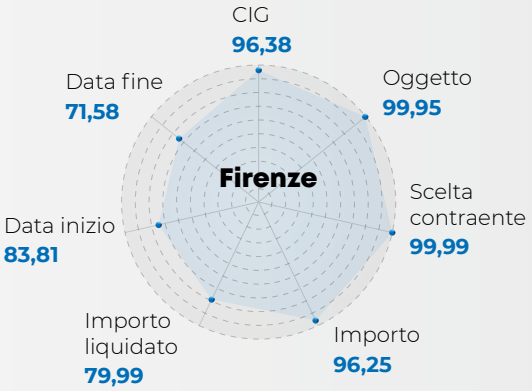
Comune	Numero contratti	Popolazione	Numero contratti ogni 1000 persone	Percentuale media completezza campi
Venezia	7.186	261.362	27,5	94,01%
Verona	10.354	252.520	41,0	93,94%
Torino	4.256	872.367	4,9	92,41%
Firenze	19.787	358.079	55,3	89,71%
Milano	21.494	1.242.123	17,3	88,90%
Roma	33.545	2.617.175	12,8	85,93%
Napoli	3.732	962.003	3,9	85,32%
Bari	4.866	315.933	15,4	80,32%
Catania	1.976	293.902	6,7	79,82%
Genova	15.347	586.180	26,2	78,95%
Bologna	11.065	371.337	29,8	69,96%
Palermo	549	657.561	0,8	40,85%

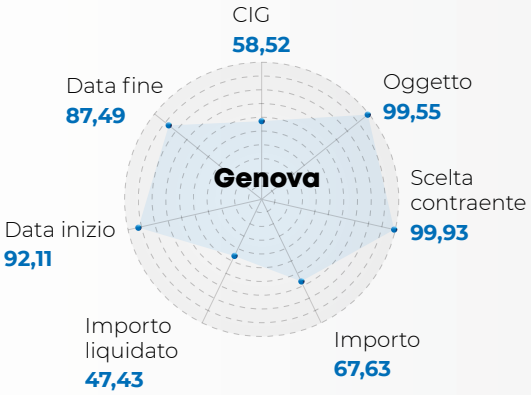
gli ultimi posti nella classifica. Sono preceduti dal Comune di Genova che pubblica l'importo liquidato meno della metà delle volte, e l'importo concordato solo in due casi su tre. È sempre poi il Comune di Palermo che pubblica l'oggetto con minor frequenza fra tutte le città, infatti è mancante circa una volta su 10. Ad ogni modo la scelta del contraente e l'oggetto sono le informazioni più presenti. Complessivamente è Venezia che pubblica i contratti più completi, anche se insieme a Catania e Palermo compila il CIG 3 volte su 4. Sono però Genova e Napoli a pubblicare meno il CIG: la prima delle due città lo compila il 58% delle volte sul totale, la seconda nel 40% dei casi.

