BENE COMUNE

Dati…di qualità

La Rivista, Numeri, Vero o falso?

 Giuseppe Notarstefano | 30 Gennaio 2017

Dati, mole immense di dati che circolano attraverso il web, dati enormi (Big Data), dati accessibili a tutti (Open Data). La modernità ci aveva spiegato che le teorie vanno provate, empiricamente. Ma cosa c'è dietro i dati? Come si producono? Quando sono attendibili? Come leggergli?

Dati, mole immense di dati che circolano attraverso il web, dati enormi (Big Data), dati accessibili a tutti (Open Data)… un autentico diluvio di dati! La modernità ci aveva spiegato che le teorie vanno provate, empiricamente. Qualcuno ci aveva detto che la loro validità va falsificata, ma sempre empiricamente. Empiricamente significa sottoporre un'ipotesi (teorica, derivante dalla speculazione o dall'esperienza o dalla osservazione riflessiva della realtà da parte dello studioso) alla "prova dei fatti", verificare le teorie per renderle più "oggettive". Ciò diventa particolarmente vero quando tali fatti sono relativi a ciò che accade della vita sociale, in quella delle persone così come in quella delle organizzazioni e delle istituzioni, quando ciò la conoscenza è soprattutto una ricostruzione aggregata. I fatti possono essere interessanti in quanto singoli casi da osservare e comparare nella loro specificità ed originalità, diversamente la nostra osservazione si rivolge ai fenomeni aggregati, quindi i collettivi.

La scienza, e ancor prima la pratica statistica, nasce proprio per questo: per fornirci una rappresentazione, sufficientemente chiara e comprensibile, di ciò che accade ad una serie di individualità (le unità di osservazione) considerate insieme, pertanto capaci di fornire uno sguardo di insieme sul mondo reale. I dati si possono pertanto dire statistici se sono riferiti a fenomeni aggregati, se sintetizzano informazioni relative a collettivi, in breve se ci aiutano a ragionare in termini generali e complessi. Ma ottenere tali tipologie di dati non è affatto scontato, non si tratta di una raccolta "passiva" o di una inoperosa ricezione di osservazioni e misure che esistono già nella realtà. Affatto!

I dati sono l'esito di un processo di produzione che attiva un dispositivo logico-tecnico che prevede una fase di ideazione-progettazione (il disegno che definisce il chi il cosa il come il quando e il dove della misurazione), una fase di estrazione dell'informazione attraverso l'osservazione o la consultazione delle unità di osservazioni attraverso opportuni strumenti di misurazione (schede di intervista o questionari, ma oggi anche rilevazioni ottiche attraverso

fotocellule o sistemi d rilevazione GPS…), quindi la raccolta e la costruzione di supporti informativi sintetici (finalmente i dati!) che consentono l'interpretazione anche ai soggetti che nulla sanno di tutto ciò che li precede. L'utilizzatore dei dati, quando non ha potuto aver sotto controllo tutto il processo di produzione, può essere rassicurato nel trattamento degli stessi dalla presenza di due requisiti che devono accompagnare i dati stessi: l'esistenza di meta-dati e la qualità dei dati.

Quando leggiamo una tabella o un grafico (che sono le prime forme più elementari di elaborazione dei dati, non già i dati!!!) *facciamo sempre attenzione all'oggetto che troviamo nel titolo*, alla unità di misura con cui vengono espressi i dati, alla scala del grafico nel sistema di assi cartesiani, alla fonte dell'indagine che ci informa sul soggetto che ha prodotto il dato: l'insieme di questi informazioni, insieme a tutte quelle che le precedono come le definizioni, le classificazioni e le tecniche costituiscono la preziosa meta-informazione.

Particolarmente nell'ambito sociale per misurare un fenomeno, supponiamo la disoccupazione, lo devo definire devo metter a fuoco il concetto, esplicitarlo, "ritagliarlo" attraverso una definizione che mi premette di rilevarlo, identificare le tipologie di unità che devo rilevare sempre dentro il famoso dispositivo logico-tecnico che di concretizza nell'indagine campionaria o nel censimento generale. Esiste una ricca letteratura in questo campo della cosiddetta "metodologia della ricerca" che riguarda particolarmente le scienze sociali e che è l'indispensabile (e talvolta supponente) compagna di strada della metodologia statistica. Quest'ultima poi ci fornisce gli strumenti (tipicamente matematici) che ci legittimano nell'interpretazione corretta dei dati.

Innanzitutto per aggregare devo avere un popolazione di oggetti (unità statistiche) tra loro omogenei, comparabili "sommabili". Ciò è garantito dall'ipotesi dell'indipendenza in senso statistico che viene verificata rispetto alle diverse metriche e tipologie di variabili e che è alla base anche degli schemi di selezione casuale e probabilistica della popolazione.

