METODOLOGIE FISICHE PER LE SCIENZE UMANE
(Prof. Paolo Rossi – Anno Accademico 2009/10)

1. Analisi qualitativa e analisi quantitativa

In qualunque disciplina, indipendentemente dal fatto che essa abbia per oggetto fenomeni naturali o fenomeni storico-sociali, accanto ad aspetti che possono essere analizzati esclusivamente mediante criteri di tipo qualitativo ne esistono altri che ammettono anche una trattazione quantitativa.
Esiste una consolidata tendenza a stabilire una sorta di gerarchia tra questi due tipi di analisi, e per di più la gerarchia, quasi paradossalmente, sembra invertirsi quando si passa dal contesto delle scienze "naturali" (nelle quali l'analisi qualitativa rischia di essere talvolta considerata priva di autentica dignità scientifica, o nella migliore delle ipotesi appare come una sorta di volgarizzazione di ciò che, più rigorosamente, dovrebbe essere enunciato in modo formalizzato e suffragato da un adeguato corredo di dati quantitativi) a quello delle scienze "umane", nelle quali spesso l'analisi quantitativa è vista con sospetto come una forma di indebita riduzione della varietà del reale a pochi, e pertanto comunque inadeguati, parametri misurabili. Più avanti torneremo con qualche dettaglio sulle critiche "filosofiche" che possono essere mosse a un'analisi quantitativa nel contesto delle scienze umane, cercando di prenderle in esame con la dovuta attenzione.
In generale tuttavia l'analisi qualitativa e quella quantitativa forniscono chiavi di lettura differenti, e quasi sempre complementari, di fenomeni complessi e pertanto suscettibili di interpretazioni anche fortemente dipendenti dal punto di vista assunto dall'osservatore. Dalla sintesi dei risultati qualitativi e quantitativi si può giungere, nella maggior parte dei casi, una miglior comprensione del fenomeno o del processo che si intende studiare, indipendentemente dalla natura dello stesso.

Vorrei considerare in dettaglio il caso di due discipline tra loro molto diverse, ma accomunate dal fatto di trovarsi entrambe, pur su versanti opposti, alla frontiera tra le scienze naturali e quelle sociali: mi riferisco alla biologia e all'economia.
Nel caso della biologia, non v'è alcun dubbio che, pur trattandosi di una scienza "naturale", essa non possa prescindere da un uso massiccio dell'analisi qualitativa: basti pensare a tutta la biologia descrittiva, su cui da sempre si basa il sistema classificatorio dei generi e delle specie, all'ecologia, all'etologia (scienza del comportamento animale), alla stessa antropologia fisica. Ma esistono altri aspetti nei quali non soltanto l'aspetto quantitativo dell'analisi diventa dominante, ma addirittura non si può più prescindere dall'uso di formalismi anche sofisticati dal punto di vista matematico: pensiamo alla biologia molecolare, che richiede un'adeguata e massiccia strumentazione concettuale di origine chimico-fisica, oppure alla genetica, in cui accanto allo studio sperimentale giocano un ruolo teorico fondamentale la statistica e il calcolo delle probabilità, mentre l'analisi computerizzata è un ingrediente imprescindibile per la sequenziazione e lo studio delle proprietà del genoma. Ciò che voglio sottolineare qui è la sempre maggior complementarità tra i due aspetti: i metodi quantitativi della genetica diventano essenziali per la ricostruzione filogenetica dell'evoluzione delle specie e quindi per una corretta riformulazione degli schemi di classificazione; parametri quantitativi sono sempre più importanti per la comprensione degli equilibri (e degli squilibri) degli ecosistemi; modelli originati nel contesto della fisica dei sistemi complessi sono utilizzati per la determinazione dei meccanismi che producono la struttura delle proteine e addirittura per la progettazione di nuove molecole complesse di uso anche farmacologico.
Sta prendendo sempre più piede, e sempre più spazio, una "biologia dei sistemi" che tenta in modo interdisciplinare di utilizzare nozioni provenienti dalle cosiddette "scienze dure" per introdurre elementi quantitativi e di modello in contesti largamente dominati, per loro stessa natura, da metodologie di ricerca essenzialmente osservative e qualitative.

Se guardiamo invece all'economia, vediamo che tale disciplina, pur appartenendo a pieno titolo al contesto delle scienze sociali, non può prescindere per moltissimi aspetti dalla dimensione quantitativa, non soltanto per la presentazione e l'analisi dei dati, ma anche (e forse soprattutto) per la modellazione dei processi. Da questo punto di vista si nota anzi una preoccupante distanza tra l'atteggiamento metodologico e culturale di molta parte degli economisti italiani, che concentrano la loro attenzione sugli aspetti qualitativi dei fenomeni economici, o al più su quegli aspetti di tipo quantitativo che possono ridursi all'analisi di *trend* e di serie statistiche, e gli studi di economia più tipici del mondo anglosassone, nei quali gli aspetti modellistici e la strumentazione matematica anche molto sofisticata (come la teoria dei giochi, vedi il premio Nobel per l'economia al matematico Nash) giocano un ruolo essenziale. Ancora una volta, non si tratta di stabilire una gerarchia ma di cogliere la complementarità: proprio l'enfasi su una modellazione troppo astratta e formale ha portato in molti casi, fino a tempi molto recenti, gli economisti teorici di scuola anglosassone a enfatizzare le capacità di autoregolazione di un "mercato" totalmente libero in cui gli operatori sarebbero portati a comportarsi in modo "razionale" (ossia sulla base del calcolo delle probabilità), e quindi a trascurare gli aspetti qualitativi e "non calcolabili" dei comportamenti economici, che non poco hanno inciso anche sulla recente crisi finanziaria mondiale.

In realtà è affascinante osservare (e ne discuteremo più avanti) che le teorie fisico-matematiche più avanzate sono in grado di tener conto, nei limiti in cui ciò è possibile (e vedremo quali sono tali limiti) anche degli aspetti di "non calcolabilità" che sono caratteristici dei sistemi complessi.
Sulla base degli esempi che abbiamo preso fin qui in considerazione, e di molti altri che si potrebbero trarre da quasi tutte le discipline, dall'astrofisica alla geologia, dalla linguistica alla sociologia, indipendentemente dalla loro collocazione nell'albero dei saperi, sembrerebbe di poter trarre due considerazioni di carattere abbastanza generale:

i) da un lato appare abbastanza evidente che l'analisi qualitativa gioca comunque un ruolo peculiare ogni volta che il fenomeno in esame sia materia di osservazione ma non di "esperimento", ovvero ogni volta che la complessità del contesto renda irrepetibili *in vitro* o, come si comincia a dire oggi, *in silico*, condizioni atte a riprodurre il fenomeno stesso in condizioni controllate, e pertanto a verificare o falsificare la validità del modello proposto sulla base degli esiti della simulazione o della prova sperimentale
ii) dall'altro il raffinamento dei metodi quantitativi, derivante soprattutto dallo studio dei "sistemi complessi", e il continuo potenziamento dei mezzi di calcolo elettronico e delle tecniche di simulazione numerica rendono oggi plausibile, in particolare, ma non soltanto, grazie alla simulazione *in silico* di situazioni di complessità anche notevole, un'utile applicazione di metodi quantitativi anche in circostanze, come quelle descritte al punto i), nelle quali fino a non molto tempo fa tali metodi potevano apparire ingenuamente riduzionisti.

Obiettivo di questo corso non è, se non in piccola parte, quello di discutere in dettaglio le caratteristiche e le tecniche d'uso delle (vecchie e nuove) metodologie quantitative che, originatesi spesso nel contesto della fisica teorica, si sono dimostrate utili per la trattazione e l'analisi di fenomeni e di situazioni anche molto differenti da quelle di partenza. Dopo l'illustrazione di alcuni principi di base dell'analisi quantitativa, talmente generali da dover costituire patrimonio comune di chiunque voglia seriamente analizzare un qualunque insieme di dati soggetti ad assumere forma numerica, ci si propone di illustrare alcune situazioni esemplari, e tecnicamente non troppo sofisticate, scelte soprattutto come evidenze dell'applicabilità di metodi quantitativi anche a contesti apparentemente molto refrattari, e del carattere universale di alcuni aspetti della dinamica dei sistemi. Discuteremo quindi in particolare alcuni fenomeni di tipo sociale (teoria delle reti, dinamica delle popolazioni, evoluzione linguistica e demografica) e alcuni fenomeni tratti dall'universo storico-letterario (frequenze di vocaboli, orizzonti cognitivi, diffusione dei cognomi).
Concluderà il corso una breve ricognizione dell'uso dei metodi della fisica sperimentale in contesti propri delle scienze umane (conservazione e studio dei beni culturali, archeometria)

2. I modelli matematici e il loro uso

Una domanda che può apparire del tutto legittima (e alla quale è quindi doveroso offrire una risposta articolata) è la seguente: perché, trattandosi della gestione e dell'analisi di dati di tipo quantitativo e fondamentalmente numerico, parlare di "metodologie fisiche" e non piuttosto di "metodologie matematiche" o al più di "metodologie statistiche"? La risposta sta (quasi) tutta nel concetto di "modello matematico", che è uno strumento nato, e assai comunemente sviluppato e usato, proprio nel contesto delle scienze fisiche.
Un modello matematico consiste in un insieme di relazioni formalizzate (e quindi spesso esprimibili tramite formule algebriche e con l'uso del concetto di "funzione") tra i valori misurabili di quantità differenti, che gli studiosi utilizzano per descrivere rapporti di causa-effetto, o anche soltanto di correlazione, tra le quantità stesse, e per predire di conseguenza l'evoluzione di alcuni parametri del sistema ("variabili dipendenti") quando sia già nota o prevedibile la variazione di altri parametri ("variabili indipendenti").
Variabile indipendente per eccellenza è il fattore tempo, il cui scorrimento è componente essenziale di ogni processo, ma non può a sua volta essere condizionato da alcun processo (se trascuriamo gli effetti relativistici che non sono certamente in alcun modo pertinenti alle nostre discussioni, salva l'eventualità che si vogliano qui discutere le pur affascinanti prospettive offerte dai viaggi spaziali).
Se analizziamo un orario ferroviario noi troviamo un'associazione quantitativa puntuale tra tempi e spazi che ci permette di conoscere con ragionevole esattezza (trascurando i ritardi) la velocità media dei treni nelle differenti tratte. Questa è certamente un'informazione di tipo quantitativo, ma non è sicuramente un modello matematico. Invece la relazione inversa, che ci dice che lo spazio percorso da un mezzo in un tempo assegnato è direttamente proporzionale alla velocità (media) del mezzo, è già, pur nella sua assoluta ed elementare semplicità, un "modello matematico". Conoscendo il tempo impiegato in media dai treni tra Milano e Roma, e sapendo che Firenze si trova all'incirca a metà strada, noi possiamo "azzardare" una previsione dell'orario d'arrivo a Firenze se conosciamo quello della partenza da Milano, basandoci su un'approssimazione (in questo caso quella detta "lineare") che per quanto rozza, ci fornisce, con una certa ma controllabile imprecisione, un'informazione quantitativa che, in mancanza del suddetto orario, non potremmo avere altrimenti.
Quando leggiamo che il petrolio finirà (o comunque la sua estrazione diventerà totalmente antieconomica) entro il 2050, o che la temperatura media aumenterà di due gradi prima della fine del secolo, o che la popolazione mondiale, prima di assestarsi o addirittura diminuire, raggiungerà i nove miliardi di individui, non stiamo leggendo le profezie di Nostradamus o le congetture speculative di qualche filosofo, ma il risultato dell'applicazione di un modello matematico. Ciò non significa che le affermazioni che ho menzionato siano intrinsecamente più attendibili di profezie esoteriche o di speculazioni filosofiche: in particolare, trattandosi di estrapolazioni, e non di interpolazioni come nell'esempio ferroviario, le procedure di validazione del modello sono molto più sofisticate e il grado di attendibilità è sempre comunque molto inferiore. Ma è importante, prima di esercitare una qualunque critica, comprendere bene la differenza concettuale che intercorre tra l'uso di un modello matematico e qualunque altra tecnica di previsione del futuro, dalla sfera di cristallo ai sondaggi d'opinione dei cosiddetti "esperti".
La principale caratteristica di un "buon" modello matematico consiste innanzitutto nella sua capacità di incorporare, e descrivere con ragionevole accuratezza, l'insieme di tutti i dati già conosciuti riguardanti il fenomeno che si vuole studiare. Se analizziamo una sequenza temporale di dati, è importante che la sequenza sia "lunga", perlomeno in relazione alla scala temporale rispetto alla quale ci interessa fare previsioni (non possiamo sperare di fare ipotesi ragionevoli sul numero di libri che saranno pubblicati l'anno prossimo se sappiamo soltanto quanti ne sono stati pubblicati il mese scorso), che i dati siano completi e accurati (non possiamo stimare quanti professori andranno in pensione nel prossimo quinquennio se non sappiamo quanti professori sono in servizio e qual è la loro attuale distribuzione in età).

Soprattutto però è essenziale che il nostro modello, se applicato a un qualunque troncamento temporale di una sequenza già nota, "predica" accuratamente le parti successive della sequenza, che già conosciamo ma (provvisoriamente) non abbiamo utilizzato nella costruzione del modello stesso.
Tutto ciò ancora non basta, perché esistono sistemi complessi per i quali, pur essendo note le leggi dinamiche che governano i processi all'interno del sistema, si può dimostrare in forma di teorema che l'evoluzione complessiva non è predicibile oltre un certo grado di precisione, che rapidamente diminuisce man mano che si allungano i tempi entro i quali si vorrebbe far valere la predizione.
Un esempio ormai "classico" di questo problema è la meteorologia. Le leggi fisiche che governano i moti dell'atmosfera sono ben note da molto tempo (si tratta in realtà di una singola formula, nota come "equazione di Navier-Stokes", che mette in relazione un insieme limitato di parametri misurabili quali temperatura, pressione, densità dell'aria, velocità del vento). Ciò nonostante, anche considerando superabile la difficoltà di introdurre negli strumenti di calcolo una rappresentazione sufficientemente accurata e dettagliata della superficie terrestre (anche la presenza di boschi o grandi edifici è importante nella determinazione del microclima), rimane sempre l'ostacolo teorico dell'instabilità matematica dell'evoluzione del sistema. Molti avranno sentito citare il famoso esempio per cui "il battito d'ali di una farfalla in Amazzonia può causare un temporale in Africa equatoriale": quest'immagine rappresenta efficacemente in modo "pittorico" il fatto che un minimo cambiamento nelle condizioni iniziali produce, a grande distanza nello spazio e nel tempo, effetti largamente divergenti tra loro.
Situazioni di imprevedibilità teorica si verificano, per motivi anche tecnicamente differenti da quello appena illustrato, in molte altre situazioni naturali e sociali, dai terremoti ai crolli di borsa, dagli incendi nei boschi ai collassi delle reti infrastrutturali. Come vedremo è tuttavia oggi possibile modellare, e quindi "capire", molto più che nel passato anche recente, la dinamica di questi fenomeni, comprendendone gli elementi e le caratteristiche che li accomunano al di là delle evidenti differenze di contesto, ed è anche possibile utilizzare il tipo di comprensione così acquisita, non per un'impossibile "previsione", ma se non altro per evitare l'eterogenesi dei fini, ossia la messa in atto di provvedimenti che, essendo volti ad evitare ciò che non può essere evitato, talvolta rischiano soltanto di aggravarne gli effetti (un esempio noto sono certe procedure di "manutenzione" dei boschi che, certamente evitando un buon numero di piccoli incendi, rendono tuttavia più tragico e disastroso l'effetto dei grandi incendi che è impossibile prevenire, in quanto derivanti da un accumulo di circostanze, statisticamente improbabile ma non matematicamente impossibile).

Non tutti i modelli matematici, peraltro, sono legati all'evoluzione temporale e hanno quindi il tempo come variabile indipendente. Quando analizziamo e modelliamo la connettività di una rete non effettuiamo un'analisi di tipo diacronico, ma ci limitiamo ad aspetti sincronici, che sono tuttavia rilevanti per rispondere a domande relative alla rapidità e alla stabilità delle connessioni, o anche alla vulnerabilità (fisica o informatica) della rete stessa. Quando analizziamo la frequenza dei vocaboli in un testo, o modelliamo la distribuzione delle citazioni delle pubblicazioni scientifiche, rispondiamo a domande importanti, nei rispettivi contesti, ai fini di un giudizio valutativo, e forniamo informazioni che vanno a confrontarsi con gli esiti di analisi di tipo qualitativo, e che quindi devono essere "attendibili" nel senso che abbiamo cercato di definire sopra, e quindi richiedono una grande capacità "descrittiva" del modello, ma non comportano alcuna capacità "predittiva" dello stesso. Anche in questo caso è comunque fondamentale la distinzione tra interpolazione ed estrapolazione, ed anche in questo caso è proprio sulla seconda che possono appuntarsi e concentrarsi le critiche più incisive.

Avendo chiarito, almeno a grandi linee, che cosa significhi il modello matematico di un fenomeno e quali possano essere le finalità (descrittive e/o predittive) della sua costruzione, abbiamo il dovere di approfondire ulteriormente le possibili limitazioni, materiali e concettuali, della modellazione, confrontandoci anche con un "dogma" talvolta sottaciuto, ma quasi sempre dato per scontato, della concezione "fisicalista" della scienza, quello del falsificazionismo popperiano.

Non v'è alcun dubbio sul fatto che, come abbiamo già visto, in tutte le discipline (non soltanto "umanistiche") per le quali non è possibile ipotizzare la ripetibilità dell'esperimento, anche le procedure volte alla verificazione/falsificazione delle proposizioni che si vogliono "scientifiche" (e quindi in particolare volte alla verificazione/falsificazione dei modelli matematici dei fenomeni in esame) risultano di dubbia applicabilità. Ma se è vero che non possiamo chiedere a Dante Alighieri di scriverci una differente versione della Commedia, è altrettanto vero che non possiamo ripetere il Big Bang, la deriva dei continenti o l'evoluzione degli ominidi. Non per questo pensiamo che la cosmologia, la tettonica globale o la paleoantropologia non siano "scienze", e non per questo ci rifiutiamo di fornire una modellazione matematica (ovviamente a diversi livelli di approssimazione) dei fenomeni (o di alcuni dei fenomeni) che sono studiati da queste discipline. Sembra quindi opportuno rilassare un poco la nozione di "falsificabilità" come criterio di verità e scientificità di una proposizione, accettando pragmaticamente una nozione di verità provvisoria e storicizzata, per cui si classificherà come più "vera" la proposizione (e nel nostro caso il modello) capace di descrivere meglio i fatti noti e soprattutto di rendere meglio conto, e con la maggiore economia di pensiero possibile, del maggior numero di fatti, e di dati, oggi a disposizione.
Nell'analisi quantitativa esistono, come vedremo, precisi criteri, anch'essi quantitativi, per misurare l'aderenza di un modello ai dati disponibili, e in tal senso è quindi possibile stilare una "graduatoria" di bontà della modellazione. Dovrebbe essere chiaro a questo punto che non è in alcun modo possibile attribuire una "verità" meta-fisica a questa procedura, e quindi ipostatizzarne gli esiti come se si trattasse di un'acquisizione permanente e indiscutibile, anche perché, nel caso di modelli di tipo predittivo, potrà essere la storia futura, in luogo dell'esperimento, a svolgere il compito di verificare o falsificare il modello, semplicemente confermandone o smentendone le previsioni. Ma, ancora una volta, non ci troviamo di fronte a una peculiarità assoluta delle scienze umane: il giorno in cui ci accorgessimo che l'Atlantico ha cominciato a restringersi anziché continuare ad allargarsi anche le nostre più consolidate dottrine in materia di tettonica a placche dovrebbero essere rimesse in discussione (e per carità di patria scientifica non vorrei neppure prendere in considerazione l'ipotesi che ci giungessero improvvisamente segnali di un cambiamento di rotta delle galassie lontane!).

Resta, in qualche modo insormontabile, l'obiezione "filosofica" a un approccio quantitativo alle scienze umane che si fonda sulla nozione di "irriducibile varietà del reale", che è forse nient'altro che il riflettersi sulla materia "umana" delle discipline di un'idea della "natura umana" incentrata sull'idea di "libero arbitrio". Non è qui questione della fondatezza o meno di tale idea: vorrei soltanto osservare che questo tipo di obiezioni sembra molto legato a un'ormai del tutto scomparsa concezione meccanicista dei fenomeni studiati dalle scienze naturali. Le nozioni di probabilità e di indeterminazione sono ormai parte integrante del bagaglio culturale anche delle cosiddette "scienze dure" (infelice traduzione dell'infelice locuzione *hard science*), come dovrebbe risultare ormai chiaro anche da una rilettura critica di molta parte della discussione precedente. Non so se le mie "scelte" avvengano in virtù dell'esercizio di un mio effettivo "libero arbitrio", però so per certo che le mie scelte (ma anche quelle della famosa farfalla dell'Amazzonia) non possono essere predette (né condizionate) con precisione da nessuno strumento fisico, matematico o informatico. Viceversa so per certo che in media ogni cinque secondi nel mondo muore di fame un bambino, e che nessun esercizio di libero arbitrio può modificare in modo sostanziale, perlomeno sulla scala di tempo della mia personale esistenza, questa dolorosa verità statistica. Il motivo per cui non si può predire il futuro risolvendo, come immaginava Asimov in uno dei sui cicli più famosi, le equazioni della "psicostoria" non sembra risiedere tanto nella libertà dell'uomo quanto nelle proprietà matematiche degli "attrattori strani" che governano l'evoluzione dei sistemi complessi.

3. Alcune nozioni elementari di statistica

Nell'analisi dei fenomeni naturali è certamente importante acquisire il maggior numero possibile di dati sperimentali onde evitare di trarre deduzioni indebite dall'osservazione di fenomeni che avvengono in realtà soltanto in condizioni molto peculiari o per il verificarsi di eventi casuali possibili ma certamente improbabili. Quest'affermazione è tanto più vera quando prendiamo in esame fenomeni che appartengono alla sfera delle "scienze umane". In quest'ambito infatti è quasi sempre impossibile che un singolo fenomeno possa essere ricondotto a una singola legge generale, in quanto la complessità del sistema fa sì che, anche qualora sia possibile individuare una qualche "legge", essa si applicherà sempre in combinazione (spesso non controllabile) con molti altri fattori circostanziali i cui effetti non possono essere trascurati e pertanto, non sarà comunque quasi mai possibile giungere a una "predizione" di tipo "meccanico" del comportamento del sistema.
Pertanto il ruolo, già straordinariamente importante, assunto dalla statistica nello studio dei fenomeni naturali sarà infinitamente esaltato nel momento in cui si intraprenda uno studio quantitativo di fenomeni storico-sociali. Qualunque "verità" si possa cercar di asserire in questo contesto non potrà comunque essere che una "verità statistica", cioè verificabile non in relazione al singolo fenomeno ma soltanto in relazione a un insieme sufficientemente ampio di fenomeni tra loro confrontabili. Paradossalmente quest'affermazione risulta del tutto adeguata anche all'enunciazione delle leggi relative al comportamento delle particelle elementari nel corso delle loro interazioni fondamentali: nessuno è in grado di dire che cosa farà una singola particella, ma siamo in grado di predire che cosa farà un insieme di particelle con una precisione tanto più grande quanto più grande è il numero delle particelle prese in esame (in circostanze confrontabili)
Queste considerazioni generali e preliminari ci impongono a questo punto di introdurre un adeguato armamentario di nozioni tipiche della scienza statistica, onde creare un linguaggio comune e dotarci della strumentazione tecnicamente necessaria per effettuare in modo scientificamente corretto le analisi, che anche nei casi più elementari devono comunque corrispondere a paradigmi condivisi.