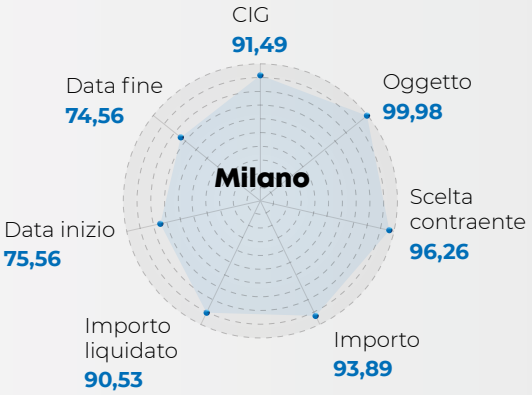


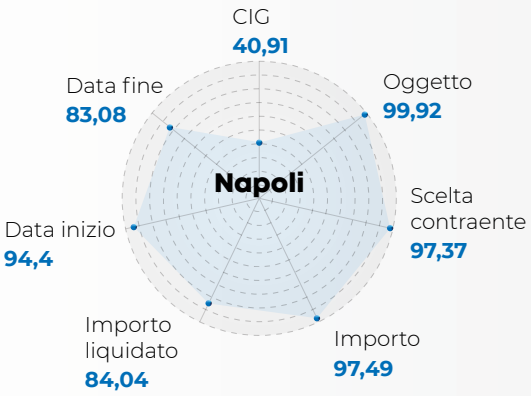


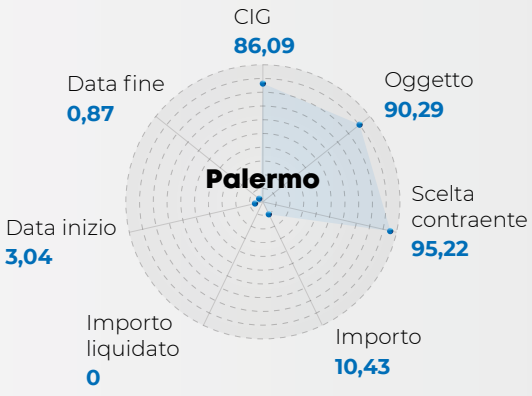


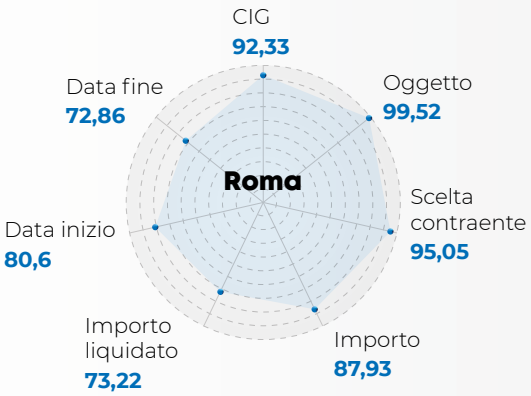


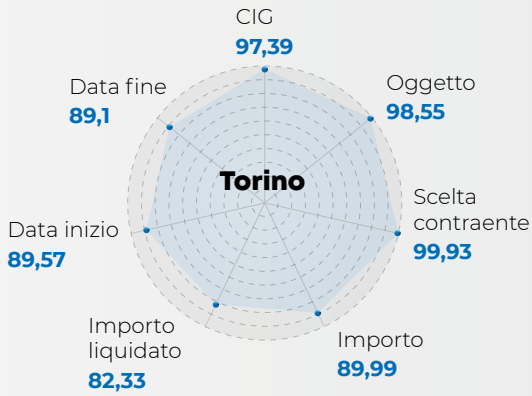


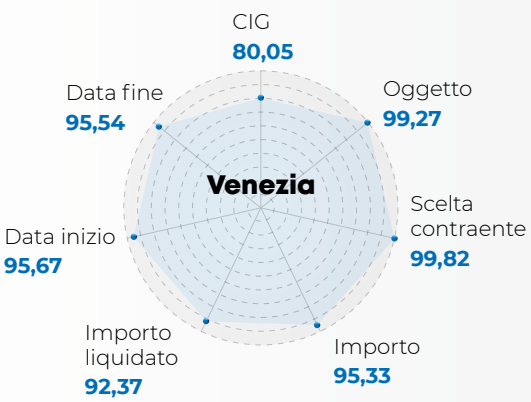


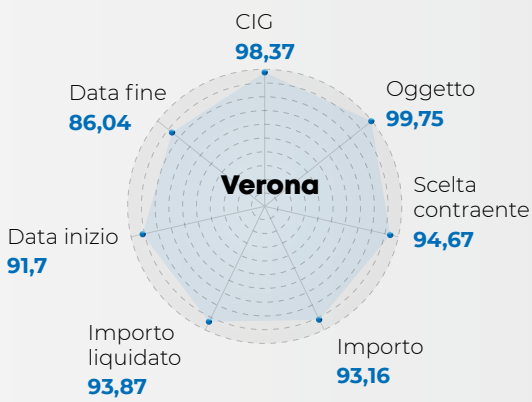

CIG
40,91
Oggetto
99,92
Scelta contraente
97,37
Importo
97,49
Importo liquidato
84,04
Data inizio
94,4
Data fine
83,08
Napoli