La significatività dell'informazione è connessa alla rappresentatività del campione che costituisce il supporto su cui poi vengono costruite le sintesi statistiche. L'altro elemento che condiziona la significatività è dato dall'esistenza di una bassa o elevata variabilità, proprietà che si osserva empiricamente e che definisca alcuni importanti requisiti delle sintesi statistiche.

Parlando di significatività, accuratezza e variabilità *dobbiamo introdurre l'altro requisito che riguarda la qualità dei dati.* Pensando ai dati come "prodotti", gli studiosi così come i soggetti produttori di statistiche hanno importato dal mondo della produzione dei beni e dei servizi, il criterio della Qualità. Esso costituisce l'orizzonte ma definisce anche le coordinate, e dunque i principi e criteri, con cui devono essere "prodotti" i dati. Il discorso potrebbe farsi complicato, ci limiteremo pertanto a dire che i dati sono di buona qualità se

sono accurati ossia aderenti al fenomeno che intendono rappresentare, se sono tempestivi, se sono comparabili nel tempo e nello spazio, se sono chiari e si possono trattare ed utilizzare senza particolari "barriere".

Poiché produrre statistiche è un processo che richiede competenze e risorse esso è affidato, o meglio condensato in una delle funzioni dello Stato, che sin dalla costituzione degli stati nazionali ma particolarmente nelle moderne democrazie, ha promosso la formazione di sistemi di statistica ufficiali. Accanto ad essi ci sono ovviamente i dati che vengono prodotti, in genere in forma sperimentale, dagli studiosi e dai centri di ricerca particolarmente universitari.

L'irrompere di fattori problematici come l'elevato costo *relativo alla progettazione e organizzazione delle indagini statistiche*, la cui gran parte è determinata nei "piani statistici" approvati dai parlamenti e posti in essere dagli istituti centrali di statistica, ma anche la necessità di ridurre la pressione o "carico" statistico su alcuni soggetti che hanno l'obbligo di sottoporsi alle rilevazioni della statistica ufficiale (pensiamo ai censimenti della popolazione e delle abitazioni!) unito al sorgere di nuove opportunità come quella offerta dalla straordinaria diffusione delle nuove tecnologie nel campo della comunicazione e dell'informazione, ha suggerito a molti studiosi ed istituzioni di cominciare a produrre sempre meno indagini e ad utilizzare, trasformandoli opportunamente, i dati amministrativi o altri dati personali (magari delicatamente estratti dai soggetti grazie alle regole che vengono definite dalle normative di tutela e garanzia dell'informazione soprattutto personale che oggigiorno esistono in tutti gli ordinamenti).

Ciò ha aperto una voragine nel mondo della statistica, all'interno della quale in tanti si stanno tuffando (con intenti più o meno speculativi!) per raccogliere, gestire ed "offrire" spesso a titolo oneroso questa mole di grandi dati (Big Data) che giacciono nelle miniere nascoste dell'informazione virtuale.

C'è anche una questione relativa alla ricerca empirica che andrebbe considerata, ma che io enuncio molto sinteticamente: avere molti dati e poter disporre di supporti statistici ampi esonera spesso gli studiosi alla ricerca di teorie generali e di modelli idonei a prevedere e comprendere maggiormente il mondo reale (per orientarlo normativamente e politicamente); ci si limita sempre più ad adattarsi alle tendenze, rassegnandosi a comprensioni sempre più locali e sempre più parziali.

Ci sono anche diversi aspetti positivi e alcuni aspetti problematici che vorrei brevemente accennare, rinviando per un approfondimento alla lettura di un altro lavoro (Cfr. Giuseppe Notarstefano, La sfida della realtà: una nuova statistica "civile", La Società – n.4 / 2015).

Il primo è il tema del rapporto tra statistica e democrazia, un legame originario e che costituisce uno dei temi su cui occorrerebbe un maggiore monitoraggio e controllo da parte dei cittadini: ritengo in tal senso molto positivo l'impegno di tutti quei cittadini e i movimenti che lavorano con competenza e passione per ottenere trasparenza attraverso gli Open Data, soprattutto da parte della pubblica amministrazione, ma non solo.

Il secondo è più delicato e riguarda la fruizione dell'informazione, particolarmente statistica: il problema oggi non è relativo alla carenza di dati, ma piuttosto al loro eccesso. Diventano rilevanti i criteri per la selezione e gli elementi per la fruizione e lettura. In tal senso occorre un vasto programma informativo che aiuti e formi i cittadini, sin dagli anni della scuola primaria, a maneggiare con competenza e scioltezza i dati statistici, sviluppando maggiormente tanto l'attitudine al dato e alla sua rappresentazione matematica, ciò che gli inglesi chiamano "numeracy", che la propensione ad utilizzare maggiormente i dati nella formazione dei giudizi e delle valutazioni delle politiche pubbliche e delle azioni sociali.

"There are three kinds of lies: lies, damned lies, and statistics" affermava il premier britannico Benjamin Disraeli. Forse si apre un'era in cui dovremo dire "*There are three kinds of statistics: statistics, good statistics, and big data*".