Nella trattazione statistica di un fenomeno di qualsivoglia natura, la prima e più fondamentale nozione che deve essere introdotta e formalizzata è quella di "distribuzione". A partire da una "popolazione" (parola generica che indica un qualunque insieme di soggetti d'indagine, dagli abitanti di una nazione ai vocaboli di un testo) caratterizzata dalla presenza di "variabili" (proprietà misurabili, per le quali esistano più valori distinti e ammissibili, degli elementi della popolazione) chiamiamo distribuzione di una certa variabile nella popolazione il numero delle ricorrenze all'interno della popolazione di ciascuno dei valori ammissibili della variabile in esame. Ad esempio la distribuzione in altezza (in cm) di un gruppo di persone è il numero di persone appartenenti al gruppo che hanno un'altezza (opportunamente arrotondata al cm) corrispondente a ciascuno dei valori possibili. Si noti che i valori della variabile possono essere "discreti", ovvero ben separati l'uno dall'altro (come i cognomi presenti in una popolazione urbana) oppure "continui" (come le altezze dell'esempio precedente), nel qual caso la definizione della distribuzione non può prescindere da un meccanismo di "discretizzazione" (nel nostro esempio l'arrotondamento al cm), che deve essere tarato con grande cura per ottenere risultati significativi (se arrotondassimo le altezze al metro potremmo al massimo separare la popolazione, peraltro in modo assai arbitrario, in due grandi gruppi,"alti" e "bassi"). Per qualunque caratteristica suscettibile di conteggio è possibile definire la relativa distribuzione, ma dovrebbe essere chiaro che non tutte le distribuzioni sono ugualmente interessanti, e fa parte del lavoro di ricerca identificare le variabili la cui distribuzione può rivestire un particolare interesse.

La nozione di distribuzione ci permette di definire meglio la distinzione fondamentale che viene comunemente operata tra le due principali branche della scienza statistica, parlando rispettivamente di "statistica descrittiva" e di "statistica inferenziale".

4. La statistica descrittiva e la teoria dell'errore

La statistica descrittiva ha come scopo fondamentale la raccolta, la presentazione e l'analisi dei dati relativi alla popolazione che costituisce l'oggetto dello studio. Pertanto essa consiste essenzialmente nella costruzione delle distribuzioni e nella valutazione delle loro principali caratteristiche, quali la media, la varianza , le correlazioni (tutte queste nozioni saranno definite meglio in seguito).
La statistica inferenziale ha invece come obiettivo quello di fare affermazioni relative alla natura teorica (la legge di probabilità, le legge di evoluzione temporale) delle distribuzioni osservate.
Se tale natura teorica (che normalmente si concretizza in un modello matematico) può essere identificata tenendo sotto controllo la possibilità di errore, diventa allora possibile "inferire" da tale conoscenza "predizioni" sul comportamento di sistemi simili, o sull'evoluzione futura del sistema in esame. È intrinseco alla natura stessa delle inferenze statistiche il carattere probabilistico di tali predizioni, che ne costituisce il limite teorico ed empirico: la consapevolezza di tale limite è condizione essenziale per un uso concettualmente appropriato e praticamente adeguato dell'analisi statistica.
Un'ulteriore evoluzione della statistica inferenziale è rappresentata dalla cosiddetta "statistica esplorativa", che si applica ai casi in cui non sono *a priori* ben individuate le ipotesi sulla legge di probabilità che governa il fenomeno e le variabili rilevanti alla sua analisi, e in cui è pertanto necessaria la formulazione e la verifica di congetture di carattere molto generale: la statistica esplorativa a sua volta sta alla base delle procedure di *data mining*.
Tra le principali caratteristiche della distribuzione P(x) di una qualunque variabile numerica x v'è in primo luogo la sua media, definita dalla relazione

$$<x> = \sum x\, P(x) / \sum P(x),$$

dove è bene notare che il denominatore $\sum P(x)$ non è altro che il numero totale degli elementi.
Grandissima importanza riveste di solito anche la varianza (scarto quadratico medio dalla media), che è una misura della "larghezza" della distribuzione, definita dalla relazione

$$(\Delta x)^2 = \sum (x - <x>)^2\, P(x) / \sum P(x).$$

Per molte distribuzioni la maggioranza degli elementi della popolazione hanno valori della variabile che stanno a una distanza dal valor medio non maggiore di una "deviazione standard" Δx.
Altre nozioni importanti sono quelle di "mediana" (il valore di x tale per cui metà degli elementi della popolazione ha un valore inferiore alla mediana e l'altra metà ha un valore superiore) e di "moda" (il valore più probabile di x, ovvero il valore per cui P(x) è massimo).

Un'importante applicazione della statistica è la teoria dell'errore, il cui impiego nella valutazione dei dati empirici risulta fondamentale per una loro corretta valutazione.
Quando si valuta l'entità quantitativa di un qualunque fenomeno, esistono numerosi e diversi motivi per cui la valutazione empirica non può essere considerata "esatta". Nelle misure fisiche dobbiamo tener conto innanzitutto dell'imprecisione e dell'inaccuratezza strumentale (l'orologio può andare avanti o indietro, e ci indica un'unità minima, ad esempio il secondo, e non le sue frazioni), ma in numerosi altri casi l'errore deriva innanzitutto dalla limitatezza del campione utilizzato (pensiamo per esempio a un sondaggio d'opinione). L'effetto immediato di tale imprecisione si osserva nel fatto che misure ripetute della medesima quantità daranno risultati tra loro (più o meno) differenti.
I diversi risultati di una misura della stessa quantità (ad esempio l'altezza media di una popolazione stimata facendo la media su tanti piccoli gruppi diversi) costituiscono essi stessi una distribuzione, che può essere analizzata valutando i parametri che abbiamo già definito nel caso generale.
Diremo allora che la media della distribuzione delle misure è il "valore atteso" della misura (ossia quello che ci sembra il valore più probabile per la quantità che realmente ci interessa), mentre l'"errore" da attribuire alla misura stessa (ossia l'intervallo di valori entro il quale è comunque plausibile che possa trovarsi il valore "vero" della quantità che ci interessa) è legato alla varianza della distribuzione, e si considera misurato dalla deviazione standard divisa per la radice quadrata del numero di misure effettuate.

Si noti bene che questa definizione risulta appropriata soltanto nel caso in cui l'errore di cui stiamo parlando possa definirsi "casuale", ovvero derivante dall'esito di operazioni che in quanto tali non condizionano il risultato se non per effetto del calcolo delle probabilità. Esiste tuttavia, e deve essere tenuta attentamente in conto, la possibilità di un errore "sistematico", derivante da procedure che , in quanto tali, potrebbero condizionare in qualche modo l'esito del risultato. I sondaggi sono un esempio tipico di "misure" facilmente soggette a un errore sistematico: per fare un esempio estremo pensiamo a un sondaggio d'opinione in cui venissero intervistati soltanto maschi: sarebbe del tutto scorretto, per quanto possa essere vasto il campione sondato (incluso il caso in cui venissero intervistati "tutti" i maschi), inferire dall'esito di questa procedura l'opinione dell'intera popolazione. Le possibilità di errore sistematico sono molteplici, e andrebbero sempre valutate con cura: nel caso delle scienze umane poi occorre tenere presente che, quando si valuta un insieme di documenti, in particolare risalenti a epoche lontane, gli eventi storici hanno già operato su quei documenti una "selezione" talvolta casuale, ma più spesso culturale, ideologica o estetica, e pertanto il nostro giudizio, anche quando si fonda su basi quantitative, è certamente condizionato da un non trascurabile errore sistematico.

Le procedure per la stima degli errori casuali, e per la possibile identificazione di errori sistematici, possono essere, in casi complessi, molto sofisticate, e in questa sede non intendiamo approfondire gli aspetti tecnici di tali procedure, limitandoci ad alcune considerazioni di carattere abbastanza generale. Notiamo in primo luogo che dalla teoria dell'errore discende immediatamente la nozione di "cifre significative": si considerano significative, nella presentazione del risultato di una misura, soltanto le cifre fino a quella che può essere modificata restando all'interno dell'intervallo definito dalla stima dell'errore, per cui ad esempio se l'errore in un sondaggio è stimato al 2%, ha senso dire che una certa opinione è condivisa del 57% degli intervistati, ma ha ben poco senso dire che tale percentuale è del 57,3% (e non ha alcun senso dire che è del 57,34%!): l'introduzione delle cifre "non significative" non rende più precisa l'informazione, ma in compenso la rende più "illeggibile".
Un'altra importante caratteristica dell'errore nelle misure è la sua proprietà di "propagazione": quando la stima di una quantità risulta dalla composizione di dati differenti, l'errore nella stima sarà percentualmente maggiore degli errori attribuiti separatamente ai singoli dati, e tipicamente l'errore relativo (percentuale) finale sarà dato dalla somma degli errori relativi parziali. Se misuro una velocità come rapporto tra lo spazio percorso e il tempo impiegato a percorrerlo, e la misura dello spazio ha un errore del 3% e quella del tempo ha un errore del 2%, l'errore sulla velocità risulterà inevitabilmente del 5%. Questo significa anche che, nelle misure composte, è perfettamente inutile cercare di raffinare la misura di una delle variabili se la misura delle altre è di per sé affetta da errori significativamente più grandi: misurare con la precisione del grammo il peso di dieci persone prese a caso tra mille non ci aiuterà a conoscere meglio il peso medio dell'intera popolazione. Anche per questo motivo è spesso bene diffidare della "falsa accuratezza" di certe stime nel campo delle scienze umane. Si pensi ai bilanci pubblici, nei quali i dati finanziari sono riportati con la precisione del centesimo, mentre i presupposti sociopolitici sulla base dei quali le cifre sono state calcolate sono spesso conosciuti con imprecisioni fino a diversi punti percentuali.
Un importante risultato della teoria dell'errore è il cosiddetto teorema del limite centrale, che assicura che, sotto ipotesi abbastanza generali, la distribuzione dei risultati di una misura tende, quando il numero delle misure sia veramente molto elevato, ad assumere una forma caratteristica universale (curva a campana, o curva di Gauss), la cui espressione matematica (distribuzione normale, o di Gauss) è la seguente:

$$P(x) = N/\sqrt{(2\pi\ \Delta x\text{^}2)}\ \exp\{ - (x-\langle x\rangle)\text{^}2/2\ \Delta x\text{^}2\},$$

dove N è il numero totale delle misure effettuate. Si tratta di un risultato di grande rilevanza teorica e pratica, perché permette di ricavare come corollari numerose altre proprietà delle distribuzioni dei risultati di misure, e anche perché un significativo scostamento dalla distribuzione normale può in certi casi essere indice della presenza di errori sistematici.

5. La statistica inferenziale: regressione ed estrapolazione

Nell'ambito della statistica inferenziale, come abbiamo già accennato, è fondamentale essere in grado di costruire modelli matematici che riproducano e "giustifichino" le relazioni empiriche che intercorrono tra le variabili, individuando leggi di probabilità per le distribuzioni, leggi di evoluzione temporale o altre relazioni funzionali.
Abbiamo già discusso il caso della distribuzione normale, la cui forma risulta determinata una volta noti due soli parametri empirici (media e varianza). Un altro tipo di distribuzione che compare spesso, soprattutto nella descrizione dell'evoluzione temporale dei fenomeni, è l'esponenziale, che rappresenta la forma matematica dei processi in cui il valore della variabile dipendente raddoppia (o si dimezza) in un preciso intervallo di valori della variabile indipendente. Quando la variabile indipendente è il tempo, il suddetto intervallo è chiamato "tempo di raddoppiamento" (o di dimezzamento). Si parla anche, in questo caso, di crescita (o decrescita) in proporzione geometrica. Un esempio classico di tale andamento è la crescita nel tempo delle popolazioni secondo Malthus[1], ma è assai facile trovare altri esempi di andamento esponenziale sia nel mondo naturale che in quello sociale: in particolare il decadimento delle sostanze radioattive segue con precisione assoluta e immodificabile una legge esponenziale e proprio grazie a questo fatto, misurando nei reperti archeologici la concentrazione attuale di specifici isotopi (ad esempio il carbonio 14), per i quali è possibile presumere il valore della concentrazione iniziale, si arriva a stabilirne la datazione.
La spiegazione della frequenza con cui si trovano distribuzioni esponenziali è abbastanza semplice: infatti tali distribuzioni emergono come soluzioni della più semplice delle equazioni che regolano matematicamente i rapporti di causa-effetto, quella che si basa sulla teoria della "risposta lineare", che nella sua più semplice formulazione consiste nell'affermazione che l'effetto è direttamente proporzionale alla causa. Nell'esempio delle popolazioni, la risposta lineare consiste nel fatto che il numero delle nascite è, in condizioni normali, proporzionale al numero dei componenti della popolazione stessa; del tutto analogo è il ragionamento nel caso dei decadimenti radioattivi.

In molti casi per i quali esista una ragionevole congettura sulla forma matematica generale del modello (fenomeni lineari, fenomeni ad andamento esponenziale, fenomeni governati da leggi di potenza) risulta possibile, direttamente o indirettamente (ad esempio operando sui logaritmi dei valori delle variabili) ricondurre il problema dell'individuazione dei parametri (ossia delle costanti caratteristiche) del modello stesso all'effettuazione della cosiddetta "regressione lineare".
La regressione lineare è l'operazione matematica che permette di ottimizzare i parametri a e b della relazione $y\_i = a + b\ x\_i$ che dovrebbe esprimere in forma matematica il collegamento (lineare, ossia di proporzionalità diretta) tra i valori i-esimi assunti da una variabile (dipendente) y e i corrispondenti valori di una variabile (indipendente) x. L'indice progressivo i, variabile tra 1 e N, ordina e numera i risultati delle N osservazioni empiriche effettuate.
Senza dimostrazioni matematiche ci limitiamo a ricordare le seguenti formule fondamentali della regressione lineare:

$$a = [\langle x^2\rangle\langle y\rangle - \langle x\rangle\langle xy\rangle]/\Delta \qquad b = [\langle xy\rangle - \langle x\rangle\langle y\rangle]/\Delta$$

dove vale $\Delta = \langle x^2\rangle - \langle x\rangle^2$ e tutte le medie $\langle\ldots\rangle$ sono calcolate come somme delle quantità indicate, estese a tutti i valori dell'indice i da 1 a N e divise per N.

Esistono tecniche di controllo, sulle quali non ci dilungheremo in questa sede, per stabilire se la regressione lineare fornisce una buona rappresentazione della relazione che intercorre tra le due variabili prese in esame. Un criterio molto semplice per stabilire *a priori* se esiste tra due variabili una relazione di proporzionalità diretta consiste tuttavia nello studiare la loro "correlazione".

[1] T.R. Malthus, *An essay of the principle of the population as it affects the future improvement of the society*, 1798

La nozione di correlazione statistica è di grande interesse e può essere utilizzata in numerosissimi contesti. La definizione matematica di correlazione tra due insiemi di valori x_i e y_i è la seguente:

$$C(x\_i, y\_i) = \frac{\sum (x\_i - \langle x \rangle)(y\_i - \langle y \rangle)}{\sqrt{[\sum (x\_i - \langle x \rangle)^2 \sum (y\_i - \langle y \rangle)^2]}}$$

È relativamente facile dimostrare che la correlazione C così definita è un numero compreso tra -1 e 1, e che i valori estremi (correlazione 1 e anticorrelazione -1) rappresentano situazioni in cui la regressione lineare si applica a pieno titolo in quanto esiste una proporzionalità diretta tra le due variabili, mentre nei casi intermedi la correlazione è parziale e addirittura, quando C assume valori prossimi allo zero, si parla di assenza di correlazione e le due variabili sono da considerarsi in realtà indipendenti l'una dall'altra.

La determinazione dei parametri del modello, effettuata mediante la regressione lineare o eventualmente mediante sue opportune generalizzazioni, sta ovviamente alla base di tutte le procedure di inferenza, le quali a loro volta vanno distinte in due grandi categorie distinte: procedure di interpolazione e procedure di estrapolazione.
L'interpolazione consiste nella determinazione congetturale dei valori che la variabile dipendente potrebbe assumere in corrispondenza di valori della variabile indipendente intermedi tra quelli effettivamente osservati. Nell'esempio ferroviario, l'interpolazione permette di stimare gli orari di passaggio nelle differenti stazioni, note le loro distanze dalla stazione di partenza, l'orario di partenza (corrispondente al parametro a della regressione lineare) e la durata complessiva del viaggio fino alla stazione finale, grazie alla quale si può calcolare la velocità media (il cui inverso è il parametro b della regressione lineare).
La procedura d'interpolazione è di regola sia concettualmente che matematicamente molto solida, e le stime così generate hanno un errore legato all'errore medio con cui sono noti i dati empirici e alla precisione con cui la regressione riproduce i dati stessi, precisione che può essere valutata con grande accuratezza.
Tutt'altro discorso vale per l'estrapolazione, che consiste nello stimare i valori che la variabile dipendente potrebbe assumere in corrispondenza di valori della variabile dipendente che si trovano all'esterno dell'intervallo dei valori per cui si hanno i dati. Il procedimento matematico è simile, perché anche in questo caso si tratta di individuare una funzione che riproduca i valori noti di y al variare di x con sufficiente accuratezza, e poi ricalcolarla per i valori di x che costituiscono l'oggetto della stima, ma l'attendibilità del risultato è in genere molto inferiore, in quanto esistono quasi sempre molti modelli che, in un intervallo limitato, hanno comportamenti molto simili (e che quindi potrebbero essere considerati, per quell'intervallo, "buoni modelli") ma che al di fuori dell'intervallo manifestano comportamenti anche pesantemente divergenti tra loro. Ovviamente la procedura di estrapolazione è, per sua stessa natura, tanto più inattendibile quanto più ci si allontana dall'intervallo per il quale si possiedono i dati.

6. L’uso delle tabelle e dei grafici

Chiunque debba presentare dati quantitativi deve porsi l’obiettivo di rappresentarli in una forma che sia facilmente comprensibile e interpretabile. È evidente che lunghe liste di numeri non possono soddisfare quest’esigenza, perché è difficile memorizzare sequenze e quindi effettuare mentalmente confronti tra grandi quantità di dati numerici.
Ciò detto, è spesso comunque indispensabile presentare tabelle numeriche. In tal caso è importante che in esse siano contenuti soltanto dati essenziali e sintetici, organizzati in modo tale da permettere rapide comparazioni. A ogni colonna dovrebbe corrispondere una delle variabili prese in esame, riservando la prima colonna (da sinistra) alla variabile “indipendente” (ad esempio la data).
Ogni volta che ciò sia possibile, l’intervallo tra i valori delle variabile indipendente, riportati riga per riga nella prima colonna, dovrebbe essere sempre lo stesso, cosicché la “distanza” tra le righe sia sempre proporzionale alla distanza tra i valori della variabile stessa. Ad esempio, se si tratta di date, i valori della prima colonna dovrebbero essere giorni, o mesi, o anni successivi, senza “buchi” e, se per una o più delle date non sono disponibili i valori delle variabili dipendenti, o di alcune di esse, si dovrebbe indicare tale indisponibilità nelle corrispondenti righe e colonne. Per i motivi detti in premessa, è bene evitare di presentare i dati relativi a variabili strettamente collegate tra loro, i cui valori siano ricavabili l’uno dall’altro (come sarebbe ad esempio il prezzo in lire e in euro, o il valore energetico in Calorie e Joule): in questo caso è bene effettuare una scelta e presentare solo i valori della variabile più significativa, eventualmente indicando in nota la relazione con cui ottenere gli altri valori . Nella legenda della tabella (ad esempio nella riga in cui è indicato il nome della variabile) occorre sempre specificare l’unità di misura in cui sono espressi i valori, anche quando sembra ovvia: raramente ciò che è ovvio per chi scrive lo è altrettanto per chi legge.
I valori numerici andrebbero sempre riportati in un formato conforme a quanto consegue dalla teoria dell’errore, presentata in precedenza. In particolare non ha alcun senso riportare cifre “non significative”, ossia le cifre successive a quella il cui valore può essere modificato restando all’interno dell’intervallo predetto dall’errore stimato. Per essere concreti, se abbiamo stimato un errore prossimo all’1%, non più di tre cifre possono essere significative, e così via. Quando è noto ed è rilevante, si dovrebbe indicare anche (dopo il segno “più o meno”) il valore dell’errore stimato.
Per fare un esempio concreto, se parlassimo della popolazione terrestre “oggi”, non avrebbe comunque alcun significato specificare il numero indicando le cifre successive a quella dei milioni, in quanto anche se fossimo in possesso di censimenti “perfetti” (cosa che ovviamente non avviene), non potremmo ignorare il fatto che ogni giorno sulla Terra nascono e muoiono, in modo del tutto casuale, centinaia di migliaia di persone.
Un’alternativa spesso preferibile alla presentazione di tabelle di dati consiste nella predisposizione di istogrammi o di grafici. Gli istogrammi sono costituiti da rettangoli di lunghezza proporzionale al valore della variabile dipendente, affiancati verticalmente o anche orizzontalmente in sequenza secondo l’ordine dei valori della variabile indipendente. L’uso del colore (o del tratteggio) permette di costruire istogrammi in cui sono riportati i valori di più variabili: tale scelta è particolarmente efficace quando i valori delle variabili sono espressi nelle stesse unità di misura, e sono quindi direttamente confrontabili, ma talvolta anche nel caso di quantità misurate da unità differenti la rappresentazione nello stesso istogramma può aiutare a mettere in evidenza una relazione di proporzionalità: si pensi ad esempio a un istogramma in cui, anno per anno, si riportino i Km percorsi da un mezzo a motore e i litri di carburante consumati. Un’alternativa interessante alla presentazione di valori relativi a variabili tra loro affini mediante barre affiancate o sovrapposte si ha quando le variabili dipendenti da presentare siano soltanto due, e in qualche modo tra loro alternative (ad esempio maschi e femmine, o entrate e uscite). In questo caso conviene riportare i segmenti proporzionali ai valori delle due variabili, riferiti allo stesso valore della variabile indipendente, dai lati opposti di uno stesso asse, verticale o anche orizzontale, creando una figura la cui simmetria, o mancanza di simmetria, rispetto all’asse stesso è già indicazione dell’equilibrio (o squilibrio) tra le variabili.