CIG
86,09
Oggetto
90,29
Scelta contraente
95,22
Importo
10,43
Importo liquidato
0
Data inizio
3,04
Data fine
0,87
Palermo


CIG
92,33
Oggetto
99,52
Scelta contraente
95,05
Importo
87,93
Importo liquidato
73,22
Data inizio
80,6
Data fine
72,86
Roma


CIG
97,39
Oggetto
98,55
Scelta contraente
99,93
Importo
89,99
Importo liquidato
82,33
Data inizio
89,57
Data fine
89,1
Torino


CIG
80,05
Oggetto
99,27
Scelta contraente
99,82
Importo
95,33
Importo liquidato
92,37
Data inizio
95,67
Data fine
95,54
Venezia


CIG
98,37
Oggetto
99,75
Scelta contraente
94,67
Importo
93,16
Importo liquidato
93,87
Data inizio
91,7
Data fine
86,04
Verona

Conclusione

L'Italia è un'eccellenza a livello internazionale per quel che riguarda la pubblicazione dei dati sui contratti pubblici. Tra l'altro, nel 2018 la Banca Dati Nazionale dei Contratti Pubblici (BDNCP) dell'Autorità Nazionale Anticorruzione si è classificata al primo posto nella competizione Better Governance through Procurement Digitalization 2018[17] – categoria National Contract Register, alla quale hanno partecipato i principali registri dei contratti pubblici europei.
I dati sui contratti pubblici italiani sono utili al riutilizzo e favoriscono innovazione, trasparenza ed imprenditorialità: Synapta può testimoniarlo avendone fatto una delle principali (anche se non l'unica) fonte dati della propria piattaforma di ricerca e business intelligence sul public procurement, *ContrattiPubblici.org*.
Inoltre, la pubblicazione di Dati Aperti per combattere e soprattutto prevenire la corruzione e le cattive pratiche amministrative nasce da un bel caso di hacking istituzionale.

Nell'agosto del 2012, un gruppo di funzionari e tecnici ANAC si è inventato, per rispettare le scadenze imposte dalla Legge Anticorruzione, un meccanismo di pubblicazione decentrato (sui siti di ogni PA) e basato su standard esistenti, che avrebbe dovuto essere temporaneo, ma ha retto meglio di altre iniziative al passare del tempo.[18]
Inoltre, dal nostro studio emerge che la maggior parte delle PA pubblicano i dati previsti dalla normativa e la qualità è generalmente alta.

Insomma, tutto bene?

Quelle riassunte sopra sono tutte buone notizie. E sono vere. Tuttavia, il nostro report mostra e noi riteniamo sia grave che un numero non trascurabile di PA abbia potuto continuare per anni a pubblicare dati di bassa qualità. Per bassa qualità intendiamo dati pubblicati così male da vanificare del tutto la possibilità di comprendere a quale contratto pubblico quei dati si riferiscano.

17 - *https://www.anticorruzione.it/portal/rest/jcr/repository/collaboration/Digital%20Assets/anacdocs/Home*

18 - Gli XML della L.190/2012 vanno infatti pubblicati sul sito di ciascuna PA, in indirizzi accessibili pubblicamente, e vanno segnalati via PEC all'ANAC. Nel 2012, la PEC era uno standard emergente per le comunicazioni con la PA e tra PA, mentre lo standard XML era molto di moda all'interno della comunità tecnica, come formato abbastanza leggibile dagli esseri umani, ma processabile automaticamente dalle macchine, ed in particolare come formato per garantire l'interoperabilità e lo scambio di dati. E così, rapidamente e con costi contenuti, si è creato uno dei più grossi dataset distribuiti e machine readable nel nostro paese.