Quando invece si tratta di rappresentare i valori percentuali di diverse variabili che concorrono a formare la totalità dei dati (come ad esempio nel caso di un risultato elettorale) un'alternativa di immediata efficacia visiva è la cosiddetta "torta", consistente in un cerchio diviso in spicchi (spesso colorati) la cui larghezza (in rapporto all'intera circonferenza) è pari alla percentuale attribuita a ciascuna variabile.

L'istogramma può in realtà essere visto una forma, rozza e didascalicamente efficace, di grafico. I grafici sono, in generale, uno degli strumenti più efficaci per la rappresentazione e l'interpretazione di dati numerici o comunque quantitativi. Alla base dell'idea stessa di grafico sta la nozione di piano cartesiano, ossia di una coppia di assi tra loro ortogonali (l'asse orizzontale è detto asse delle ascisse, o delle *x*, quello verticale è l'asse delle ordinate, o delle *y*) che dividono il piano in quattro regioni distinte. Ogni punto del piano è individuato in modo univoco specificando la sua distanza da ciascuno dei due assi: il valore della distanza (orizzontale) dall'asse delle *y* ((positivo se il punto è a destra dell'asse, negativo se è a sinistra) è il valore dell'ascissa, alla quale di solito si associa la variabile indipendente, mentre i valori delle distanze (verticali) dall'asse delle *x* (positive se i punti sono sopra l'asse, negative se sono sotto), sono associati a una o più variabili dipendenti.
La rappresentazione mediante un grafico permette in molti casi di individuare più facilmente un'eventuale relazione funzionale tra le variabili dipendenti e quella indipendente (ad esempio un andamento lineare, associato alla proporzionalità diretta), e anche di ricavare "graficamente" il valore di parametri che andrebbero altrimenti stimati mediante calcoli relativamente complessi.
Alcune regole e convenzioni aiutano a costruire grafici chiari e comprensibili, a cominciare dal fatto che è sempre bene indicare lungo gli assi coordinati il nome delle quantità ad essi associate e la relativa unità di misura adottata. La scelta dell'unità di misura lungo l'asse delle *y* è particolarmente importante per una buona resa grafica: occorre evitare che tutti i dati siano "appiattiti" intorno all'asse delle *x* (cosa che avviene se si è scelta un'unità di misura troppo grande) o viceversa che alcuni dati finiscano "fuori dal grafico" (come accade se si sceglie un'unità troppo piccola). Occorre poi indicare, lungo gli assi, sequenze di valori tipici ed equispaziati (tipicamente i valori che corrispondono a multipli significativi delle unità di misura, ad esempio unità, decine, o centinaia).
Nel caso in cui alle variabili sia associato un errore quantitativamente importante, i dati devono essere riportati sul piano non come punti ma sotto forma di "barre" che coprano l'intervallo di valori previsto dall'errore (ad esempio se la variabile *y* quando *x* vale 3 assume il valore 14 con un errore di 2, il dato deve essere rappresentato da un segmento verticale posto in corrispondenza dell' ascissa 3, esteso dall'ordinata 12 all'ordinata 16 e marcato a metà da un trattino orizzontale).
In taluni casi può essere utile unire con una linea i punti che rappresentano i valori riportati nel grafico. Quando si è in grado di formulare un modello matematico che descriva il fenomeno studiato, è possibile rappresentare sul grafico, mediante una linea continua, l'andamento della funzione prevista dal modello, che in generale potrà dirsi "buono" se la linea della funzione passa all'interno di tutte le barre d'errore.
Non sempre è conveniente utilizzare, su uno o su entrambi gli assi cartesiani, una scala lineare. Un esempio tipico è rappresentato dai fenomeni che hanno un andamento esponenziale, crescente (come accade in certi periodi per le popolazioni) o decrescente (come nel caso delle curve di sopravvivenza, o nelle code di molte distribuzioni). In casi di questo genere conviene usare sull'asse delle *y* una scala logaritmica (per cui i valori successivi equispaziati, in luogo di 0,1,2,3,… sono, ad esempio, 1, 10, 100, 1000,…). Su questa scala i fenomeni esponenziali hanno un andamento grafico lineare, che è molto più facilmente riconoscibile anche a colpo d'occhio.
Un altro caso molto interessante anche per le nostre successive discussioni è quello delle leggi di potenza, per cui la relazione funzionale tra le variabili è del tipo $y = A\, x^b$, dove A è una costante e b un numero reale (l'"esponente" dell'andamento a potenza). In questo caso conviene usare, nella rappresentazione grafica, una scala doppiamente logaritmica, per sfruttare il fatto che la relazione indicata si traduce in una relazione tra i logaritmi delle variabili che ha la forma

$$\mathrm{Log}(y) = \mathrm{Log}(A) + b\ \mathrm{Log}(x).$$

Di conseguenza una legge di potenza prende, se rappresentata su scala doppiamente logaritmica, una forma grafica lineare, e l'esponente b può essere facilmente stimato misurando la pendenza della retta che passa per i punti indicati nel grafico.
Abbiamo già discusso, nel caso degli istogrammi, come rappresentare i dati nel caso di una contrapposizione binaria. Può essere utile considerare anche il caso in cui si debba rappresentare la ripartizione percentuale tra tre possibili scelte (come nel caso di un sistema politico prevalentemente, ma non strettamente, maggioritario, o quando la risposta prevista per un quesito sia SI/NO/NON SO). In questo caso è particolarmente efficace la rappresentazione grafica detta "*plot* di Dalitz", che sfrutta il teorema di geometria euclidea per cui, su un piano, la somma delle distanze di un punto dai tre lati di un triangolo equilatero è costante. In questo modo a ogni terna di percentuali la cui somma sia 100 corrisponde in modo univoco un singolo punto all'interno di un triangolo equilatero, e il modo in cui si dispongono e si raggruppano i punti quando si analizzano molte terne differenti di valori offre già a livello visivo informazioni molto precise e significative sulle caratteristiche e la variabilità del sistema sotto esame.

Vogliamo in questa sede esaminare anche un utilizzo del piano cartesiano che rappresenta una sorta di compromesso tra l'analisi qualitativa e quella quantitativa. Si tratta della possibilità di rappresentare in forma grafica l'effetto combinato di differenti comportamenti sociali. Immaginiamo di avere due atteggiamenti sociali sostanzialmente indipendenti l'uno dall'altro, e rispetto ai quali esista una gamma di atteggiamenti classificabili su una scala che vada dal positivo al negativo, passando per posizioni neutre (pensiamo ad esempio a concetti come laicismo e progressismo: si può essere da totalmente laici a niente affatto laici, e da totalmente progressisti a niente affatto progressisti, ma i due atteggiamenti non sono, né in linea di principio né in pratica, correlati in alcun modo l'uno all'altro). Se ora sul piano cartesiano attribuiamo a ciascun soggetto analizzato un valore numerico, positivo o negativo, in relazione a ciascuno dei due atteggiamenti, e riferiamo l'asse delle x e quello delle y alle due variabili in esame, a ogni individuo analizzato corrisponderà un punto del piano, più o meno lontano dal centro (che rappresenta la neutralità rispetto a ciascuno dei due atteggiamenti) e comunque collocato in uno dei quattro quadranti (nel nostro esempio quello dei laici progressisti in alto a destra, quello dei laici conservatori in basso a destra, quello dei clericali progressisti in alto a sinistra e quello dei clericali conservatori in basso a sinistra). Il modo di raggrupparsi dei soggetti sulla base di questa rappresentazione offre un'informazione combinata che il grafico rende particolarmente intelligibile e analizzabile.
Un esempio classico di questa metodologia si trova nel saggio di Carlo M. Cipolla sulla stupidità[2]. Cipolla pone sull'asse delle x la misura del guadagno che un individuo può trarre dalle proprie azioni, e sull'asse delle y la misura del guadagno che altre persone possono trarre dalle azioni dell'individuo in esame. In questa rappresentazione i quattro quadranti rappresenteranno (in misura esaltata dalla distanza dal centro) quattro tipi psicologici distinti: lo sprovveduto, l'intelligente (che si avvantaggia procurando un beneficio anche ad altri), il bandito e infine, appunto, lo stupido, che è l'individuo che produce un danno agli altri danneggiando anche se stesso: una definizione straordinariamente puntuale ed efficace e che permette anche di "misurare" il grado di stupidità.
Traiamo un altro esempio molto stimolante da un saggio di C. Formenti[3] sull'analisi dell'impatto politico-sociale dei nuovi *media*. Collocando sull'asse delle x lo spettro degli atteggiamenti che vanno dall'individualismo al collettivismo, e sull'asse delle y le posizioni di filosofia politica che vanno dalla democrazia classica alla post-democrazia diretta e mediatica, Formenti arriva a classificare con grande accuratezza, individuando in modo significativo contiguità e distanze, gli atteggiamenti di un gran numero di analisti e di scuole di pensiero nei confronti del rapporto tra nuovi media e nuove forme della politica. Ma gli esempi potrebbero moltiplicarsi, e ognuno è intitolato a "inventare" le proprie coordinate fondamentali: l'efficacia della scelta è misurata dalla qualità dell'analisi che ne risulta.

[2] Carlo M. Cipolla, *Le leggi fondamentali della stupidità umana*, in "Allegro ma non troppo", il Mulino, Bologna 1988
[3] Carlo Formenti, *Cybersoviet*, Cortina, Milano 2008

7. Sistemi privi di scala

Molti fenomeni naturali e sociali sono caratterizzati dalla presenza di un a"scala intrinseca".
Parliamo di scala intrinseca quando le distribuzioni dei valori numerici delle variabili quantitative tipiche dei fenomeni studiati tendono a concentrarsi in un intervallo abbastanza ben definito, mentre i valori lontani da tale intervallo sono estremamente improbabili, se non addirittura impossibili.
Un semplicissimo esempio di questo concetto è dato dalla distribuzione delle altezze degli esseri umani: quando diciamo che la scala intrinseca del corpo umano è il metro, intendiamo sottolineare il fatto che non esistono individui alti dieci metri o dieci centimetri. e che la distribuzione delle altezze di una popolazione sufficientemente numerosa tenderà ad assumere la forma di una curva a campana, con una media certamente compresa (per popolazioni adulte) tra 1,50 e 2,00 metri, e una deviazione standard di poche decine di centimetri. Il rapporto tra la massima e la minima altezza storicamente registrate per individui adulti è un numero inferiore a 5.
Dovrebbe essere già chiaro a questo punto che gli esempi potrebbero moltiplicarsi a volontà, e che il loro studio dal punto di vista statistico (tema sul quale torneremo) trova naturalmente nei concetti di media e di varianza i primi e più semplici strumenti d'indagine.
Esistono tuttavia, sia in natura che nell'ambito delle scienze umane, anche numerosi fenomeni che non possiedono affatto una scala intrinseca. La caratteristica più evidente delle distribuzioni che descrivono il comportamento quantitativo dei sistemi privi di scala sta nel fatto che, non esistendo valori intrinsecamente favoriti o proibiti, le distribuzioni non si concentrano in intervalli particolari, ma, fatta salva la rarissima e peculiare eventualità in cui tutti i valori possibili siano equiprobabili (presente in pratica soltanto in certi fenomeni matematici), nelle situazioni realistiche ci saranno valori "grandi" che compaiono con scarsa frequenza e valori "piccoli" che compaiono con molto maggior frequenza, e in mancanza di una scala intrinseca la riduzione della frequenza al crescere del valore tenderà ad avvenire con la stessa regolarità in tutti gli intervalli di valori, per cui a una data variazione percentuale (positiva) del valore corrisponderà in ogni intervallo una costante e determinata variazione percentuale (negativa) della frequenza. Questo tipo di relazione si traduce in una proprietà matematica ben precisa delle distribuzioni, che prendono la forma di "leggi di scala" o "leggi di potenza". Abbiamo già indicato il fatto che la più semplice rappresentazione di una legge di potenza è data in forma grafica adottando una doppia scala logaritmica, per cui la curva rappresentativa della legge diventa una retta, la cui inclinazione misura l'esponente della legge stessa. Notiamo fin d'ora due caratteristiche fondamentali delle leggi di potenza, originate dalle loro proprietà matematiche ma strettamente collegate alla natura e al significato profondo di tali leggi.
In primo luogo, nel caso di una distribuzione basata su una legge di potenza è certamente ancora possibile definire il valor medio della distribuzione, ma poiché non esiste una scala intrinseca neanche tale valor medio ha un significato intrinseco: esso rappresenta semplicemente la scala empirica della particolare manifestazione concreta del fenomeno in esame, e differenti istanze dello stesso fenomeno potranno quindi essere caratterizzate da valori medi anche molto diversi tra loro, mentre, come vedremo meglio in seguito, ci si aspetta che l'esponente della legge sia, per tutte le manifestazioni dello stesso fenomeno, sempre sostanzialmente lo stesso.
La seconda proprietà fondamentale delle distribuzioni prive di scala è la perdita di significato del concetto di varianza e di deviazione standard, che per la gran parte di queste distribuzioni risulta addirittura impossibile, nella formulazione astratta della legge, definire matematicamente, in quanto la definizione conduce a integrali divergenti. Per quanto si è detto in precedenza sul fatto che in un sistema privo di scala nessun valore è intrinsecamente impossibile o proibito (fatte salve le ovvie limitazioni pratiche della realtà empirica) questo risultato non dovrebbe stupire più di tanto, ma certamente ci deve mettere in guardia nei confronti di un qualunque uso "ingenuo" delle statistiche per questo tipo di sistemi, per i quali dovrebbe essere ormai evidente che la nozione di "normalità" statistica (anche nel senso indicato dal teorema del limite centrale) non ha alcun significato.

Anche se i fenomeni caratterizzati dall'assenza di una scala intrinseca rivestono spesso un grande interesse teorico e/o pratico, il loro riconoscimento e il loro studio hanno avuto inizio soltanto in tempo relativamente recenti, e spesso soltanto l'applicazione delle metodologie più moderne della fisica ne ha illuminato la natura profonda e "spiegato" le dinamiche.
Il primo esempio storicamente noto di individuazione di una "legge di scala" in un fenomeno storico-sociale è la legge di Pareto[4], consistente nell'affermazione che la distribuzione del reddito annuale tra gli individui di una popolazione obbedisce a una legge di potenza. Questa legge è spesso volgarizzata con la formula 80-20, per cui il 20% della popolazione percepirebbe l'80% del reddito totale, ed è stata generalizzata con la stessa formula a numerosi altri contesti, di cui citiamo soltanto alcuni esempi, tratti spesso, ma non solamente, dal contesto economico:

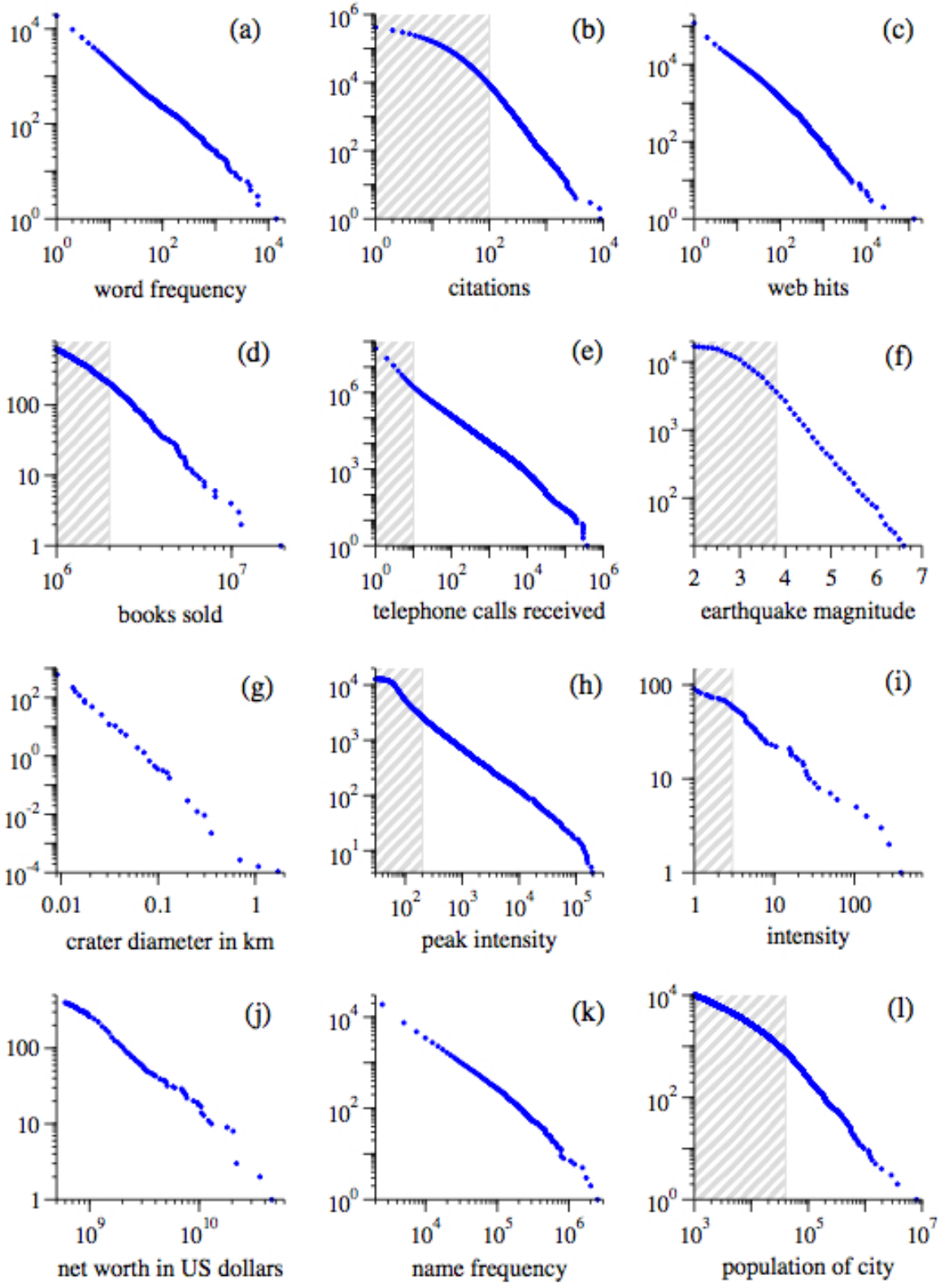
- l'80% delle vendite è effettuato dal 20% dei venditori
- l'80% dei ricavi di treni e aerei è generato dal 20% delle rotte
- l'80% del deficit della Sanità è prodotto dal 20% delle ASL
- l'80% dei reclami proviene dal 20% dei clienti
- l'80% del tempo di esecuzione di un programma è utilizzato dal 20% delle istruzioni
- l'80% degli utilizzi di un applicativo riguarda il 20% delle sue funzionalità
- l'80% dei visitatori di un sito vede solo il 20% delle pagine
-

Nel contesto delle scienze umane le leggi di scala sono quindi spesso dette "distribuzioni di Pareto", ma sono anche note come "leggi di Zipf" in quanto il secondo importante esempio storico di tali leggi in un contesto sociologico, individuato forse per la prima volta da Auerbach[5] ma reso famoso appunto da Zipf[6], è la distribuzione delle città secondo il numero dei loro abitanti. È ovvio che esistono poche città con moltissimi abitanti, moltissime città e villaggi con pochi abitanti, che intercorrono molti ordini di grandezza tra la popolazione delle città più grandi (che si misura in decine di milioni di individui) e quella dei più piccoli villaggi (che si misura in poche unità) e che non esiste una scala intrinseca per questo fenomeno. Non è tuttavia ovvio *a priori* ciò che si verifica empiricamente, ossia il fatto che la distribuzione del numero delle città che hanno dato numero di abitanti in funzione del numero stesso, se rappresentata su scala doppiamente logaritmica, mostra un chiaro andamento lineare. Lo stesso Zipf[7], che era un linguista, riprendendo un'osservazione di Estoup[8] produsse un terzo esempio molto conosciuto di legge di scala nei fenomeni sociali, la distribuzione della frequenza con cui le parole compaiono nei testi scritti.
L'elenco dei fenomeni per i quali è stata trovata evidenza empirica (e spesso motivazione teorica) per l'esistenza di una legge di scala è ormai assai vasto, e la lista si allunga di giorno in giorno.
A puro scopo esemplificativo menzioniamo, nell'ambito delle scienze della natura, l'intensità dei terremoti, le dimensioni dei crateri lunari, il numero delle specie biologiche nei differenti ordini animali e vegetali, l'entità delle estinzioni di massa, la portata dei fiumi, gli incendi nei boschi, la formazione delle valanghe.
Ma ciò che più ci interessa in questa sede è il fatto che un grande numero di leggi di questo genere è stato individuato anche nel contesto delle scienze umane. Ricordiamo qui brevemente, salvo poi analizzare meglio nel seguito alcuni dei casi più interessanti, diversi esempi tratti dall'informatica (distribuzione delle dimensioni dei *files*, del numero degli *hits* delle pagine *web,* teoria delle reti), dall'onomastica (frequenza dei nomi propri di persona, distribuzione dei cognomi), dall'economia (fluttuazioni nell'andamento della borsa, dimensioni delle imprese, numero delle copie vendute di un libro, di un disco o di altri prodotti), dalla sociologia della ricerca (numero di articoli scritti, numero delle citazioni) e dalla realtà sociale in genere (ingorghi del traffico, scoppi delle guerre, tendenze della moda).

---

[4] V. Pareto, *Cours d'Economie Politique*, Droz, Généve 1896
[5] F. Auerbach, *Das Gesetz der Bevölkerungskonzentration*, in Petermanns Geographische Mitteilungen 59, 74-76 (1913)
[6] G. K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Reading, MA 1949
[7] G. K. Zipf, *op. cit.*
[8] J. B. Estoup, *Gammes stenographiques*, Institut Stenographique de france, Paris 1916

8. Esempi empirici di leggi di scala



Nei grafici allegati (tutti in scala doppiamente logaritmica), tratti dall'interessante articolo di rivista di M.E.J. Newman[9], sono riportati i dati empirici relativi a una dozzina di fenomeni (naturali e sociali) che vale la pena di esaminare con qualche dettaglio per comprendere meglio l'ubiquità, ma anche la variabilità fenomenologica, delle dinamiche che stiamo studiando.
In tutti i grafici è stata utilizzata per semplicità di illustrazione la "distribuzione cumulativa", che si ottiene dalla distribuzione p(x) sommando, per ogni valore di x, tutti i valori della distribuzione maggiori o uguali a x. In questo modo si eliminano in gran parte le forti fluttuazioni legate al numero piccolo (e quindi largamente casuale) degli elementi che si trovano empiricamente presenti nella distribuzione per valori molto grandi di x, senza perdere con questo la principale caratteristica strutturale (ovvero il comportamento a potenza) della distribuzione stessa.
(a) si riferisce alla legge di Zipf propriamente detta, ossia la distribuzione della frequenza con cui le parole sono usate in un testo scritto. La variabile x rappresenta il numero di volte che un certo vocabolo compare nel testo, e la variabile y indica il numero dei vocaboli che compaiono x volte. Nell'esempio riportato l'analisi è effettuata sul testo inglese di *Moby Dick* di Melville, e i vocaboli più usati (che nel grafico figurano a partire da destra) sono in ordine di frequenza *the, of, and, a* e *to*; distribuzioni simili sono state osservate per molti altri testi e in altre lingue.
(b) riporta la frequenza delle citazioni di lavori scientifici, per la quale la distribuzione secondo una legge di potenza fu osservata per la prima volta da Price[10] nel 1965; i dati sono tratti dallo *Science Citation Index* e riguardano lavori pubblicati nel 1981, le cui citazioni sono state contate da Redner[11] nel 1997.
(c) rappresenta la distribuzione cumulativa del numero di *hits* ricevuti da siti *web* in un singolo giorno (1 dicembre 1997) da un sottoinsieme di 60.000 utenti del servizio Internet di A.O.L..[12]
(d) è la distribuzione del numero totale di copie vendute negli U.S.A. per i 633 *bestseller* (esclusa la Bibbia) che hanno venduto più di due milioni di copie tra il 1895 e il 1965.[13]
(e) è la distribuzione del numero di chiamate interurbane ricevute da 51 milioni di clienti U.S.A. di AT&T in un singolo giorno. Distribuzioni simili si riscontrano per i messaggi *e-mail*.
(f) mostra la distribuzione cumulativa della magnitudine (Richter) dei terremoti in California tra il gennaio 1910 e il maggio 1992, secondo il *Berkeley Earthquake Catalog*: l'asse orizzontale ha una scala lineare in quanto la scala Richter è già di per sé una scala logaritmica.
(g) è la distribuzione dei crateri lunari (per Kmq) secondo il loro diametro.
(h) è l'intensità delle eruzioni solari (tra il 1980 e il 1989).
(i) è la distribuzione dell'intensità di 119 conflitti tra il 1816 e il 1980, misurata mediante il conteggio del numero dei morti in battaglia rapportato al totale delle popolazioni coinvolte.[14]
(j) è la distribuzione della ricchezza degli individui più ricchi degli U.S.A. nell'ottobre 2003.
(k) è la distribuzione in frequenza degli 89.000 cognomi più comuni negli U.S.A. nel 1990.
(l) è la distribuzione cumulativa della dimensione delle popolazioni delle città degli U.S.A nel 2000 secondo l'*US Census Bureau.*

Un'utile e ampia presentazione, di tipo divulgativo, di molti fenomeni naturali e sociali per i quali è stata individuata, e talvolta anche giustificata sulla base di modelli fisico-matematici, l'esistenza empirica di leggi di scala è fornita dai saggi di Buchanan.[15]

[9] M.E..J. Newman, *Power laws, Pareto distributions and Zipf's law,* Contemporary Physics 46, 323-351 (2005)
[10] D. J.. de S. Price, *Networks of scientific papers*, Science 149, 510-515 (1965)
[11] S. Redner, *How popular is your paper? An empirical study of the citation distribution*, Eur.Phys J.B4, 131-134(1998)
[12] L.A. Adamic and B.A. Huberman, *The nature of markets in the WWW*, Quart. Jour.. of E.Commerce 1, 512 (2000)
[13] A. P. Hackett, *70 Years of Best Sellers, 1895-1965*, R.R. Bowker Co., New York 1967
[14] M. Small and J.D. Singer, *Resort to Arms: International and Civil Wars, 1816-1980*, Sage Pub., Beverley Hills 1982
[15] M. Buchanan, *Ubiquità,* Mondadori, Milano 2001; M. Buchanan, *Nexus*, Mondadori, Milano 2003

9. Origine dinamica delle leggi di scala e teoria delle reti

Esistono differenti meccanismi dinamici atti a giustificare la comparsa di una legge di scala. Alcuni di questi meccanismi sono descritti nell'articolo di Newman[16], in cui si insiste in particolare sulla validità, in numerosi casi, delle spiegazioni basate sul cosiddetto "processo di Jule", la cui forma più generale è la seguente. Supponiamo di avere un sistema composto da una collezione di oggetti ("generi"), con la possibilità che possano talvolta comparire nuovi oggetti. Tutti gli oggetti sono caratterizzati dal valore numerico di una particolare proprietà (il numero delle "specie"), e ci si chiede sotto quali ipotesi la distribuzione del numero delle "specie" possa obbedire a una legge di potenza. Si dimostra che ciò avviene se, nell'intervallo tra comparsa di un nuovo genere e quella del successivo, un certo numero di nuove specie è aggiunto all'intero sistema, attribuendo le nuove specie a ciascun genere in proporzione al numero di specie che il genere già possiede.

Un altro meccanismo dinamico oggi largamente invocato per spiegare la comparsa di leggi di scala nei più vari e diversi campi della scienza è la criticità auto-organizzata, un concetto inizialmente proposto da Bak, Tang e Wiesenfeld[17] nel 1987. I fenomeni critici (transizioni di fase di seconda specie) sono noti da molto tempo in fisica statistica, ed è anche ben noto che tali fenomeni sono caratterizzati dalla comparsa di leggi di scala nelle relazioni che intercorrono tra le diverse variabili termodinamiche: gli esponenti che compaiono in queste leggi di scala sono appunto detti "esponenti critici" e la moderna teoria del gruppo di rinormalizzazione di Wilson[18] offre una solida base teorica alla comparsa di tali leggi e tecniche di calcolo del valore numerico degli esponenti stessi. Tuttavia nei sistemi fisici la comparsa della criticità (e quindi delle leggi di scala) è strettamente legata al fatto che determinati parametri esterni (in particolare la temperatura) assumano un valore preciso. Nello spazio dei parametri della teoria il "punto critico" è un punto da cui è "facile" allontanarsi e quindi difficilmente il sistema muove "spontaneamente" verso di esso.
Esiste tuttavia la possibilità teorica che per certi sistemi dinamici esista un punto critico "attrattivo" (ossia tale per cui qualunque trasformazione del sistema tenda a farlo evolvere verso la criticità). In tal caso sarebbe del tutto "normale" per un sistema di questo tipo trovarsi quasi sempre in condizioni di criticità, e le relazioni tra le variabili del sistema, in condizioni di equilibrio dinamico, assumerebbero quindi la forma di leggi di scala, senza la necessità che i parametri esterni abbiano valori particolari (e in quanto tali difficilmente giustificabili in circostanze "casuali").

L'idea di criticità auto-organizzata si colloca all'interno di un più vasto ambito di ricerca che riguarda lo studio e la teoria dei "sistemi complessi", definiti genericamente come sistemi composti da parti interconnesse che esibiscono nel loro complesso proprietà non riconducibili alle proprietà individuali delle parti componenti, e quindi necessariamente derivanti dal carattere non lineare dell'interconnessione tra le parti. Dovrebbe essere assolutamente chiaro che (oltre a numerosi fenomeni fisici) la maggior parte dei fenomeni biologici e sociali rientra nella definizione di "sistema complesso", ma lo studio quantitativo (e spesso anche qualitativo) dei sistemi complessi è in generale assai difficile, e quindi l'aver ricondotto i fenomeni sociali a tale tipologia non significa che se ne abbia in ogni caso una miglior comprensione: ad esempio capiamo ancora ben poco dei fenomeni caotici, a partire dalla turbolenza.
I sistemi caratterizzati da criticità auto-organizzata, malgrado la loro significativa diffusione e il loro notevole interesse teorico e pratico, rappresentano comunque un sottoinsieme molto particolare dell'insieme dei sistemi complessi.

---

[16] M.E.J. Newman*, op. cit.*
[17] P. Bak, C. Tang and K. Wiesenfeld, *Self-organized criticality: an explanation of 1/f noise*, Physical Review Letters 59 (1987) 381-384
[18] K.G. Wilson and J. Kogut, *The renormalization group and the epsilon expansion*, Physics Reports 12 (1974) 75-199

Una delle caratteristiche più interessanti dei fenomeni descritti da una legge di scala è la proprietà di autosimilarità, ovvero il fatto che ogni volta che ci si restringe a una parte (non troppo piccola) del sistema in esame si ritrovano le stesse caratteristiche (in particolare le stesse distribuzioni) possedute dal sistema nel suo complesso: tale proprietà è caratteristica anche delle strutture frattali.

Un importantissimo esempio dell'utilizzo delle idee e delle nozioni che abbiamo fin qui presentato è la teoria delle reti sviluppata principalmente da Barabási[19] e dai suoi collaboratori, e applicabile allo studio di numerose situazioni reali, dalle reti neurali a quelle sociali. La teoria delle reti è per molti aspetti un'evoluzione della teoria dei grafi casuali. Numerosi aspetti qualitativi e quantitativi della teoria sono oggi confrontabili abbastanza facilmente con l'evidenza empirica grazie alla possibilità di analizzare il comportamento del *World Wide Web* e delle sue parti.
Uno degli aspetti più interessanti di questa teoria consiste nell'avere messo in evidenza l'importanza dei cosiddetti "legami deboli" ai fini della generazione di comportamenti globalmente coerenti in sistemi costituiti da molte parti che in genere interagiscono soltanto localmente l'una con l'altra. Un legame debole è un "ponte" tra sottoinsiemi del sistema che, in quanto tali, sono connessi anche fortemente al loro interno ma hanno solo un piccolo numero di connessioni con altri sottoinsiemi. Si dimostra tuttavia che è sufficiente un numero anche molto limitato di tali connessioni (purché non si vada al di sotto di una soglia "critica") affinché l'informazione necessaria per lo stabilirsi di un comportamento globale possa diffondersi nell'intero sistema.
Questo principio è brillantemente illustrato dalla cosiddetta "teoria dei sei gradi di separazione", che asserisce che, prese due persone a caso nell'intero pianeta, è (pressoché) sempre possibile stabilire tra loro una connessione basata su relazioni umane dirette che non richiede più di sei passaggi per giungere da un individuo all'altro (è il cosiddetto effetto *smallworld*). Molti *test* sperimentali hanno confermato questa stima della connessione della rete sociale, mentre un'applicazione dello stesso tipo di analisi al *World Wide Web* ha mostrato che il numero di passaggi massimo richiesto per andare da un qualunque documento presente nella Rete a un qualunque altro documento è circa pari a diciannove: si tratta di un valore abbastanza elevato (nella maggior parte delle reti di altro genere che sono state studiate tale valore oscilla tra i due e i quattordici gradi di separazione) ma comunque piccolo rispetto alle dimensioni della Rete stessa, e ai miliardi di nodi che ne fanno parte.
Oltre la presenza di legami deboli, un altro aspetto cruciale per la comparsa dell'effetto *smallworld* è la presenza di "connettori", ovvero di nodi della rete caratterizzati da un grande numero di connessioni ad altri nodi. I connettori funzionano come gli *hub* aeroportuali: è oggi possibile andare da un qualunque piccolo aeroporto a un altro qualunque piccolo aeroporto raggiungendo uno *hub* connesso alla località di partenza e passando da questo (nella peggiore delle ipotesi con uno scalo intermedio in un altro *hub*) a uno *hub* connesso con la località che si vuole raggiungere.
La presenza di connettori in una rete è a sua volta strettamente legata all'invarianza di scala del sistema (che nelle reti si manifesta in modo chiaro nell'autosimilarità, per cui ogni porzione di rete ha le stesse proprietà topologiche dell'intera rete). Notiamo infatti che la distribuzione del numero di connessioni possedute da ciascun nodo, non avendo una scala caratteristica, non è una gaussiana ma è governata da una legge di potenza, e ciò significa che la probabilità di trovare nodi con un grandissimo numero di connessioni, pur essendo ovviamente e rapidamente decrescente, tuttavia non è esponenzialmente soppressa, e anche i connettori conservano una ragionevole probabilità di formarsi (ad esempio con un processo di Jule) non appena il numero dei nodi risulti abbastanza elevato.

[19] A.-L. Barabási, *Link. La scienza delle reti*, Einaudi, Torino 2004

10. Universalità

Una delle più significative proprietà che si possono manifestare nei sistemi privi di una scala è l'universalità. Con questo nome si designa la caratteristica per cui differenti distribuzioni empiriche possono ricondursi alla stessa forma matematica (e i loro grafici risultano quindi sovrapponibili), indipendentemente dal valore preso dai parametri legati al contesto specifico (come ad esempio la dimensione del campione esaminato, o l'ambito territoriale cui il campione appartiene), alla sola condizione di adottare per ciascuna delle distribuzioni un'opportuna unità di misura ossia, in altri termini, a condizione di scegliere per ciascun caso particolare una "scala" specifica al contesto, sfruttando l'invarianza di scala che ci assicura che tale scelta non è predeterminata dalla natura del fenomeno in esame.
Notiamo (e lo capiremo meglio considerando esempi specifici) che nel caso delle leggi di potenza del tipo già discusso, che può essere rappresentato in generale dalla relazione

$$\mathrm{Log}(y) = \mathrm{Log}(A) + b\ \mathrm{Log}(x),$$

è sempre possibile eliminare il termine Log(A) semplicemente introducendo come nuova variabile la quantità "riscalata" y' = y/A, per cui la relazione diventa Log(y') = b Log(x), e quindi possiamo asserire di essere in presenza di universalità quando, studiando differenti manifestazioni dello stesso fenomeno, o fenomeni in qualche modo analoghi, troviamo lo stesso valore dell'esponente b.
La teoria dei fenomeni "universali", anche nei casi in cui può essere formulata a partire da qualche principio generale, o comunque cerca di fondarsi su modelli dinamici (non limitandosi a una pura e semplice rappresentazione dei dati fenomenologici), non ha quindi per obiettivo la determinazione dei parametri relativi al contesto particolare, ma in genere cerca soltanto di "predire" gli esponenti che compaiono nelle leggi di scala.
Consideriamo la più semplice e ovvia delle leggi di scala, quella relativa alle aree delle figure geometriche simili. Sappiamo che per ogni cerchio l'area è proporzionale al quadrato del raggio, per ogni quadrato essa è uguale al quadrato del lato, relazioni analoghe tra lato e area valgono per tutti i poligoni regolari. L'universalità in questo caso si traduce nella relazione più generale "per tutte le figure geometriche tra loro simili (e non soltanto per poligoni regolari) l'area S della figura è direttamente proporzionale al quadrato di una lunghezza L caratteristica della figura stessa", ovvero

$$S = K*L^2$$

dove K è una costante dipendente dalla forma della figura e dalla caratteristica prescelta per misurare L. Come si vede la legge di scala non offre indicazioni su K, che dipende dal "contesto", ma per l'esponente vale sempre b=2, indipendentemente da ogni altra proprietà delle figure simili.

È molto importante capire che l'universalità non è una proprietà di tutte le leggi di potenza: se fenomeni analoghi e riconducibili a una dinamica comune, pur obbedendo a una legge di potenza, non presentano lo stesso esponente, non possiamo parlare di universalità in senso stretto, a meno che non sia il meccanismo dinamico stesso a giustificare la variabilità dell'esponente, legandone il valore alla variabilità di un qualche parametro fondamentale del processo: in tal caso l'universalità è per così dire "ristretta" a quei fenomeni che, oltre a essere governati da una dinamica simile, sono anche caratterizzati da un comune valore del parametro che governa l'esponente.

È anche molto importante capire che le leggi di potenza non sono certo l'unico tipo di relazione matematica tra variabili compatibile con l'invarianza di scala e con il concetto di universalità.
Per essere concreti consideriamo il caso della distribuzione detta "lognormale", che può essere scritta nella forma generale

$$\mathrm{Log}\,(y) = a + b\,[\mathrm{Log}\,(x) + c]^2$$

Notiamo subito che, utilizzando le variabili riscalate $y' = y/\exp(a)$ e $x' = x*\exp(c)$, possiamo di nuovo trovare una forma "universale" della distribuzione se i casi in esame sono tutti caratterizzati dallo stesso valore di b.

Esistono numerosi esempi di distribuzioni lognormali nell'ambito delle scienze naturali, ma nel contesto delle scienze sociali un esempio molto interessante è fornito dall'analisi della distribuzione delle citazioni degli articoli scientifici effettuata di recente da Radicchi, Fortunato e Castellano[20]. Radicchi e collaboratori hanno mostrato che, normalizzando per ciascuna disciplina il numero delle citazioni ricevute da ciascun articolo al numero medio delle citazioni per articolo della disciplina stessa, la distribuzione delle citazioni, ovvero la percentuale di articoli che hanno ricevuto esattamente quel numero (normalizzato) di citazioni, assume una forma universale e indipendente dalla disciplina, e tale forma corrisponde appunto a quella di una distribuzione lognormale.
L'accuratezza del risultato ottenuto, può essere facilmente verificata esaminando la figura allegata, tratta appunto dall'articolo succitato: si noti nella legenda la grande varietà delle discipline per le quali l'universalità è stata empiricamente osservata.

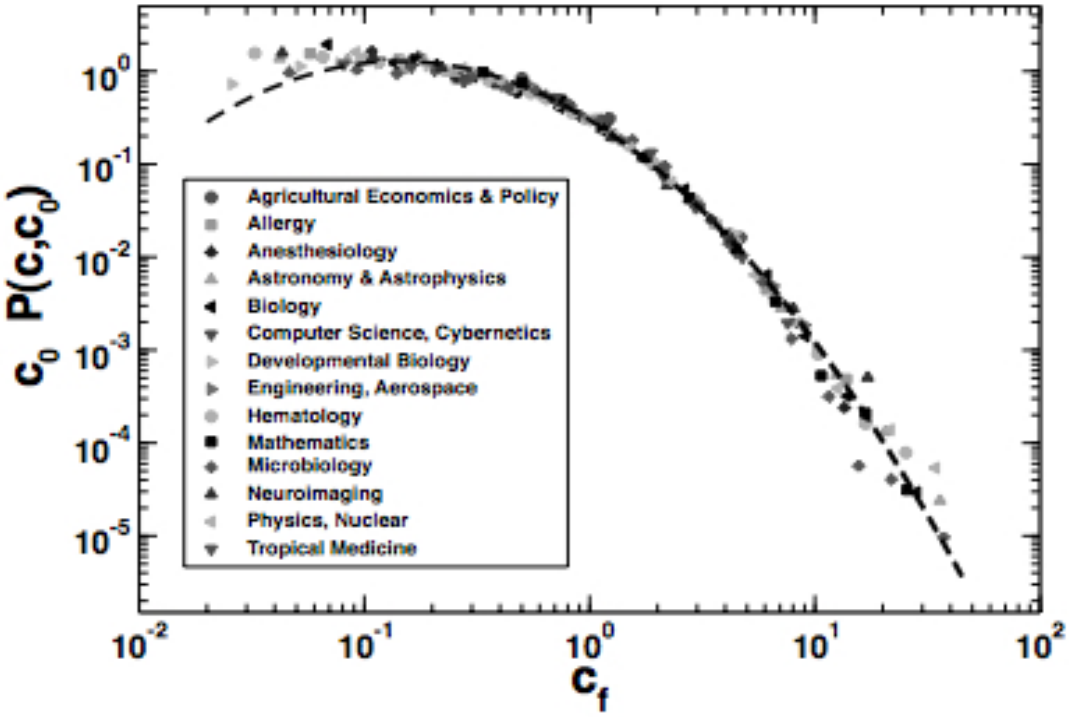


Fig. 2. Rescaled probability distribution $c_0\ P(c,c_0)$ of the relative indicator $c_f = c/c_0$, showing that the universal scaling holds for all scientific disciplines considered (see Table 1). The dashed line is a lognormal fit with $\sigma^2 = 1.3$.

Un altro esempio molto illuminante di universalità non legata a una legge di potenza è fornito dalla teoria statistica della consanguineità tra gli antenati. Le proprietà statistiche degli alberi genealogici sono state studiate per la prima volta nel 1999 da Derrida, Manrubia e Zanette[21]. La distribuzione delle ripetizioni di antenati, dopo un numero sufficiente di generazioni, acquista una forma universale, sulla base della quale è anche possibile calcolare la frazione di individui della più antica generazione considerata che risulta mediamente assente in un dato albero genealogico (pari al 20,3% della popolazione), e l'esponente della legge di potenza che governa l'andamento della distribuzione per bassi valori del numero delle ripetizioni.

---

[20] S. Radicchi, S. Fortunato, C. Castellano, *Universality of citation distributions: toward an objective measure of scientific impact*, Proceedings National Academy of Science USA 105 (2008) 17268-17273
[21] B. Derrida, S.C. Manrubia and D.H. Zanette, *Statistical Properties of Genealogical Trees,* Phys. Rev. Lett. 82 (1999) 1987-1990; *Distribution of repetitions of ancestors in genealogical trees*, Physica A 281 (2000) 1-16

L'analisi di Derrida e collaboratori è basata sulla simulazione numerica e lo studio analitico di un modello teorico di popolazione chiusa con riproduzione sessuale e generazioni non sovrapposte. L'unica verifica sperimentale di quest'analisi finora presentata consiste nello studio, effettuato dagli stessi autori, dell'albero genealogico di re Edoardo III d'Inghilterra (1312-1377), ricostruito per dieci generazioni. Sarebbe molto interessante analizzare un campione molto più vasto di genealogie, stabilite per individui vissuti in tempi significativamente differenti (dal Medioevo ai nostri giorni).