Questi dati riguardano adempimenti di legge per finalità anticorruzione e la mancata o errata pubblicazione dovrebbe essere tempestivamente individuata e conseguentemente corretta.
Tra l'altro, mentre per strumenti di business intelligence come *ContrattiPubblici.org* è già molto utile disporre della stragrande maggioranza dei dati riguardanti i contratti pubblici, per prevenire la corruzione dovremmo avere la ragionevole certezza che tutti i dati su tutti i contratti fossero online. Invece, purtroppo, sarebbe oggi molto facile omettere la pubblicazione di una manciata di dati, magari proprio quelli che riguardano casi "da nascondere". E non ci sono strumenti efficaci per individuare ad esempio queste omissioni, ovvero errori fatti con dolo, nel mezzo dei tanti errori fatti per trascuratezza.
Per tornare a dare buone notizie, d'altro canto, possiamo aggiungere che soggetti come Synapta - ma altrettanto potrebbe fare qualunque privato cittadino, associazione o altra azienda - hanno a disposizione gli strumenti dell'accesso civico per richiedere la pubblicazione dei dati mancanti o la correzione di quelli scorretti. E questo lo facciamo spesso e sicuramente lo racconteremo più in dettaglio in appositi report. Anche sulla base delle risposte spesso positive e collaborative che abbiamo ricevuto dagli enti a fronte di richieste di accesso civico molto specifiche, tuttavia, riteniamo che si possa e si debba fare molto di più per rendere gli enti consapevoli della qualità - a volte bassa - dei dati che pubblicano. Molti enti pubblicano male senza accorgersene davvero e ricevono poco feedback (o feedback di tipo troppo formale) in merito dagli enti preposti al controllo.
Ci pare anche giusto sottolineare in queste conclusioni che non è solo con la "minaccia" di una richiesta di accesso civico che vorremmo convincere gli enti a pubblicare i loro dati.
Riteniamo infatti che strumenti di riutilizzo dei dati, come il nostro portale *ContrattiPubblici.org*, siano parte di una strategia che incoraggia a pubblicare tempestivamente dati di buona qualità.
Chi pubblica bene i propri dati, infatti, riceve come effetto collaterale positivo la possibilità di analizzarli con strumenti come il nostro portale.
Ed in generale migliora probabilmente la qualità dei suoi sistemi informativi.

In conclusione, i dati relativi ai contratti della pubblica amministrazione rappresentano uno dei più grandi ed importanti dataset open in Italia. E sono oggettivamente una best practice nel panorama dei Dati Aperti italiani, essendo online a disposizione di chiunque voglia (con una certa fatica, vi assicuriamo!) raccoglierli. Ma sono anche dati che hanno ancora tutti i problemi di qualità dei dati descritti in questo report, sui quali è necessario investire anni uomo di lavoro, come ha fatto Synapta, per poterne trarre strumenti di analisi e business intelligence affidabili. C'è dunque ancora molto lavoro da fare e chi volesse farlo assieme a noi è benvenuto se ha voglia di farcelo sapere a **contrattipubblici@synapta.it**.

DATA QUALITY

Appendice

PROVENIENZA DEI DATI

Abbiamo raccolto gli Open Data
sugli acquisti di beni e servizi e l'affidamento
di lavori di tutta la PA italiana

04

L'articolo 1 comma 32 della legge 190 / 2012[19] prevede per le pubbliche amministrazioni che queste «*sono in ogni caso tenute a pubblicare nei propri siti web istituzionali: la struttura proponente; l'oggetto del bando; l'elenco degli operatori invitati a presentare offerte; l'aggiudicatario; l'importo di aggiudicazione; i tempi di completamento dell'opera, servizio o fornitura; l'importo delle somme liquidate*», e che poi queste informazioni devono essere «*pubblicate in tabelle riassuntive rese liberamente scaricabili in un formato digitale standard che consenta di analizzare e rielaborare, anche a fini statistici, i dati informatici*».

Questi dati si presentano come file XML[20] su direttiva dell'Autorità Nazionale Anticorruzione (ANAC). I dati sono poi pubblicati nelle sezioni "*Amministrazione Trasparente*" nel sito ufficiale di ogni pubblica amministrazione[21]. Data la **licenza aperta**[22] che vige su questi dati, è possibile scaricarli per la consultazione e analisi statistiche.

Con un'intensa e costante attività di *web crawling* e *web scraping* - ovvero l'utilizzo di software specifico rispettivamente per scaricare in modo metodico tutte le pagine di un sito e per l'estrazione di dati da un sito web - delle sezioni "Amministrazione Trasparente" dei siti delle PA, è stato possibile raccogliere una grande quantità di dati sugli acquisti. Poi, dopo essere stati scaricati, sono stati strutturati, puliti, indicizzati, interconnessi e infine resi disponibili per la consultazione e l'analisi nel portale online *Contrattipubblici.org*.

19 - Legge, 6 novembre 2012, n. 190, Gazzetta ufficiale | *http://www.gazzettaufficiale.it/eli/id/2012/11/13/012G0213/sg*

20 - *http://dati.anticorruzione.it/schema/TypesL190.xsd*
http://dati.anticorruzione.it/schema/datasetAppaltiL190.xsd
http://dati.anticorruzione.it/schema/datasetIndiceAppaltiL190.xsd

21 - Qui a titolo di esempio il link alla relativa sezione del Comune di Torino | *http://www.comune.torino.it/amministrazionetrasparente/bandi-gara/procedure-tabellare/index.shtml*

22 - Licenza CC BY 3.0.

Allegato A

Su ognuno dei campi dei contratti sono stati eseguiti i controlli qui elencati:

- **Oggetto**
  - Completezza
    - Dato mancante
- **CIG**
  - Completezza
    - Dato mancante
- **Scelta contraente**
  - Completezza
    - Dato mancante
- **Importo pattuito e importo liquidato**
  - Completezza
    - Dato mancante
    - Scritto “NaN”, “annullato”, “gara deserta”, “chiusa d’ufficio” ecc.
    - Scritto “N.A.”, “nd” ecc.
    - Scritto “importo”
  - Coerenza
    - L’importo liquidato è più del doppio dell’importo pattuito
    - C’è l’importo liquidato ma non l’importo pattuito
    - C’è una data di inizio ma non c’è un importo pattuito
  - Precisione
    - Troppe poche cifre dopo la virgola “100.1” o “100” invece che “100.12”
    - Troppe cifre dopo la virgola “1000.12000000”
- **Data di fine e di inizio**
  - Completezza
    - Dato mancante
    - Zeri o caratteri speciali
    - Formati vari per il dato mancante come “xxxx-xx-xx” , “aaaa-mm-gg” o “gg-mm-aaaa”
  - Coerenza
    - C’è una data di inizio ma non c’è un importo pattuito
    - C’è una data di fine ma non una di inizio
    - La data di fine è precedente a quella di inizio

Allegato B

Tabella riassuntiva con le percentuali con le quali un criterio di misura è stato soddisfatto sul totale dei contratti, considerando tutti gli enti e non solo i Comuni.

Campo	Completezza	Coerenza	Precisione
Importo	93.68%	98.27%	88.28%
Importo liquidato	85.77%	92.64%	86.14%
Data inizio	85.26%	93.21%	-
Data fine	77.30%	98.85%	-
CIG	92.47%	-	-
Oggetto	99.12%	-	-
Scelta contraente	92.05%	-	-
Identificativo della PA (p.iva o cf)	96.47%	-	-

PROGETTO DI:
SYNAPTA
CONTRATTIPUBBLICI.ORG

AUTORI:
SOFIA BENEDETTA ROSATI
ALESSIO MELANDRI
FEDERICO MORANDO

PROGETTO GRAFICO DI:
BIANCO TANGERINE