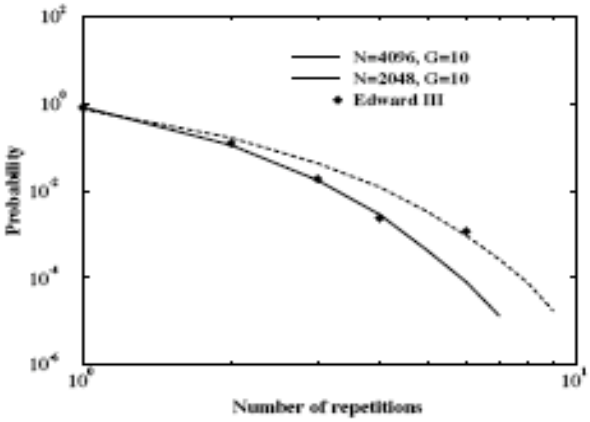
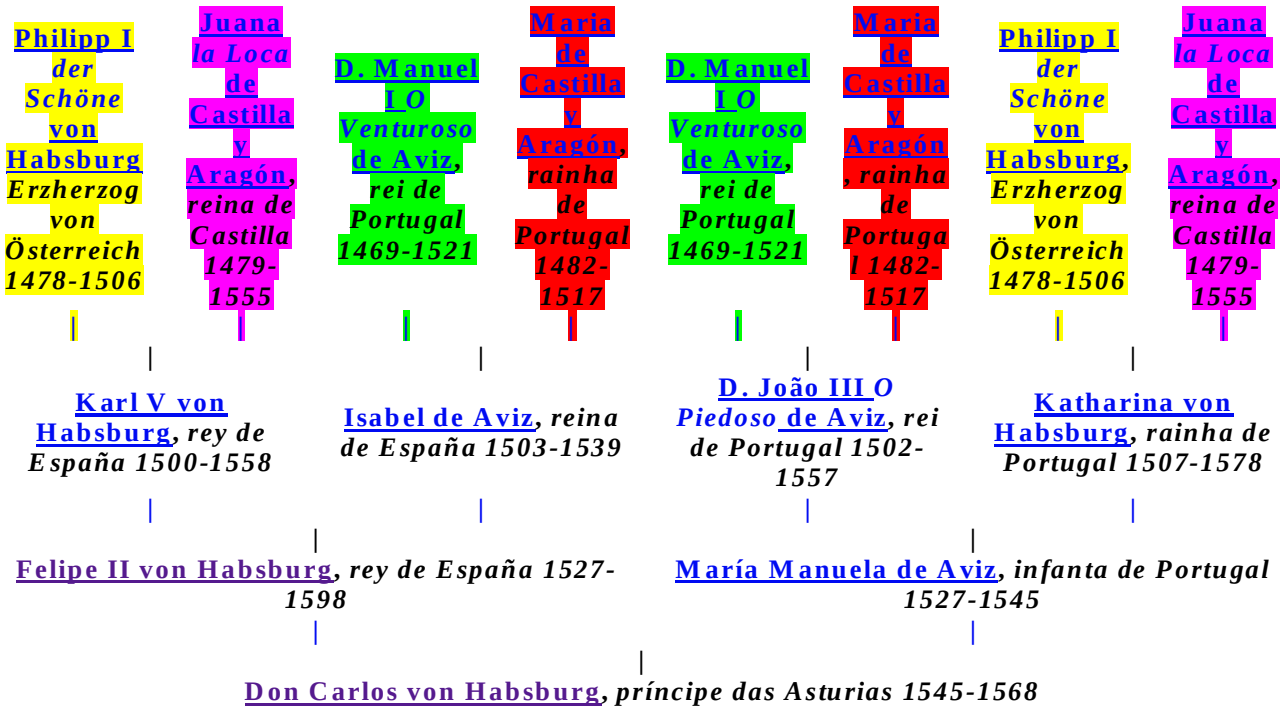


FIG. 1. Probability of ancestor repetitions in the genealogical tree of the king Edward III [5]. The continuous and dashed lines represent the results of simulations of $F(r)$ in a closed population with 2^{11} and 2^{12} individuals for our model. Averages have been performed over the ten first generations of 10^3 independent trees.

Per chiarire quanto possa essere importante il fenomeno della ripetizione degli antenati anche nel corso di poche generazioni un esempio molto significativo è quello di Don Carlos d'Asburgo, nipote dell'imperatore Carlo V. Alla sesta generazione i suoi antenati erano 12 (contro i 32 teorici), e alla decima generazione soltanto 83 (contro i 512 che si sarebbero avuti in assenza di ripetizioni).



Un filone di ricerche strettamente connesso al precedente, almeno in linea di principio, anche se finora mancano studi volti a mettere in connessione diretta i due tipi di analisi, è quello relativo alla identificazione (o meglio alla datazione) del cosiddetto "antenato comune più recente" (*most recent common ancestor,* in sigla MRCA). Si tratta in sintesi di stimare quando potrebbe essere vissuto l'ultimo individuo di cui tutti gli esseri umani attualmente viventi sono discendenti. Fino a non molto tempo fa si riteneva di dover risalire fino al Paleolitico Superiore, ma recenti simulazioni numeriche applicate a modelli più sofisticati, che tengono conto della grande diffusione geografica della specie umana ma anche delle dinamiche migratorie degli ultimi secoli e dei loro effetti sul mescolamento delle popolazioni, hanno portato ad abbassare considerevolmente questa data e a stimare[22] la collocazione temporale del MRCA tra il terzo e il primo millennio avanti Cristo, mentre restringendo l'analisi alla sola popolazione europea occidentale si scenderebbe addirittura al 1000 dopo Cristo. Trattandosi comunque in ogni caso di analisi statistiche basate su modelli teorici, non è possibile in linea di principio escludere la possibilità che esista tuttora un qualche gruppo umano vissuto in condizioni di particolare isolamento e che quindi un MRCA che includa questo gruppo abbia una collocazione temporale molto più antica di quella congetturata. Soltanto *test* genetici estesi all'intera popolazione umana potrebbero permettere una ricostruzione "sperimentale" della intera storia genetica della nostra specie.

Alcuni commenti sono necessari ad evitare fraintendimenti del concetto di MRCA. In primo luogo deve essere chiaro che i meccanismi della riproduzione sessuata (per cui ogni individuo eredita da ciascuno dei due genitori soltanto metà del patrimonio genetico) portano a una rapida diluizione del patrimonio genetico di un singolo antenato, e dopo un numero non molto elevato di generazioni (basti pensare che per ogni millennio ve ne sono più di trenta!) è perfettamente possibile che di un particolare antenato (e quindi anche del MRCA) non sia rimasta traccia genetica in gran parte della popolazione (o anche in tutta l'umanità). Eccezioni a questa regola sono rappresentate da quei geni che si trasmettono per via puramente maschile (come quelli presenti nel cromosoma Y) e sono quindi presenti in tutti i discendenti maschi di un antenato comune, o per via puramente femminile (come il DNA mitocondriale) e sono quindi presenti in tutte le discendenti femmine di un'antenata comune. Su questa base sono state introdotte le nozioni di "Adamo Y-cromosomico" e di "Eva mitocondriale", che sono antenati comuni di tutti gli esseri umani oggi viventi, ma che per definizione si collocano in un passato molto più remoto: si stima che "Adamo" sia vissuto circa 60.000 anni fa, mentre "Eva mitocondriale" (la cosiddetta "Eva africana") sarebbe vissuta circa 140.000 anni fa. Non si deve nemmeno pensare che l'esistenza di questi antenati comuni comporti che la popolazione umana, all'epoca in cui essi vissero, fosse particolarmente esigua, o priva di discendenza. Nell'esempio di "Adamo" l'unica ipotesi formalmente non confutabile è che tra i discendenti di tutti gli altri maschi suoi contemporanei, dopo un periodo (probabilmente) breve, si siano riprodotte solo le femmine (fatto peraltro non statisticamente improbabile e che si trova all'origine del fenomeno dell'estinzione dei cognomi, che discuteremo in seguito). Dobbiamo quindi presumere che il MRCA avesse molti contemporanei di entrambi i sessi, alcuni dei quali non hanno discendenti viventi, mentre altri hanno ancora un numero più o meno grande di discendenti, ma nessuno dei quali, a parte il MRCA, è antenato di tutti gli esseri umani attualmente viventi.

Per questo motivo è stato introdotto, e studiato, il concetto di "punto degli antenati identici" (*identical ancestors point*), definito come il momento temporale in cui tutti coloro che erano a quel tempo viventi possono essere divisi in due soli gruppi: quelli che oggi non hanno più nessun discendente e quelli che sono antenati comuni di tutti gli esseri umani attuali.

Dovrebbe essere chiaro a questo punto che tutti questi studi, pur vendo una forte componente quantitativa, e richiedendo un uso intensivo di modelli matematici e di simulazioni numeriche, si trovano al punto d'incontro, e richiedono il contributo interdisciplinare, di un grande numero di discipline diverse, che vanno dalla genetica delle popolazioni alla demografia storica, dalla geografia antropica alla storia dei fenomeni migratori.

---

[22] D.L.T. Rohde, S. Olson, J.T. Chang, *Modelling the recent common ancestry of all living humans*, Nature 431 (2004) 562-566

11. La distribuzione dei cognomi e gli studi di genetica

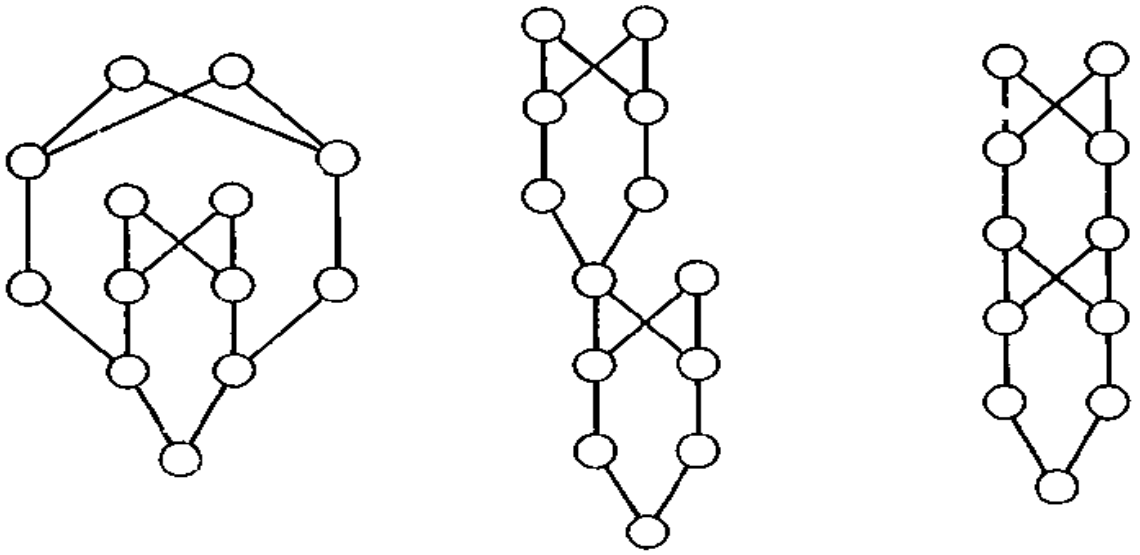
Lo studio dell'origine e della distribuzione dei cognomi è per sua stessa natura intrinsecamente interdisciplinare. Un'analisi non settoriale della questione coinvolge infatti non solo le discipline storiche e demografiche, ma anche la genetica delle popolazioni. In questa sede vogliamo ripercorrere con qualche dettaglio le linee di sviluppo e i principali risultati che sono emersi in quest'ambito dagli studi svolti in prevalenza dai genetisti, e negli ultimi anni anche da ricercatori che usano i concetti e applicano i metodi della fisica statistica.

Il principale motivo di interesse per lo studio dei cognomi nell'ambito della biologia umana nasce dal fatto che i cognomi, pur essendo di origine culturale e non biologica, si trasmettono di generazione in generazione secondo regole ben precise e legate al comportamento riproduttivo, esattamente come avviene per i geni, e più specificamente per i cosiddetti "geni neutri", che pur facendo parte integrante del corredo genetico non hanno influenza sul fenotipo e pertanto non sono soggetti a nessun tipo di pressione selettiva. La loro distribuzione è quindi governata soltanto dalle leggi della statistica, che devono però tener conto dell'esistenza di fenomeni anche importanti di mutazione e di migrazione (e ciò vale tanto per i cognomi quanto per i geni).
In particolare, nelle culture in cui il cognome dei figli è mutuato da quello paterno, le sue leggi di trasmissione sono esattamente le stesse che governano il cromosoma Y, che essendo posseduto soltanto dal sesso maschile viene ereditato soltanto dal padre e resta invariato nella linea di discendenza maschile diretta con la sola eccezione delle mutazioni (e delle false paternità).
La prima idea di utilizzo dei cognomi nella biologia delle popolazioni risale, certo non incongruamente, a George Darwin[23] (figlio del ben più famoso Charles) che nel 1875 propose di analizzare i casi di matrimoni tra persone portatrici dello stesso cognome per stimare la proporzione dei matrimoni tra cugini primi e valutare gli effetti biologici della consanguineità. Egli suppose che il numero dei matrimoni isonimi che non fossero tra cugini primi dovesse essere proporzionale alla frequenza del cognome nella popolazione, e quindi frequente soltanto per i cognomi comuni. Sulla base della frequenza dei 50 cognomi più comuni in Inghilterra nel 1853 egli stimò che il numero atteso per i matrimoni isonimici tra persone non imparentate fosse prossimo all'uno per mille, e che la deviazione osservata da tale frequenza fosse da attribuirsi ai matrimoni tra cugini primi, che egli stimò al 4,5% per l'aristocrazia, al 3,5% per la borghesia, al 2,5% per la popolazione rurale e al 2% per la popolazione urbana.
L'idea fu abbandonata per circa trent'anni, ripresa da Arner[24] nel 1908, di nuovo abbandonata, e riproposta infine nel 1960 da Shaw[25] che sottolineò l'opportunità offerta dal sistema spagnolo che attribuisce a ciascun individuo due cognomi, corrispondenti nell'ordine al primo cognome del nonno paterno e di quello materno, per cui la frequenza dell'identità tra i due cognomi permetterebbe una più rapida misura (statistica) della consanguineità.
La vera svolta nell'utilizzo dei cognomi nello studio della genetica delle popolazioni umane si ebbe però soltanto nel 1965, quando Crow e Mange[26] proposero un modello matematico per la stima della consanguineità a partire dai cognomi. Essi definirono la consanguineità totale (Ft) di una popolazione e le sue componenti casuale (Fr) e non casuale (Fn), delle quali soltanto la prima può essere spiegata dall'accoppiamento casuale (*random mating)* degli individui presenti in una popolazione.

---

[23] G.H. Darwin, *Marriages between first cousins in England and their effects*, Journal of the Statistical Society 38 (1875) 153-184
[24] G.B.L. Arner, *Consanguineous marriages in the American population*, Columbia U. Studies in History, Economics and Public Law (XXXI) 3, New York 1908
[25] R.F. Shaw, *An index of consanguinity based in the use of the surname in Spanish speaking countries*, Journal of Heredity 51 (1960) 221-230
[26] J.F. Crow and A.P. Mange, *Measurement of inbreeding from the frequency of marriages between persons of the same surname*, Eugenics Quarterly 12 (1965) 199-203

Crow e Mange fecero quindi le ipotesi restrittive (e non certo applicabili a tutti i casi storici concreti) che tutti i cognomi siano monofiletici (ovvero che condividere il cognome significhi condividere l'antenato da cui esso deriva) e che i due sessi siano ugualmente rappresentati nei migranti, Sotto queste ipotesi essi osservarono che, per un grande numero di tipologie di parentela, l'isonimia nei matrimoni I (*marital isonymy*) è indicativa del grado di consanguineità nella popolazione indipendentemente dal grado di consanguineità nei singoli matrimoni, in quanto la probabilità che due discendenti da un antenato comune abbiano lo stesso cognome varia, nella maggior parte dei casi in misura proporzionale al grado di consanguineità.
Per l'esattezza, la probabilità P che due individui abbiano lo stesso cognome è ovviamente legata al grado di parentela (e vale 1 per i fratelli, 1/2 per la parentela zio-nipote, 1/4 per i cugini, e così via), ma anche il grado di consanguineità F tra i loro figli dipende dal grado di parentela (e vale 1/4 per i figli di fratelli, 1/16 per i figli di cugini e così via), ed è abbastanza facile convincersi che tipicamente vale la relazione F = P/4, con rare eccezioni legate a relazioni di parentela complesse e incrociate la cui rilevanza statistica è comunque assai scarsa. Di conseguenza il numero effettivo di matrimoni tra consanguinei è tanto maggiore di quello indicato dall'isonimia quanto più remota è la parentela, e il rapporto numerico cresce in ragione inversa di P, ma in compenso la consanguineità tra i discendenti di matrimoni tra parenti è ridotta in proporzione a F, e i due effetti si compensano in virtù della relazione F/P = 1/4. Pertanto la percentuale dei matrimoni isonimi è una misura diretta del grado di consanguineità Ft della popolazione nel suo complesso, che vale appunto 1/4 dell'isonimia matrimoniale osservata: Ft = 1/4 I



Nella figura, tratta da un articolo di Crow[27], sono illustrati alcuni esempi di applicazione della formula di Crow e Mange. Nel primo caso si ottiene F= 5/64, P/4 = 5/64 e quindi la formula genera il risultato corretto, mentre nel secondo caso F = 33/512, P/4 = 32/512 e nel terzo caso F = 9/128, P/4 = 8/128, quindi negli ultimi due casi la stima della consanguineità basata sull'isonimia risulta un po' inferiore al valore vero. Dovrebbe essere comunque chiaro dalla figura che si tratta di casi assai rari e complessi, e che l'errore percentuale è comunque limitato.

[27] J.F. Crow, *The estimation of inbreeding from isonymy*, Human Biology 52 (1980) 1-14

Per separare la componente casuale della consanguineità da quella non casuale, Crow e Mange notarono che, dette p(i) e q(i) le frazioni della popolazione maschile e femminile individuate dal cognome i-esimo, la frequenza dei matrimoni casuali tra coppie con lo stesso cognome è p(i) q(i) e il coefficiente di consanguineità casuale Fr vale quindi, per quanto in precedenza dimostrato, $Fr = 1/4 \sum p(i)\, q(i)$, mentre il coefficiente di consanguineità non casuale Fn si ricava facilmente dalla relazione $Ft = Fn + (1 – Fn)\, Fr$, dove a sua volta Ft è misurato dall'isonimia nei matrimoni.
A livello teorico è stato mostrato da Holgate[28] nel 1971 che il coefficiente di isonimia tende a crescere nel tempo in modelli di popolazione monogama con accoppiamenti casuali. Nel 1977 Lasker[29], osservando che il coefficiente di parentela per una coppia di genitori è esattamente il doppio del coefficiente di consanguineità dei loro figli, propose un'estensione della formula di Crow e Mange che permettesse di usare l'isonimia come misura del grado di parentela tra due popolazioni qualunque. La parentela misurata dall'isonimia vale $RI = 1/2 \sum p1(i)\, p2(i)$, dove p1(i) e p2(i) sono le frequenze con cui il cognome i-esimo compare nelle due popolazioni.
I risultati di Crow e Mange e di Lasker sono stati presto applicati a differenti popolazioni, anche con correzioni atte a tener conto degli effetti del polifiletismo e delle migrazioni[30]. Per una ricognizione dei dati raccolti fino al 1980 e per un inquadramento generale si vedano l'articolo di rivista di Lasker[31] e il volume pubblicato dallo stesso nel 1985.
Tra gli sviluppi successivi merita di segnalare soprattutto lo studio dell'isonimia matrimoniale nel contesto iberico e iberoamericano[32], particolarmente adatto per i motivi citati in precedenza.

Un recente caso di applicazione del calcolo delle probabilità allo studio dei matrimoni isonimi è quello relativo alla popolazione degli Xuetes (ebrei convertiti) delle isole Baleari tra il 1565 e il 1750, studiato da Porqueres[33] e riesaminato da Rossi[34].
Considerando soltanto le undici famiglie (su diciassette studiate) che sono nominate più di frequente, poiché i numeri delle altre sono troppo piccoli per potere considerare attendibili le analisi statistiche, e confrontando i valori empirici (Tabella I) dei matrimoni interni alla comunità con quelli determinati dal calcolo teorico delle probabilità (Tabella II), si trova una chiara evidenza di fenomeni che non possono essere interpretati come fluttuazioni e richiedono quindi una spiegazione sociologica e/o antropologica: notiamo l'eccesso di matrimoni isonimi dei Fuster, dei Forteza, dei Cortés e dei Martì, l'eccesso di matrimoni dei tipi Pinya-Picò, Picò-Mirò e all'interno del gruppo Forteza-Cortés-Martì, e per converso la scarsità di matrimoni dei componenti di quest'ultimo gruppo con i membri della maggior parte delle altre famiglie.

---

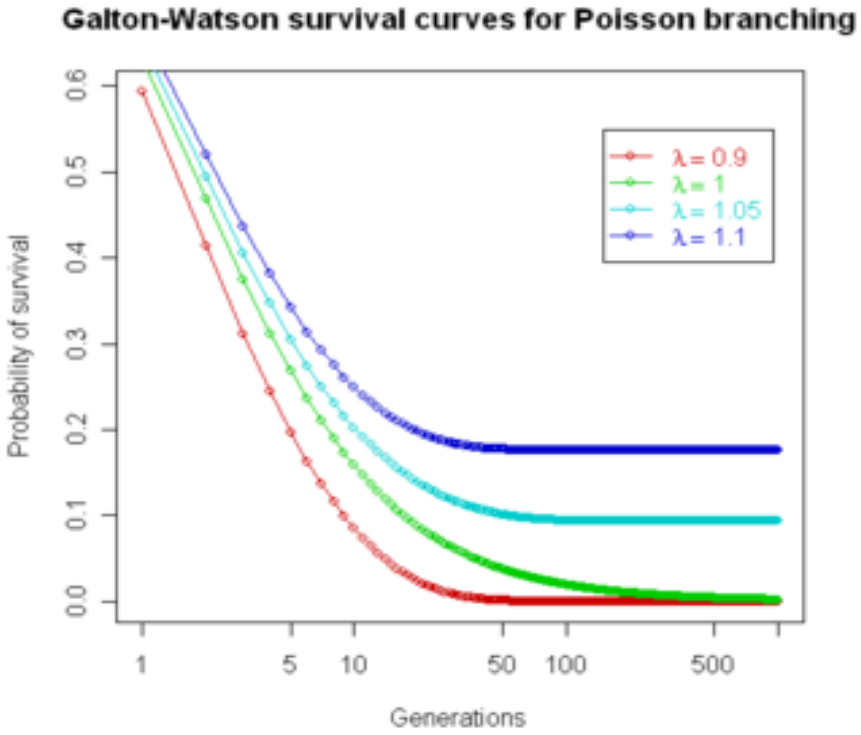
[28] P. Holgate, *Drifts in the Random Component of Isonymy*, Biometrics 27 (1971) 448-451
[29] G.W. Lasker, *A coefficient of relationship by isonymy: A Method for Estimating the genetic Relationship between Populations*, Human Biology 49 (1977) 489-493
[30] W.S. Ellis and W.T. Starmer, *Inbreeding as measured by Isonymy*, Pedigrees, and Population Size in Törbel, Switzerland, Am. J. Hum. Genet. 30 (1978) 366-376
[31] G.W. Lasker, *Surnames in the Study of Human Biology*, American Anthropologist 82 (1980) 525-538; G.W. Lasker, *Surnames and Genetic Structure*, Cambridge University Press 1985
[32] J. Pinto-Cisternas, L. Pineda and I. Barrai, *Estimation of Inbreeding by Isonymy in Iberoamerican Populations*, Am. J. Hum. Genet. 37 (1985) 373-385; M. Esparza, C. Garcia-Moro and M. Hernandez, *Inbreeding From Isonymy and repeated Pairs of Surnames in the Ebro Delta Region*, Am. J. Hum. Biology 18 (2006) 849-852
[33] E. Porqueres i Gené, *Lourde alliance. Mariage et identité chez les descendants des juifs convertis de Majorque (1435-1750)*, Kymé, Paris1995
[34] P. Rossi, *La distribution des noms de famille comme outil pour l'analyse des dinamiques migratoires,* in "Un juego de engaños. Nombres, apellidos y movilidad en los siglos XV al XVII, Madrid 2010

Tableau I – Les intermariages des familles Xuetes de Majorque											
	Fuster	Forteza	Aguilò	Cortés	Pinya	Pomar	Bonnìn	Martì	Mirò	Picò	Valls
Fuster	22	13	10	7	13	10	5	2	4	7	10
Forteza	11	19	15	17	5	7	2	10	7	1	3
Aguilò	15	18	12	15	7	6	8	11	4	2	5
Cortés	4	13	13	18	2	6	2	14	0	0	6
Pinya	17	7	11	0	6	13	10	2	3	10	7
Pomar	12	12	6	2	11	10	8	7	4	4	4
Bonnìn	14	5	7	4	9	11	4	0	8	8	5
Martì	5	16	6	20	0	5	2	10	0	0	0
Mirò	7	3	8	1	5	6	10	0	4	9	4
Picò	8	2	4	0	10	6	4	0	6	3	4
Valls	4	7	3	4	7	5	0	1	4	3	1

Tableau II – Les intermariages des Xuetes : valeur statistiquement plus probable											
	Fuster	Forteza	Aguilò	Cortés	Pinya	Pomar	Bonnìn	Martì	Mirò	Picò	Valls
Fuster	15	15	13	12	10	11	7	7	6	6	7
Forteza	14	13	11	11	9	9	6	7	5	5	6
Aguilò	14	14	12	11	10	10	7	7	5	6	6
Cortés	11	11	9	9	8	8	5	6	4	4	5
Pinya	12	11	10	9	8	8	5	6	4	5	5
Pomar	11	10	9	8	7	7	5	5	4	4	5
Bonnìn	11	10	9	8	7	7	5	5	4	4	5
Martì	9	9	8	7	6	7	4	5	3	4	4
Mirò	8	8	7	6	5	6	4	4	3	3	4
Picò	7	6	5	5	4	5	3	3	2	3	3
Valls	5	5	4	4	4	4	3	3	2	2	2

12. La distribuzione in frequenza e l'estinzione dei cognomi

Una linea di ricerca indipendente dalla precedente fu avviata anch'essa nel XIX secolo dallo studio di Galton e Watson[35] che nel 1874, stimolati dall'osservazione che numerose famiglie dell'aristocrazia britannica si erano estinte o rischiavano l'estinzione, si posero il problema di calcolare la probabilità di tale estinzione, giungendo inizialmente all'errata conclusione che la probabilità di estinzione di qualunque cognome dopo un tempo sufficientemente lungo fosse sempre 1. In realtà tale risultato vale soltanto nel caso in cui il numero medio di discendenti maschi sia minore o uguale a uno, mentre in caso contrario la probabilità è certamente maggiore di 0 ma inferiore a 1. Lo studio di Galton e Watson diede comunque il via allo studio matematico dei fenomeni di ramificazione (*branching processes*), che nel contesto dell'analisi evolutiva delle popolazioni fu ripreso da Lotka[36] a partire dal 1931.



Tuttavia soltanto l'apparizione nel 1967 dell'articolo di Karlin e McGregor[37] che formulava la teoria del comportamento delle mutazioni neutre in popolazioni finite di dimensione costante, unita alla constatazione che i cognomi possono essere considerati come alleli trasmessi lungo la linea maschile, attrasse l'attenzione degli studiosi di genetica verso lo studio della distribuzione e dell'estinzione dei cognomi, a partire dal lavoro di Yasuda *et al.*[38] pubblicato nel 1974, nel quale la teoria di Karlin e McGregor fu applicata ai dati sui cognomi raccolti tra la popolazione della valle del Parma da Cavalli-Sforza e collaboratori fin dal 1954.

---

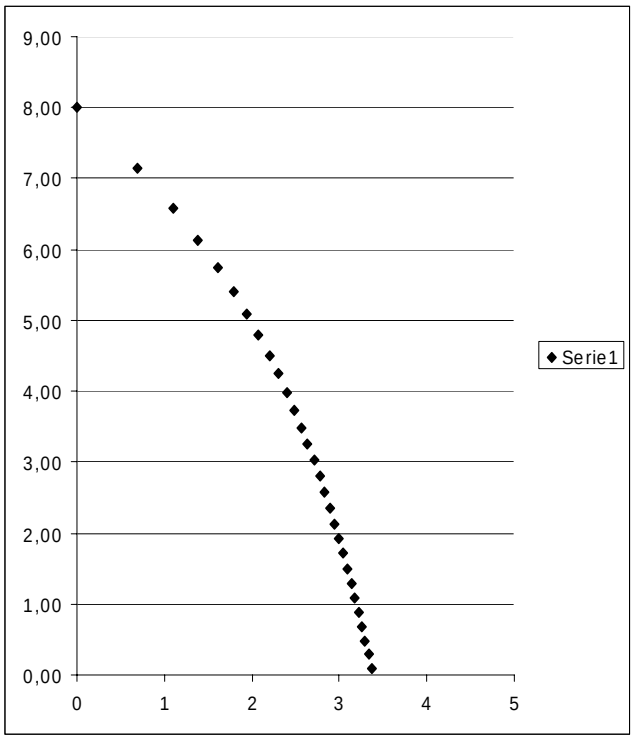
[35] F. Galton and H.W. Watson, *On the Probability of the Extinction of Families*, Journal of the Anthropological Institute of Great Britain and Ireland 4 (1874) 138-144

[36] A.J. Lotka, *The extinction of families I-II*, J. Wash. Acad. Sci. 21 (1931) 377-380, 453-459

[37] S. Karlin and J. McGregor, *The number of mutant forms maintained in a population*, Proc. 5th Berkeley Symp. Math. Stat. Prob. 4 (1967) 415-438

[38] N. Yasuda, L.L. Cavalli-Sforza, M. Skolnick and A. Moroni, *The Evolution of Surnames: An Analysis of Their Distribution and Extinction*, Theor. Pop. Biol. 5 (1974) 123

La teoria permette di predire il valore atteso della distribuzione di N*(k), il numero dei cognomi che sono rappresentati da k individui (maschi) in una popolazione di dimensione N, assumendo che in ogni transizione (passaggio da una generazione alla successiva) sia assegnata (e uguale per tutti gli individui) la probabilità ν di un cambio di cognome. La media della distribuzione è il numero atteso S di cognomi nella popolazione, e si può calcolare il valore atteso dell'isonimia casuale, ossia la probabilità che due individui presi a caso abbiano lo stesso cognome, ottenendo $I = 1/(N\nu + 1 - \nu)$. Si può dimostrare che in questo modello n $I < 1$. Il confronto tra la distribuzione teorica e quella empirica permette di verificare la bontà delle assunzioni e di fissare i valori dei parametri N e ν. Poiché le vere mutazioni di cognome sono relativamente rare, un valore elevato di ν (come quello trovato nel caso della valle del Parma, superiore a 0,2) sembra indicare l'importanza dei fenomeni di immigrazione nella distribuzione dei cognomi.
Yasuda *et al.*, nello stesso articolo, affrontarono poi il tema della distribuzione statistica della discendenza, in particolare per l'aspetto relativo alla probabilità di estinzione dei cognomi dopo un numero sufficientemente alto di generazioni. Le probabilità di estinzione da essi calcolate (sulla base di un modello di decrescita geometrica per le probabilità di nascita di n figli maschi) risultarono in ragionevole accordo con i dati empirici raccolti.
Un risultato fondamentale nello studio della distribuzione delle frequenze dei cognomi fu pubblicato da Fox e Lasker[39] nel 1983 (ma già annunciato nell'articolo di Lasker del 1980). Fox e Lasker mostrarono che i dati empirici relativi alla distribuzione in frequenza dei cognomi di 4794 individui viventi nell'area di Reading (Inghilterra) potevano essere descritti con buona precisione utilizzando una distribuzione di Pareto discreta (ossia una legge di potenza).



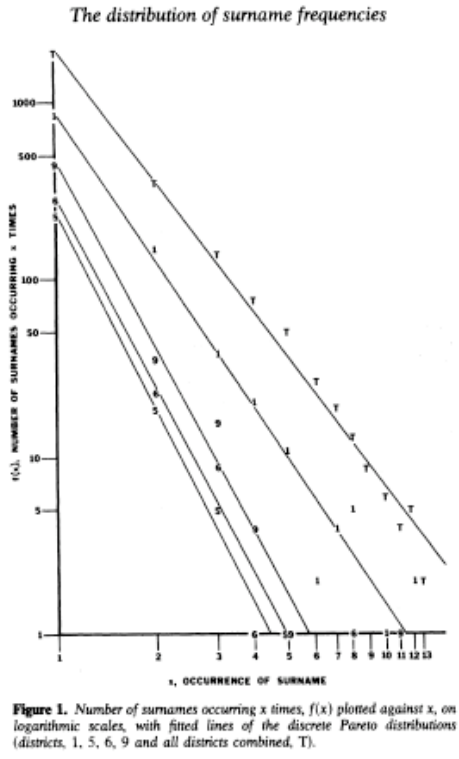


Figure 1. *Number of surnames occurring x times, f(x) plotted against x, on logarithmic scales, with fitted lines of the discrete Pareto distributions (districts, 1, 5, 6, 9 and all districts combined, T).*

Nelle due figure sono poste a confronto le due distribuzioni che risultano rispettivamente dal modello di Karlin e McGregor (nella forma semplificata di Fisher) e dal *fit* di Fox e Lasker basato sulla distribuzione di Pareto. Appare chiaro che i due modelli non sono tra loro equivalenti: una possibile connessione sarà discussa in seguito.

[39] W.R. Fox and G.W. Lasker, *The Distribution of Surname Frequencies*, Int. Stat. Review 51 (1983) 81-87

Un altro importante contributo all'indagine sulla distribuzione in frequenza dei cognomi fu portato da una serie di lavori pubblicato a partire dal 1983, Zei *et al.*[40] e specificamente dedicati allo studio della distribuzione dei cognomi in Sardegna. Fu mostrato che la distribuzione empirica mostrava un buon accordo con la distribuzione degli alleli neutri di Karlin-McGregor, e anche con la distribuzione logaritmica di Fisher, peraltro derivabile dalla precedente nel limite in cui N diventa infinito. Vale in questo caso $N^*(k) = \alpha/k \; x^k$, dove si sono introdotte le nuove variabili $x = 1 - \nu$ e $\alpha = N \nu/(1 - \nu)$, ed è possibile valutare il numero totale S dei cognomi mediante la semplice relazione $S = - \alpha \ln \nu$.
Infine fu considerata la dipendenza spaziale della distribuzione dei cognomi, notando come la variazione spaziale fosse molto più marcata della variazione temporale, e largamente indipendente dalla frequenza dei cognomi (salvo il caso di bassissima frequenza). Questi risultati sono tutti consistenti con l'ipotesi che i cognomi si comportino come alleli neutri.
Ulteriori proposte di parametrizzazione della distribuzione in frequenza dei cognomi, basate su modelli statistici dei comportamenti riproduttivi e della conseguente evoluzione dei cognomi, furono avanzati da Panaretos[41] che nel 1989 propose l'impiego della distribuzione di Yule, e da Consul[42], che nel 1991 ricavò, sia da un modello di nascita e morte che da un modello di ramificazione, una descrizione basata sulla distribuzione di Geeta, in buon accordo con i dati.

Nel lavoro di Wijsman *et al.*[43] del 1984 furono calcolate le matrici di migrazione tra nove differenti aree della Sardegna, sulla base di una matrice di parentela definita in modo analogo a quello introdotto da Lasker e valutata tra popolazioni considerate a tempi diversi (1850-1970).
Fu così evidenziata una dipendenza forte e non lineare dei coefficienti di immigrazione dalla distanza tra i territori esaminati.
Il tema dell'utilizzo dei cognomi per lo studio delle dinamiche migratorie fu ulteriormente sviluppato in alcuni lavori successivi. In particolare Piazza *et al.*[44] nel 1987 confrontarono le stime dei coefficienti di migrazione in Italia da loro ottenute analizzando le distribuzioni di cognomi estratte dagli elenchi telefonici con le corrispondenti stime ricavate dalle fonti demografiche ufficiali, mostrando che le stime basate sulle distribuzioni dei cognomi mediante il calcolo del coefficiente ν potevano fornire una stima attendibile dei coefficienti di migrazione.
Lo studio dei fenomeni migratori mediante l'analisi della distribuzione dei cognomi fu condotto in seguito anche da Darlu e Ruffié[45] per le migrazioni interne francesi, e da Degioanni *et al.*[46] per le migrazioni dall'Italia verso la Francia.

---

[40] G. Zei, C.R. Guglielmino, E. Siri, A. Moroni and L.L. Cavalli-Sforza, *Surnames as Neutral Alleles: Observations in Sardinia*, Human Biology 55 (1983) 357-365; G. Zei, R. Guglielmino Matessi, E. Siri, A. Moroni and L. Cavalli-Sforza, *Surnames in Sardinia I. Fit of frequency distributions for neutral alleles and genetic population structure*, Ann. Hum. Genet. 47 (1983) 329-352; G. Zei, A. Piazza, A. Moroni and L.L. Cavalli-Sforza, *Surnames in Sardinia III. The spatial distribution of surnames for testing neutrality of genes*, Ann. Hum. Genet. 50 (1986) 169

[41] J. Panaretos, *On the Evolution of Surnames*, Int. Stat. Review 57 (1989) 161

[42] P. C. Consul, *Evolution of Surnames*, Int. Stat. Review 59 (1991) 271; M.N. Islam, *A Stochastic Model for Surname Evolution*, Biom. J. 37 (1995) L119-126

[43] E. Wijsman, G. Zei, A. Moroni and L.L. Cavalli-Sforza, *Surnames in Sardinia II. Computation of migration matrices from surname distributions in different periods*, Ann. Hum. Genet. 48 (1984) 65-78;

[44] A. Piazza, S. Rendine, G. Zei, A. Moroni and L.L. Cavalli-Sforza, *Migration rates of human populations from surname distributions*, Nature 329 (1987) 714; A. Piazza, N. Cappiello, E. Olivetti and S. Rendine, *A genetic history of Italy*, Ann. Hum. Genet. 52 (1988) 203-213

[45] P. Darlu and J. Ruffié, *L'immigration dans les départements français étudiée par la méthode des patronymes*, Population 47 n.3 (1992) 719; P. Darlu and J. Ruffié, *Relationships between consanguinity and migration rate from surname distributions and isonymy in France*, Ann. Hum. Biol. 19 (1992) 133; P. Darlu, A. Degioanni, J. Ruffié, *Quelques statistiques sur la distribution des patronymes en France*, Population 52 n.3 (1997) 607

[46] A. Degioanni, A. Lisa, G. Zei, P. Darlu, *Patronymes italiens et migration italienne en France*, Population 51 n.6 (1996) 1153

A loro volta Zei *et al.*[47] ricostruirono le dinamiche migratorie italiane a partire dall'analisi della distribuzione geografica dei cognomi (e in particolare dallo studio dei confini tra regioni in cui gli stessi cognomi sono presenti con frequenze molto differenti), confrontando i risultati con le barriere geografiche e linguistiche e con gli esiti delle analisi genetiche delle popolazioni. I confini genetici e quelli tra i cognomi mostrarono locazioni molto simili, e correlate soprattutto alle barriere fisiche, ma anche a quelle linguistiche.
La relazione tra la trasmissione genetica (puramente verticale) e quella culturale (che è spesso anche orizzontale) è stata investigata confrontando la distribuzione dei cognomi con quella di altri marcatori culturali in una serie di lavori relativi alla Sicilia, a partire da Guglielmino *et al.*[48]

Tra i recenti sviluppi dell'indagine sull'origine e l'evoluzione biologica dei cognomi uno dei più interessanti è quello avviato dall'articolo di Sykes e Irven[49] del 2000, basato sull'osservazione che nella maggior parte delle culture, e in particolare in quelle europee, la trasmissione ereditaria del cognome avviene in forma patrilineare, in perfetta analogia con quanto avviene per il cromosoma Y, che essendo posseduto soltanto dai maschi viene ereditato esclusivamente dal padre. Fu quindi analizzato il patrimonio genetico di un significativo campione casuale di soggetti maschi portatori del cognome Sykes. Dal fatto che quasi il 50% del campione condivideva un particolare aplotipo non osservato nei gruppi di controllo, Sykes e Irven inferirono che, ammettendo un numero molto limitato di casi di nonpaternità nell'arco dei 700 anni in cui il cognome è documentato (in media 1,3% per generazione), il cognome Sykes dovesse ritenersi monofiletico (mentre fonti scritte parevano indicare un'origine multipla).
Una prima descrizione dello stato delle ricerche sul cromosoma Y e sul loro intreccio con lo studio dell'origine e della distribuzione dei cognomi si trova nell'articolo di Jobling[50] del 2001 e nella relativa bibliografia. Tra i risultati più importanti vanno ricordati quello sulla relazione tra gli aplotipi di un gruppo di 221 maschi irlandesi e l'origine (gaelica o anglosassone) dei loro cognomi, e quello relativo al patrimonio genetico dei religiosi ebraici il cui cognome (Cohen e sue varianti) indica, per la tradizione biblica, una discendenza patrilineare diretta da Aronne, fratello di Mosé (XIII sec. a.C.). In entrambi i casi la forte correlazione sembra indicare uno stretto legame tra il dato genetico e quello culturale.
Il metodo di analisi della distribuzione spaziale dei cognomi è stato raffinato nell'articolo di Manni *et al.*[51], con l'uso del metodo di raggruppamento SOM (*self-organizing maps*), che gli autori hanno applicato al caso dei cognomi dei Paesi Bassi, suggerendo che il campionamento genetico di un'area soggetta a fenomeni di immigrazione debba essere effettuato con selezione preliminare basata sul cognome, se si vogliono mettere in evidenza i tratti genetici caratteristici della popolazione da maggior tempo residente nell'area in esame.

---

[47] G. Zei, G. Barbujani, A. Lisa, O. Fiorani, P. Menozzi, e. Siri and L.L. Cavalli-Sforza, *Barriers to gene flow estimated by surname distribution in Italy*, Ann. Hum. Genet. 57 (1993) 123
[48] C.R. Guglielmino, G. Zei and L.L. Cavalli-Sforza, *Genetic and Cultural transmission in Sicily as Revealed by Names and Surnames*, Human Biology 63 (1991) 607-627; A. Rodriguez-Larralde, A. Pavesi, C. Scapoli, F. Conterio, G. Siri and I. Barrai, *Isonymy and the genetic structure of Sicily*, J. Biosoc. Sci. 26 (1994) 9-24; A. Piazza et al., *Towards a genetic history of Sicily*, J. Cult. Heritage 1 (2000) Sup.39-42; A. Pavesi, P. Pizzetti, E. Siri, E. Lucchetti and F. Conterio, *Coexistence of Two Distinct Patterns in the Surname Structure of Sicily*, Am. J. Phys. Anthrop. 120 (2003) 195-199; A. De Silvestri and C.R. Guglielmino, *Sicilian Provinces: Population Subdivisions Revealed by Surname Frequencies*, Human Biology 76 (2004) 901-920
[49] B. Sykes and C. Irven, *Surnames and the Y Chromosome*, Am. J. Hum. Genet. 66 (2000) 1417-1419
[50] M. A. Jobling, *In the name of the father: surnames and genetics,* Trends in Genetics 17 (2001) 353-357
[51] F. Manni, B. Toupance, A. Sabbagh and E. Heyer, *New Method fir Surname Studies of Ancient Patrilineal Population Structures, and possible Application to Improvement of Y-Chromosome Sampling*, Am. J. Phys. Anthrop. 126 (2005) 214-228

13. Studi empirici sulla distribuzione dei cognomi

Le più recenti messe a punto dello stato delle ricerche in forma di articolo di rivista sono quella di Colantonio *et al.*[52] del 2003, con un'amplissima bibliografia ragionata, ripartita per aree geografiche, e quella di P. Darlu[53], che aveva l'obiettivo esplicito di trasmettere le acquisizioni della ricerca effettuata dai genetisti alla comunità dei demografi storici, spesso interessati a problemi molto simili (dinamica delle popolazioni, fenomeni migratori) ma finora non particolarmente interessati allo studio della distribuzione dei cognomi.
Un'altra importante messa a punto è quella contenuta nell'articolo del 2007 di Scapoli *et al*.[54], in cui vengono riassunti e nuovamente analizzati soprattutto i dati raccolti in una ventina d'anni di attività da un ampio gruppo di ricercatori il cui principale esponente è I. Barrai.
I primi lavori del gruppo (fino al 1994) si concentrarono su alcune realtà locali italiane (Ferrara, Perugia, la già citata Sicilia)[55], indagando sia i fenomeni di consanguineità che quelli di migrazione e di isolamento dovuto alla distanza.
In seguito, quando per molti Paesi divennero largamente disponibili le versioni digitalizzate degli elenchi telefonici, il gruppo ampliò la propria indagine a tali Paesi, e a un numero elevatissimo di soggetti, fino a coprire mediamente il 9% della popolazione dei principali Stati dell'Europa occidentale (Svizzera, Germania. Italia, Austria, Paesi Bassi, Belgio Spagna, Francia)[56]
Alcune indagini del gruppo hanno riguardato anche Venezuela, Argentina, Stati Uniti.[57]

---

[52] S.E. Colantonio, G.W. Lasker, B.A. Kaplan and V. Fuster, *Use of Surname Models in Human Population Biology: A Review of Recent Developments*, Human Biology 75 (2003) 785-807
[53] P. Darlu, *Patronymes et démographie historique*, Ann. Dem. Hist. 2 (2004) 53-65
[54] C. Scapoli, E. Mamolini, A. Carrieri, A. Rodriguez-Larralde, I. Barrai, *Surnames in Western Europe: A comparison of the subcontinental populations through isonymy*, Theor. Pop. Biol. 71 (2007) 37-48
[55] I. Barrai, G. Barbujani, M. Beretta, I. Maestri and A. Russo, *Surnames in Ferrara: distribution, isonymy and levels of inbreeding*, Ann. Hum. Biol. 14 (1987) 415; I. Barrai, G. Formica, R. Barale, C. Scapoli, R. Camella and M. Beretta, *Isonymy in immigrants from Ferrara in 1981-1988*, Ann. Hum. Biol. 17 (1990) 7; I. Barrai, G. Formica, R. Barale and M. Beretta*, Isonymy and migration distance*, Ann. Hum. Genet. 53 (1989) 249; I. Barrai, C. Scapoli, R. Carella, G. Formica, R. Barale and M. Beretta, *Isonymy in records of births and deaths in Ferrara*, Ann. Hum. Biol. 18 (1991) 395; I. Barrai, G. Formica, C. Scapoli, M. Beretta. E. Mamolini, S. Volinia, R. Barale, P. Ambrosino and F. Fontana, *Microevolution in Ferrara: Isonymy 1890-1990*, Ann. Hum. Biol. 19 (1992) 371; A. Rodriguez-Larralde, G. Formica, C. Scapoli, M. Beretta. E. Mamolini and I. Barrai, *Microevolution in Perugia: isonymy 1890-1990*, Ann. Hum. Biol. 20 (1993) 261; M. Beretta, E. Mamolini, A. Ravani, C. Vullo, C. Scapoli, R. Barale, A. Rodriguez-Larralde and I. Barrai, *Comparison of Structures from Frequencies of Genes and Surnames in the Population of Ferrara*, Human Biology 65 (1993) 225
[56] I. Barrai, C. Scapoli, M. Beretta, C. Nesti, E. Mamolini and A. Rodriguez-Larralde, *Isonimy and the genetic structure of Switzerland I. The distribution of surnames*, Ann. Hum. Biol. 23 (1996) 431; C. Scapoli, A. Rodriguez-Larralde, M. Beretta, C. nesti, A. Lucchetti, I. Barrai, *Correlations between Isonimy Parameters*, Int. J. of Anthropology 12 (1997) 17; A. Rodriguez-Larralde, C. Scapoli, M. Beretta, C. Nesti, E. Mamolini and I. Barrai, *Isonimy and the genetic structure of Switzerland II. Isolation by distance*, Ann. Hum. Biol. 25 (1998) 533; A. Rodriguez-Larralde, I. Barrai, C. Nesti, E. Mamolini and C. Scapoli*, Isonimy and Isolation by Distance in Germany*, Human Biology 70 (1998) 1041; I. Barrai, A. Rodriguez-Larralde, E. Mamolini and C. Scapoli, *Isonimy and isolation by distance in Italy*, Human Biology 71 (1999) 947; I. Barrai, A. Rodriguez-Larralde, E. Mamolini, F. Manni and C. Scapoli, *Elements of the surname structure of Austria*, Ann. Hum. Biol. 27 (2000) 607; I. Barrai, A. Rodriguez-Larralde, F. Manni and C. Scapoli, *Isonimy and Isolation by Distance in the Netherlands*, Human Biology 74 (2002) 263; I. Barrai, A. Rodriguez-Larralde, F. Manni, V. Ruggiero, D. Tartari and C. Scapoli, *Isolation by Language and Distance in Belgium*, Ann. Hum. Genetics 68 (2003) 1; A. Rodriguez-Larralde, A. Gonzales-Martin, C. Scapoli, I. Barrai, *The names of Spain: A study of the isonymy structure of Spain*, Am. J. Phys. Anthropol. 121 (2003) 280; C. Scapoli, H. Goebl, S. Sobota, E. Mamolini, A. Rodriguez-Larralde, I. Barrai, *Surnames and dialects in France: Population structure and cultural evolution*, J. Theor. Biol. 237 (2005) 75
[57] A. Rodriguez-Larralde, J. Morales and I. Barrai, *Surname Frequency and the Isonimy Structure of Venezuela*, Am. J. Human Biol. 12 (2000) 352; I. Barrai, A. Rodriguez-Larralde, E. Mamolini, F. Manni and C. Scapoli, *Isonymy structure of USA population,* Am. J. Phys. Anthropol. 114 (2001) 109; J.E. Dipierri, E.L. Alfaro, C. Scapoli, E. Mamolini, A. Rodriguez-Larralde and I. Barrai, Surnames in Argentina: *A population study through isonimy* Am. J. Phys. Anthropol. 128 (2005) 199; A. Rodriguez-Larralde, C. Scapoli, E. Mamolini, I. Barrai, *Surnames in Texas: a population study through isonymy*, Human Biology 79 (2007) 215

Nel principale articolo metodologico, pubblicato da Scapoli *et al.*[58] nel 1997, sono indicati i principali parametri che costituiscono l'oggetto di studio in tutti i lavori del gruppo: l'isonimia I intesa come misura della componente casuale della consanguineità, il coefficiente α di Fisher che misura l'abbondanza dei cognomi in un gruppo, l'indice ν di Karlin e McGregor, che risulta proporzionale al coefficiente di immigrazione, e anche l'entropia della distribuzione dei cognomi espressa dalla relazione $H = \Sigma\, p(i) \ln p(i)$. Si noti che nel limite di N grande, in cui vale la teoria di Fisher, vale anche la relazione $I = 1/\alpha + 1/N$. Viene calcolato anche l'isolamento dovuto alla distanza, misurando la correlazione tra il grado di isonimia e la distanza geografica di coppie di popolazioni.
Un elemento comune a tutti i lavori, ispirato al risultato di Fox e Lasker, è la rappresentazione della distribuzione dei cognomi mediante grafici log-log, che trasformano le leggi di potenza in relazioni lineari, e la stima del corrispondente coefficiente (l'esponente nella legge di potenza) per i diversi Paesi. Nel lavoro riassuntivo sull'Europa occidentale i dati relativi agli otto Paesi esaminati sono stati aggregati, mostrando che anche nel caso generale la rappresentazione della distribuzione come legge di potenza fornisce ancora una descrizione abbastanza soddisfacente dell'andamento generale della distribuzione stessa.
Nella prima figura è rappresentata la distribuzione dei cognomi italiani; nelle figure successive sono riportate le distribuzioni dei cognomi francesi e di quelli europei, ed è analizzata la correlazione della distribuzione geografica dei cognomi con quella dei dialetti e dei gruppi linguistici.

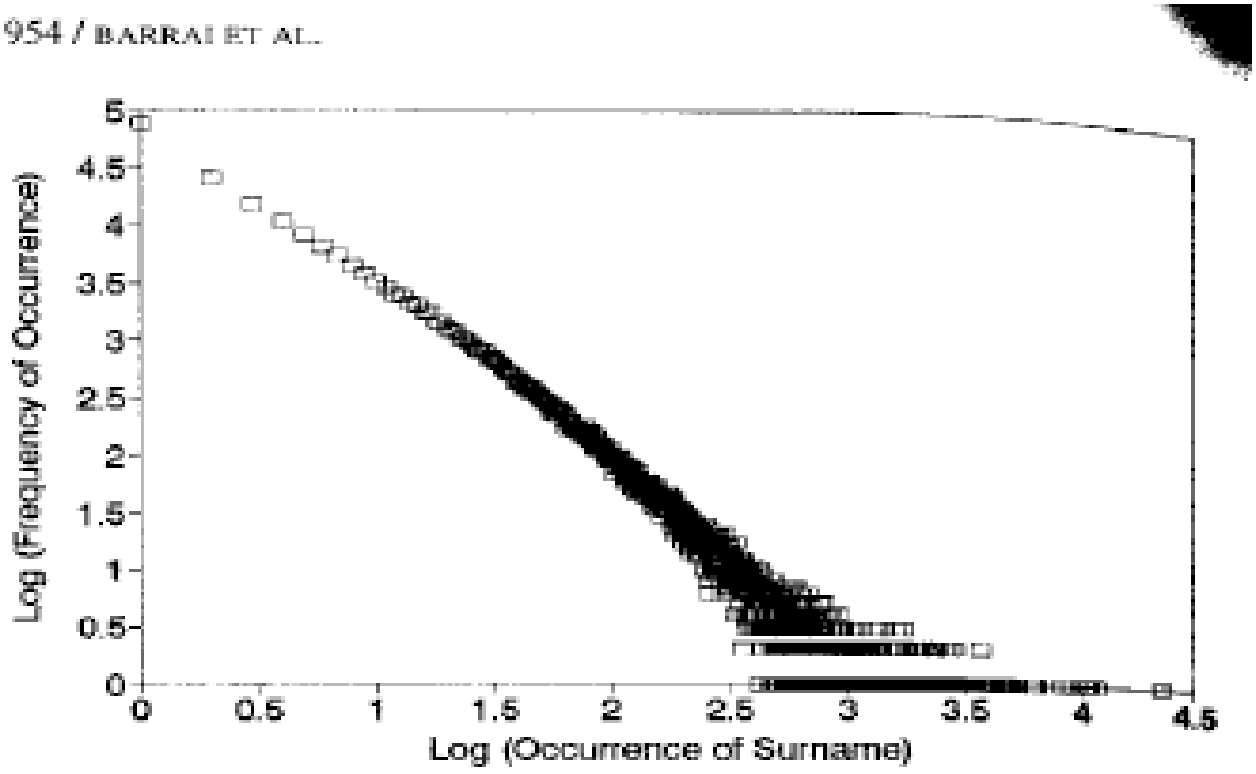


Figure 2. Log-log distribution (natural logarithms) of the occurrence of surnames in Italy in a sample of 5 million individuals in 1996.

[58] C. Scapoli, A Rodriguez-Larralde, M. Beretta, C. Nesti, A. Lucchetti, I. Barrai, *Correlations between Isonymy Parameters*, Int. J. of Anthropology 12 (1997) 17

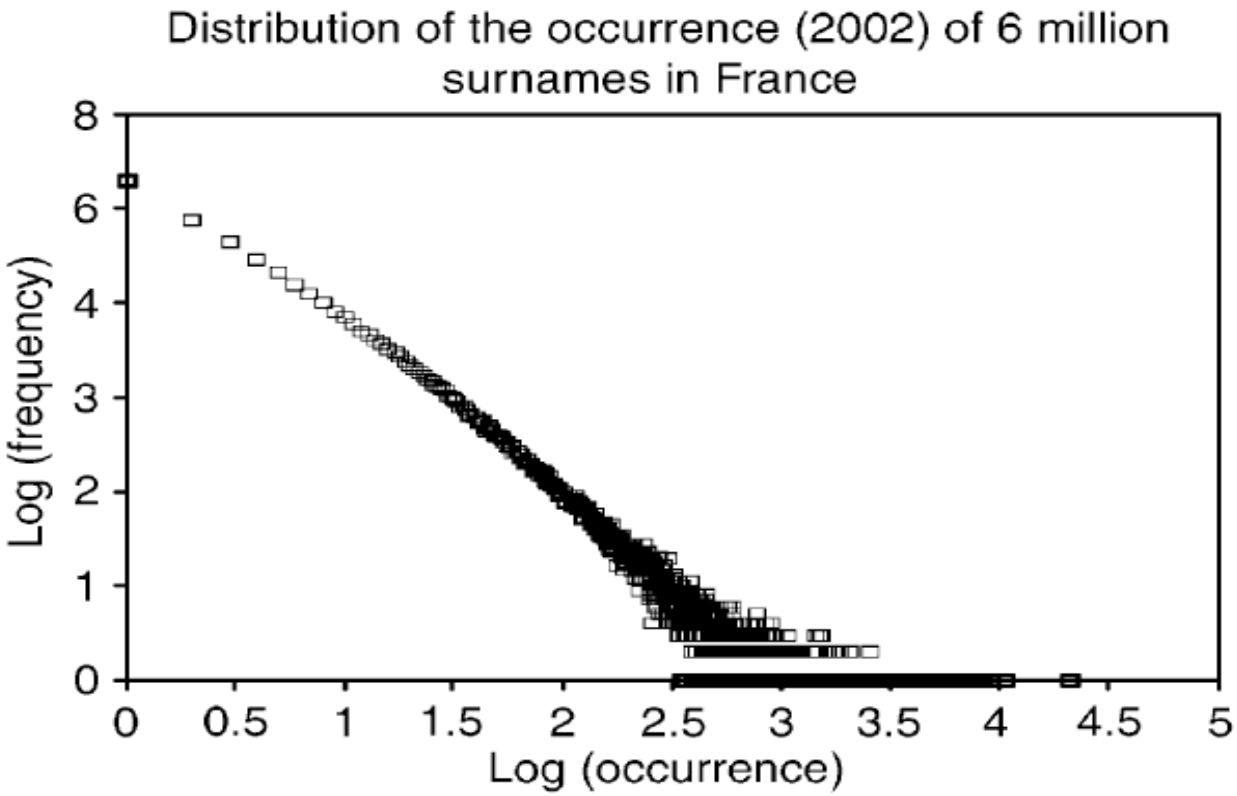
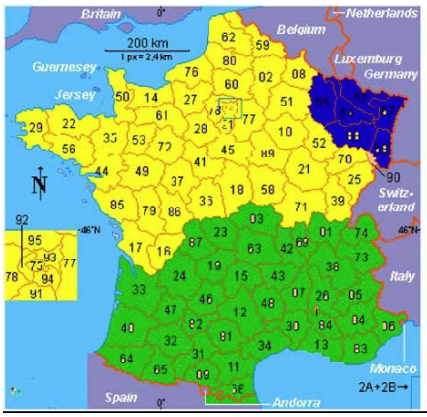
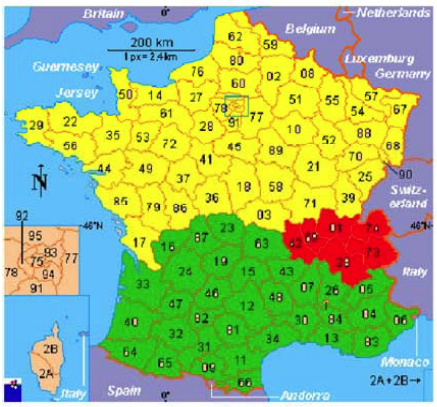


Fig. 2. The log–log distribution of the frequency of occurrence of surnames in France.



Map of France from the matrix of Lasker distances between regions.

Note the main boundary between North and South, and the cluster including Alsace and Lorraine.

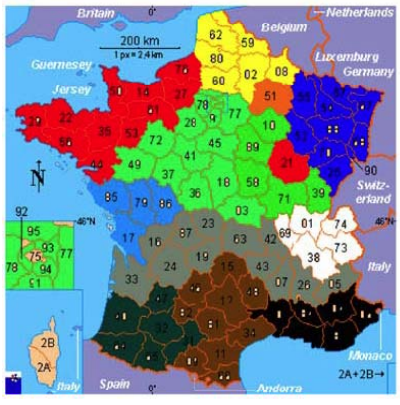


(a)

Map of France derived from the matrix of dialect distances between departments.

The departments of most of the Franco-Provençal area cluster together.

The North-South boundary mostly coincide with that identified by surname distances.



(b)

The subclusters formed by the matrix of dialect distances between departments.

Some of the subclusters correspond to well defined linguistic areas.

The Limousin area is not separated from the Auvergnat and from part of the Gascon at west and Provençal at East.

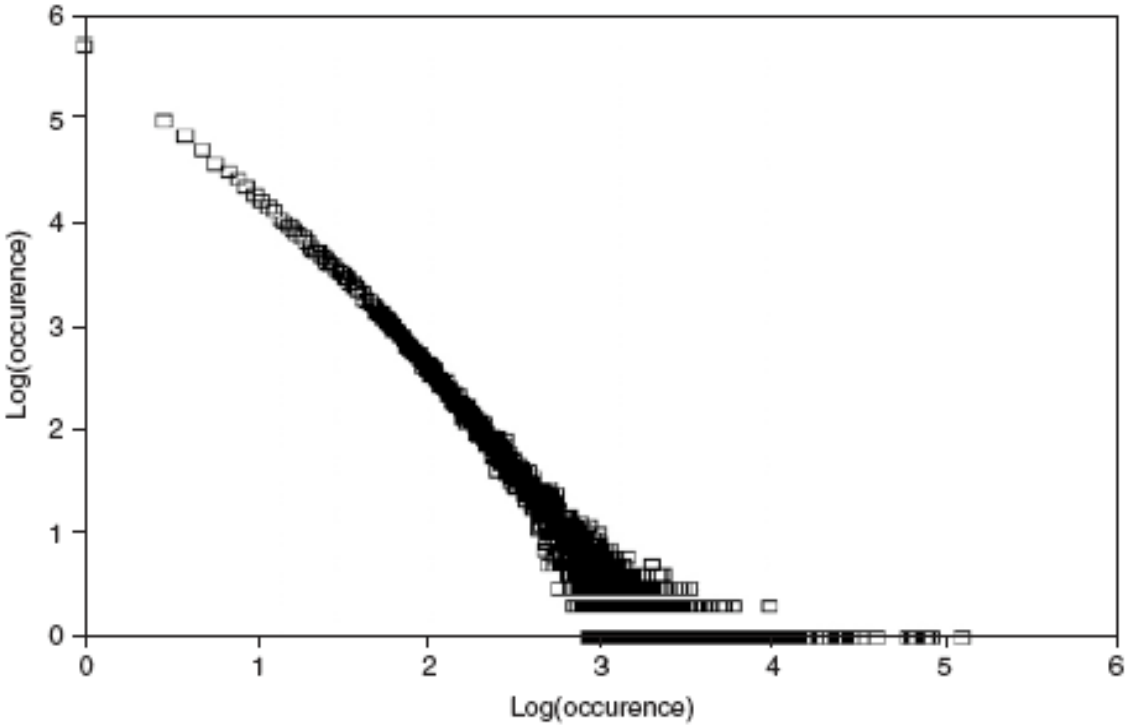
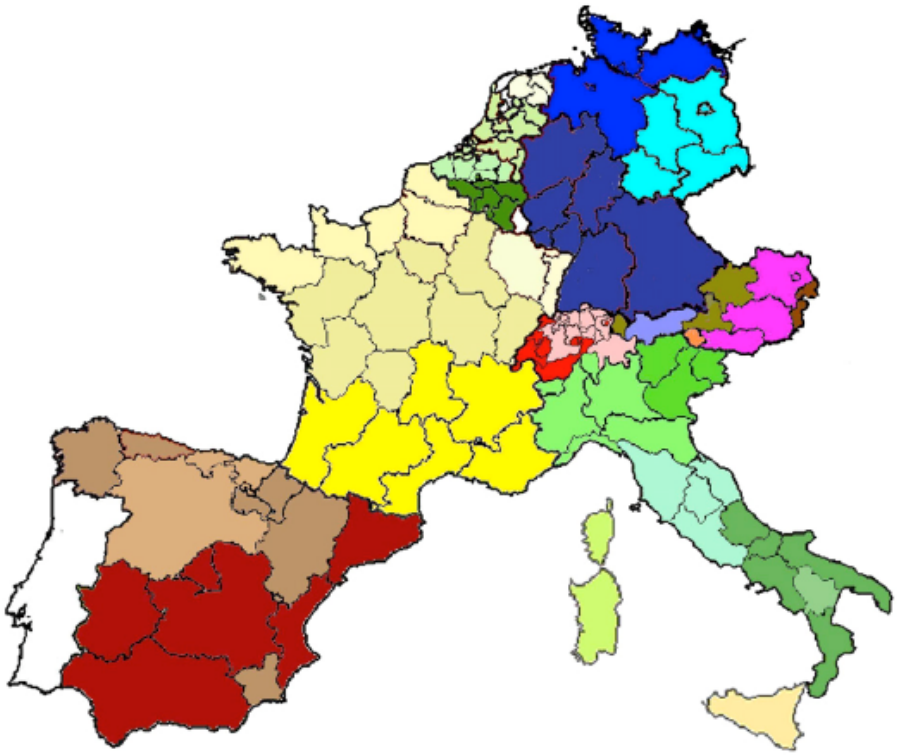


Fig. 2. The log-log distribution of the frequency of occurrence of surnames in Western Europe.



14. La distribuzione dei cognomi e la fisica statistica

L'interesse dei fisici teorici per la creazione e lo studio di modelli stocastici capaci di cogliere alcuni aspetti della dinamica dell'evoluzione biologica risale alla fine degli anni Ottanta del XX secolo, in corrispondenza dell'affermarsi della teoria neutralista dell'evoluzione (per cui la gran parte della variabilità individuale non ha effetti importanti sulla *fitness*), che si presta molto bene a rappresentazioni che sono tipiche dei sistemi studiati in meccanica statistica, e grazie ai notevoli sviluppi della teoria dei sistemi disordinati, nel cui ambito entrano abbastanza naturalmente i modelli di evoluzione neutra. Ricordiamo qui, tra i lavori che in qualche modo anticipano le problematiche di interesse in questo contesto, l'articolo di Derrida e Peliti[59] del 1991 sull'evoluzione di un modello per popolazioni di individui che si riproducono in modo asessuato, per il quale risulta possibile calcolare no solo la statistica delle genealogie e la variabilità genetica (che sono i corrispettivi del grado di consanguineità e della distribuzione in frequenza) ma anche le fluttuazioni di queste quantità, tipiche dei sistemi disordinati. Nell'articolo di Serva e Peliti[60] dello stesso anno viene invece esaminato un modello di riproduzione sessuata, trovando che in questo caso le fluttuazioni nella distanza genetica tra gli individui tendono a svanire nel limite di popolazione infinita.
Dopo la pubblicazione dei risultati di Miyazima *et al.*[61] a proposito dei cognomi nelle città giapponesi, Zanette e Manrubia[62] si indirizzarono al problema della distribuzione in frequenza dei cognomi, dimostrando che in un modello stocastico (proposto da Simon nel 1955) con popolazione (esponenzialmente) crescente e con probabilità non nulla α per la comparsa di nuovi cognomi a ogni generazione, la distribuzione tende asintoticamente nel tempo a una legge di potenza, con esponente dipendente da α, e tendente a 2 quando α tende a 0.
Un modello alternativo di spiegazione fu proposto nel 2002 da Reed e Hughes[63], che mostrarono che, se un processo stocastico caratterizzato da una crescita esponenziale è interrotto (oppure osservato) in modo casuale, la distribuzione dello stato osservato segue (almeno asintoticamente) una legge di potenza. Considerando quindi l'evoluzione della distribuzione dei cognomi come un processo di diramazione di tipo Galton-Watson con l'aggiunta di una probabilità finita per la comparsa di nuovi cognomi (per mutazione o per immigrazione), si dimostra che la distribuzione dei cognomi segue una legge di potenza, con un esponente dipendente dalla probabilità di comparsa di nuovi cognomi e dal coefficiente di crescita della popolazione.

Più di recente alcuni studi originati dall'analisi sia diacronica che sincronica dei cognomi coreani[64] hanno condotto alla formulazione di un modello di dinamica delle popolazioni formulato in termini di una *master equation*, nella quale appaiono come variabili le probabilità P(j,k;s,t) che una famiglia

---

[59] B. Derrida and L. Peliti, *Evolution in a flat fitness landscape*, Bull. Math. Biol. 53 (1991) 355
[60] M. Serva and L. Peliti, *A statistical model of an evolving population with sexual reproduction*, J. Phys. A: Math. Gen. 24 (1991) L705
[61] S. Miyazima, Y. Lee, T. Nagamine and H. Miyajima, *Power Law Distribution of Family Names in Japanese Societies*, Physica A 278 (2000) 282
[62] D. H. Zanette and S. C. Manrubia, *Vertical transmission of culture and distribution of family names*, Physica A 295 (2001) 1; S. C. Manrubia and D. H. Zanette, *At the Boundary between Biological and Cultural Evolution: the Origin of Surname Distributions*, J. Theor. Biol. 216 (2002) 461; S.C. Manrubia, B. Derrida and D.H. Zanette, *Genealogy in the Era of Genomics*, Am. Scient. 91 (2003) 158
[63] W.J. Reed and B.D. Hughes, *From gene families to incomes and internet file sizes: Why power laws are so common in nature*, Phys. Rev. E 66 (2002) 067103; W.J. Reed and B.D. Hughes, *On the distribution of family names*, Physica A 319 (2003) 579
[64] B.J. Kim and S. M. Park, *Distribution of Korean Family Names*, Physica A 347 (2005) 683;
H.A.T. Kiet, S.K. Baek, B.J. Kim, H. Jeung, *Korean Family Name Distribution in the Past*, J. Korean Phys. Soc. 51 (2007) 1812; S.K. Baek, H.A.T. Kiet and B.J. Kim, *Family name distributions: Master equation approach*, Phys. Rev. E 76 (2007) 046113

abbia k membri al tempo t se aveva j membri al tempo s, mentre compaiono come parametri le probabilità di nascita, di morte e di mutazione di cognome di un membro della famiglia al tempo t (che sono tutte proporzionali a k e hanno, per ipotesi, la stessa dipendenza dal tempo). Quest'equazione può essere formalmente risolta per valori dati di j e s, e in particolare nel caso j=1 (ossia assumendo la comparsa della famiglia al tempo s), e pertanto la distribuzione complessiva dei cognomi nella popolazione si può ricavare dalla conoscenza del numero di cognomi introdotti a ciascun tempo s. Il modello così formulato è molto generale, e comprende per esempio il modello di Zanette e Manrubia come caso particolare. È possibile rappresentare il caso di una popolazione in cui siano presenti sia mutazione che immigrazione assumendo che il numero di cognomi introdotti al tempo s abbia due componenti, di cui una costante che rende conto dell'immigrazione, e l'altra proporzionale alla popolazione totale N, per tener conto delle mutazioni. In una popolazione che cresce esponenzialmente il numero di cognomi cresce nel tempo come il logaritmo di N, mentre la probabilità che una famiglia abbia k membri è (a grandi tempi) proporzionale a 1/k : questi risultati sembrano descrivere bene il caso della Corea (e quello della Cina), coerentemente con i dati storici noti. Se invece si ammette la possibilità di mutazioni il numero dei cognomi cresce proporzionalmente a N, mentre l'esponente della distribuzione è prossimo a 2, ma comunque dipende dal coefficiente di crescita della popolazione e dal coefficiente di mutazione.
Questi risultati sono stati di recente riprodotti da De Luca e Rossi mediante l'applicazione di tecniche di gruppo di rinormalizzazione mutuate dalla teoria dei campi statistica.[65]

Il problema della discrepanza tra il modello teorico di Karlin-McGregor e Fisher da un lato, e le parametrizzazioni basate su leggi di potenza dall'altro, può essere affrontato, e concettualmente risolto, tenendo conto del fatto che tutte le distribuzioni empiriche sono basate sui risultati derivanti da campionamenti in cui viene estratto un numero finito di elementi da un insieme che è comunque a sua volta finito. La distribuzione in frequenza dei risultati di un campionamento non ha, in linea di principio, la stessa forma della distribuzione di partenza: se pensiamo a un elenco telefonico come campione di una popolazione, possiamo aspettarci che i cognomi molto comuni siano presenti in misura proporzionale alla loro effettiva frequenza nell'intera popolazione, ma molti dei cognomi presenti una sola volta nella popolazione non saranno affatto presenti nell'elenco. È comunque possibile da un punto di vista teorico ricavare la distribuzione più probabile del campione a partire dall'ipotetica forma della distribuzione presente nell'intera popolazione e dalla conoscenza del rapporto tra le dimensioni del campione e quelle dell'insieme di partenza. Per famiglie molto generali di distribuzioni che, nel limite di popolazioni infinite, tenderebbero a una legge di potenza si possono trovare distribuzioni dei campioni che rappresentano generalizzazioni della distribuzione di Fisher. Per l'esattezza esiste una famiglia abbastanza semplice di distribuzioni (le cosiddette "distribuzioni binomiali negative" per le quali vale la proprietà che la media delle distribuzioni dei loro campionamenti (purché non troppo piccoli) è essa stessa una distribuzione binomiale negativa.
Si dimostra che, mentre per campioni finiti le distribuzioni non obbediscono esattamente a una legge di potenza, quando si considera il limite di popolazioni infinite vale una legge esatta di potenza, con un esponente reale arbitrario (compreso tra –1 e –2) La distribuzione di Fisher è un caso particolare di questa famiglia, corrispondente all'esponente –1.
Le distribuzioni empiriche potrebbero quindi essere parametrizzabili mediante questo tipo di modelli, estrapolando poi dal modello la legge di potenza che dovrebbe valere per la popolazione infinita. È interessante anche notare che esiste una sequenza di combinazioni pesate dei valori empirici della distribuzione ("momenti invarianti") il cui valore più probabile non dipende dalla dimensione del campione (sempre per campioni non troppo piccoli) e che di conseguenza obbedisce a un'opportuna legge di scala, ovviamente legata all'esponente della distribuzione limite, i cui momenti sono evidentemente gli stessi che quelli dei propri campioni.[66]

[65] A. De Luca and P. Rossi, *Renormalization group evaluation of exponents in family name distributions*, Physica A 388 (2009) 3609-3614
[66] P. Rossi, *On sampling and parametrization of discrete frequency distributions* (in preparazione)

15. Il linguaggio come sistema complesso

Un'applicazione relativamente recente della teoria dei sistemi complessi ai fenomeni sociali riguarda lo studio di diversi aspetti e proprietà dei linguaggi naturali.
Abbiamo già detto che una delle prime osservazioni empiriche di leggi di scala nei fenomeni sociali fu appunto la cosiddetta legge di Zipf, relativa alla frequenza con cui le parole compaiono nei testi.
Sviluppi assai più recenti in questa stessa direzione riguardano la possibilità di stabilire una relazione tra la frequenza con cui determinati gruppi di parole compaiono nei testi letterari e il concetto (introdotto da Shannon) di entropia dell'informazione[67], l'esistenza di correlazioni a grande distanza nell'uso delle parole in *corpora* letterari ampi[68], una spiegazione dell'origine della legge di Zipf nel quadro di un modello dinamico stocastico per la generazione dei testi[69].
Detto per inciso, non sono mancati nemmeno tentativi di applicazione di analisi quantitativa delle proprietà frattali al particolare "linguaggio" costituito dalla pittura astratta, in particolare ai fini dell'autenticazione di opere di cui è apparentemente facile produrre imitazioni, come i quadri di Jackson Pollock.[70]
Altri sviluppi hanno riguardato la classificazione tassonomica delle lingue naturali, in analogia con quanto già fatto nel contesto della classificazione dei generi e delle specie biologiche. Anche in questo caso è emersa nella tassonomia una struttura di autosimilarità, per cui la frequenza del numero dei *taxa* espressa come funzione del numero dei linguaggi appartenenti a ciascun *taxon* appare obbedire a una legge di scala.[71] La distribuzione del numero dei parlanti i diversi linguaggi risulta invece di tipo lognormale, e ciò potrebbe essere conseguenza della dinamica demografica.
Un filone di ricerche linguistiche ancor più direttamente connesso ai metodi e ai modelli della fisica statistica dei sistemi complessi riguarda la dinamica stessa dei linguaggi (*language dynamics*), e in particolare i processi che portano alla formazione delle lingue, alla loro evoluzione e alla loro estinzione. Il processo di formazione del linguaggio è visto come un caso particolare di dinamica sociale, notando che i modelli di dinamica sociale usualmente postulano una popolazione di agenti e un protocollo d'interazione binaria tra gli agenti, e che in questi modelli, malgrado le interazioni siano di natura "locale" è spesso possibile la comparsa di comportamenti cosiddetti "emergenti" e di correlazioni a grande distanza, che costituiscono, come abbiamo già evidenziato in precedenza, la premessa per la comparsa di fenomeni collettivi e di leggi di scala.[72]
La dinamica evolutiva dei linguaggi è stata studiata con l'uso di modelli matematici, in particolare da N. Komarova[73] e collaboratori. Anche la dinamica della scomparsa dei linguaggi è stata analizzata mediante modelli[74] sottoposti al confronto con i dati empirici relativi a numerose realtà storico-geografiche differenti.

Notiamo qui per inciso che la connessione tra genetica e linguistica, già rilevata nello studio dei cognomi, è stata esplorata sistematicamente da Cavalli-Sforza e collaboratori[75], anche in questo caso facendo uso di sofisticate tecniche di analisi di tipo fisico-matematico, come lo studio delle "componenti principali", e la costruzione di alberi filogenetici..

---

[67] M.A. Montemurro, D.H. Zanette, *Entropic analysis of the roleof words in literary texts,* Advances in Complex Systems 5 (2002) 7-17;
R. Mansilla and E. Bush, *Increase of Complexity from Classical Greek to Latin Poetry,* preprint cond-mat/0203135
[68] M.A. Montemurro and P. Pury, *Long-range fractal correlations in literary corpora,* Fractals 10 (2002) 451-457
[69] D.H. Zanette, M.A. Montemurro, *A stochastic model of text generation with realistic Zipf distribution*, Journal of Quantitative Linguistics 12 (2005) 29-40
[70] A. Abbott, *Fractals and art: In the hands of a master,* Nature 439 (2006) 648-650
[71] D.H. Zanette, *Self-similarity in the taxonomic classification of human languages,* Advances in Complex Systems 4 (2001) 281-286
[72] V. Loreto and L. Steels, *Social dymanics: Emergence of language*, Nature Physics 3 (2007) 758-760
[73] N.L. Komarova, *Population dynamics of human language: a complex system*, in "Frontiers of engineering", pp. 89-98, National Academy Press 2006
[74] D.M. Abrams and S.H. Strogatz, *Linguistics: Modelling the dynamics of language death,* Nature 424 (2003) 900
[75] L.L. Cavalli-Sforza, P. Menozzi, A. Piazza, *Storia e geografia dei geni umani*, Adelphi, Milano 1997

16. Sistemi dotati di scala

Il fatto che fino a questo momento abbiamo concentrato la nostra attenzione su fenomeni e processi caratterizzati dall'assenza di un scala intrinseca, e quindi spesso descrivibili mediante leggi di scala, non deve indurre a pensare che questi siano gli unici, tra i fenomeni appartenenti alla sfera delle scienze umane, suscettibili di essere modellati e studiati quantitativamente.
In realtà è probabilmente vero il contrario, e di certo in moltissimi casi è abbastanza immediato riconoscere che esistono scale spaziali (a partire dalle dimensioni tipiche del corpo umano) o temporali (a partire dalla durata tipica della vita umana) tali per cui molti fatti demografici sono descrivibili mediante distribuzioni per le quali le nozioni di media e di varianza conservano tutto il loro significato e tutta la loro pregnanza concettuale. Dovrebbe essere ovvio che se, a proposito di una popolazione, attuale o storica, parliamo di "vita media" o di "età media al matrimonio", così come se parliamo di "altezza media" o di "numero medio di Calorie disponibili pro-capite", stiamo riferendoci a quantità che possono certamente oscillare a livello individuale, ma con oscillazioni che sono sempre contenute all'interno di un intervallo di valori plausibili.

Lo studio dei sistemi dotati di una scala propria, se da un lato è facilitato dalla possibilità di utilizzare un grande numero di nozioni e di strumenti di analisi sviluppati da tempo nell'ambito delle scienze statistiche, è d'altra parte reso più complesso dalla perdita della nozione di universalità, e quindi anche della possibilità di un'interpretazione dinamica e di una spiegazione causale sufficientemente generale (come la teoria dei processi di Jule, o la criticità auto-organizzata).
Dal punto di vista della statistica descrittiva questo non è un problema, mentre dal punto di vista della statistica inferenziale la mancanza di una giustificazione causale e concettuale del modello può indebolire anche fortemente la capacità predittiva dello stesso, e quindi l'attendibilità delle relative estrapolazioni.

Non è questa la sede per un'esposizione anche soltanto superficiale di risultati di statistica descrittiva e di modellazioni che, particolarmente in campi come l'economia o la demografia, sono ormai innumerevoli, e che comunque non fanno in generale uso di nozioni originalmente peculiari alle scienze fisiche. Vorrei piuttosto concentrarmi su alcuni lavori recenti che sono caratterizzati specificamente dall'impiego di nozioni nate in contesto fisico, e che dimostrano quindi la fertilità del confronto di idee e di paradigmi propri di discipline anche concettualmente molto lontane l'una dall'altra.

17. Lo spazio cognitivo nei testi letterari

Sappiamo dalle scienze cognitive che i tentativi di rappresentare le percezioni umane mediante parametri quantitativi e misurabili sono irti di ostacoli materiali e concettuali, e ciò è certamente ancor più vero quando s'intenda ragionare di percezioni spaziali su grandi scale di distanza, per le quali è molto difficile immaginare protocolli sperimentali sufficientemente attendibili.
Esiste tuttavia almeno un caso in cui le percezioni soggettive sono, per così dire, "fossilizzate" e quindi indefinitamente ricontrollabili, ed è il caso dei testi scritti, poco importa se letterari, storiografici o documentari. La possibilità di trasformare le informazioni "qualitative" relative alla percezione spaziale dell'autore in dati "quantitativi" associati alla misura di ben definite "grandezze osservabili" ci viene in questo caso suggerita dalla meccanica classica, che ci offre le nozioni di "baricentro", "momento d'inerzia" e "assi principali", la cui possibile riconversione nel linguaggio dell'analisi dei testi sarà ben presto evidente.
Possiamo infatti definire, per ciascun testo che contenga un numero sufficiente di indicazioni spaziali, un "centro di percezione" che rappresenta il luogo più probabile nel quale si è collocato (fisicamente o psicologicamente) l'autore al momento della stesura del testo, una "raggio di percezione" che rappresenta la distanza dal centro alla quale si trova l'orizzonte cognitivo dell'autore (ossia il limite della regione spaziale al di fuori della quale gli eventi hanno un interesse limitato o trascurabile) e gli assi principali della percezione, ossia le direzioni preferenziali, all'interno dello spazio cognitivo verso le quali è orientata l'attenzione di chi scrive.
Questo tipo di analisi, sia nel caso dei testi letterari che in quello dei testi storiografici, ci offre importanti indicazioni sulla "geografia mentale" del testo (quando la delimlitazione dell'orizzonte cognitivo appare come una scelta volontaria) o anche dell'autore stesso (quando risulta più o meno evidentemente chiaro che i confini della percezione sono fissati inconsciamente dalla limitata "conoscenza del mondo" (o "attenzione al mondo") dello scrivente.
Un confronto tra opere diverse dello stesso autore può mettere in evidenza un'evoluzione cognitiva o anche un'evoluzione (o involuzione) culturale, che porta ad allargare (o a restringere) l'orizzonte cognitivo. Un confronto diacronico tra autori diversi che si occupano di materia affine può aiutarci a cogliere elementi evolutivi (o involutivi) nella percezione spaziale e geografica come effetto di trasformazioni della sensibilità (anche collettiva) nel corso del tempo.

Per ragionare adeguatamente su questi concetti ci occorre presentare alcune procedure operative e alcune definizioni formali. Occorre innanzitutto individuare, nel testo che interessa, tutte le occorrenze interpretabili come indicazioni geografiche e/o spaziali, esplicite (nomi di luoghi, di regioni, di monti, di fiumi, di edifici, etc) o anche implicite (perifrasi o allusioni a luoghi, etc).
Sia $o(n)$ il numero di occorrenze (citazioni dirette o implicite) dell'n-esimo "oggetto" geografico. Definiamo $w(n) = o(n)/\sum o(n)$ il "peso" dell'n-esimo oggetto, notando che con questa definizione la somma dei pesi è necessariamente uguale a 1.
Associamo poi a ogni oggetto le sue coordinate geografiche. latitudine $a(n)$ e longitudine $b(n)$. Siamo ora in grado di calcolare, con l'ausilio della trigonometria sferica, le "coordinate pesate"

$$X = \sum w(n) \cos a(n) \cos b(n), \qquad Y = \sum w(n) \cos a(n) \sin b(n), \qquad Z = \sum w(n) \sin a(n)$$

a partire dalle quali troviamo le coordinate geografiche A e B del centro di percezione:

$$A = \arctan Z/\sqrt{(X^2+Y^2)}. \qquad B = \arctan Y/X$$

Abbiamo quindi una definizione matematica (modellata su quella del baricentro) del punto dal quale, nel contesto cognitivo del testo in esame, ha maggiormente senso misurare le distanze. La conoscenza dei dati e dei risultati sopra elencati permette di valutare tali distanze "in linea d'aria", per tutti i casi in cui sia difficile o impossibile valutare con esattezza le distanze "sul terreno".
La distanza $d(n)$ dell'n-esimo oggetto dal centro di percezione è data, in formula, da

$$d(n) = T \arccos [\cos a(n) \cos A \cos (b(n)-B) + \sin a(n) \sin A]$$

dove T è il raggio terrestre. Poiché le fluttuazioni tendono a compensarsi, nella maggior parte dei casi l'errore insito nella approssimazione risulterà trascurabile.

A questo punto abbiamo la possibilità di definire il raggio di percezione *R* come media pesata delle distanze degli oggetti geografici citati nel testo dal centro di percezione:

$$R = \sum w(n)\, d(n)$$

Un'apparente difficoltà di questa definizione è legata al fatto che esistono oggetti geografici estesi (regioni, catene montuose, fiumi, etc) per i quali il concetto di coordinate geografiche non appare immediatamente applicabile. Tale problema può essere in realtà risolto in almeno due modi (che empiricamente portano a risultati molto simili): formalmente qualunque oggetto esteso ha un suo proprio baricentro, le cui coordinate possono essere utilizzate per caratterizzare l'intero oggetto; alternativamente , se il testo lo permette, conviene utilizzare gli elementi puntuali più "tipici" dell'oggetto stesso (la città principale di un territorio, il ponte principale su un fiume, il valico principale di una catena montuosa, etc).

Una nozione lievemente più sofisticata è quella di "ellisse di percezione", che tiene conto del fatto, già accennato in precedenza, che all'interno dell'orizzonte cognitivo esistono comunque direzioni preferenziali di osservazione. In particolare, con un formalismo per il quale rinviamo al lavoro originale[76], si possono individuare matematicamente a partire dai dati sopra utilizzati due direzioni, tra loro ortogonali, dette assi principali, delle quali una ("asse maggiore") è quella lungo la quale lo "sguardo" dell'autore si spinge più lontano. Lo spazio percettivo, che in prima approssimazione era un cerchio, risulta ora geometricamente deformato e si riduce a un'ellisse, incentrata nel centro di percezione che è anche il luogo in cui si incontrano gli assi principali.

Notiamo che il raggio di percezione è la "scala" del sistema in esame, che in questo caso è legata non tanto alla media della distribuzione *w(n)* (ossia alla posizione del centro di percezione) quanto piuttosto alla sua varianza. Può essere interessante anche notare che, almeno nei casi empirici presi in esame, la distribuzione dei luoghi geografici non sembra generalmente configurarsi come una distribuzione di tipo gaussiano, pur avendo in comune con la gaussiana la forma "a campana" e la rapida decrescita della probabilità di trovare elementi man mano che ci si allontana dal centro di percezione.

In molti casi non è realmente necessario procedere al calcolo del centro di percezione, perché il testo stesso ce ne offre un'indicazione empirica o "intuitiva" (che può essere comunque interessante confrontare con quella formale). In questi casi basterà operare con la formula delle distanze (o con le distanze empiriche note) per determinare, sempre grazie all'espressione precedente, il valore del raggio di percezione e quindi l'ampiezza dello spazio cognitivo.

Quando si abbia *a priori* una conoscenza intuitiva del centro di percezione è possibile un'ulteriore semplificazione della procedura operativa, basata sugli stessi presupposti concettuali ma volta a ridurne drasticamente gli adempimenti formali e computazionali. Si tratta di ordinare tutti i luoghi geografici indicati nel testo sulla base della loro distanza crescente dal centro di percezione, e quindi di sommare il loro pesi, a partire dal peso del centro stesso, fino a raggiungere il valore di 2/3. La distanza entro la quale sono racchiusi i 2/3 delle citazioni geografiche di un testo si può considerare in molti casi come una valida e pratica definizione alternativa del raggio di percezione di un particolare testo letterario.
Vedremo nel seguito alcune applicazioni di queste nozioni a tre testi storiografici del X secolo riferiti alla stessa area geografica (Francia settentrionale) e ad alcune opere di Dante Alighieri.

[76] P. Rossi, *Measuring Large Scale Space perception in Literary Texts,* Physica A 380 (2007) 439-446

18. L'orizzonte cognitivo di tre testi altomedievali

Quasi tutta la nostra conoscenza storiografica relativa alla Francia settentrionale nel X secolo riposa (oltre che su pochi diplomi e brevi cronache locali) essenzialmente su tre testi, tra loro in qualche modo collegati sia per origine che per contenuto, e dovuti a tre autori molto differenti, ma tutti gravitanti sulla città episcopale e metropolitana di Reims: il canonico Flodoard, Richer monaco di Saint Remi e Gerbert d'Aurillac, il futuro papa Silvestro II. Analizzeremo separatamente i tre testi con l'obiettivo di individuare i rispettivi spazi e orizzonti cognitivi.

Flodoard (893/4-966), a noi noto soltanto attraverso le sue opere, è l'autore di *Annales* che coprono con continuità il periodo dal 919 al 966, con una narrazione abbastanza "fredda" e impersonale, fortemente incentrata sulle vicende che riguardano la città di Reims (che costituisce l'ovvio centro di percezione del testo) e la sua provincia ecclesiastica.
Richer (di cui non conosciamo alcun dato biografico, se non i pochi accenni contenuti nel testo) è l'autore di un testo, pervenutoci nel manoscritto originario, e usualmente denominato *Historiae* o *Historiarum libri IV*. L'ambizione di Richer, che scrisse su commissione di Gerbert, suo maestro e all'epoca arcivescovo di Reims, è quella di narrare in modo originale la storia di Francia dall'888 ai suoi tempi (998), ma in realtà si tratta, per il periodo fino al 966, di un talvolta maldestro rifacimento di Flodoard, e per il seguito di una narrazione che risulta abbastanza confusa per tutti gli eventi ai quali l'autore non ha assistito personalmente. Anche nel suo caso Reims è il centro di percezione, ma in misura meno ovvia che nel caso di Flodoard.
Gerbert (nato in Alvernia intorno al 945/50, morto a Roma il 12 maggio 1003), fu uno dei personaggi più significativi del suo tempo: monaco benedettino ad Aurillac, studioso in Spagna, *magister* a Reims, poi dal 983 in sequenza abate di Bobbio, consigliere di Ugo Capeto, arcivescovo di Reims (991) poi di Ravenna (998) e infine Papa (999-1003) col nome di Silvestro II. Non scrisse testi propriamente storiografici, ma ci ha lasciato la raccolta delle proprie *Epistolae*: si tratta di 220 lettere, scritte tra il 983 e il 997, indirizzate ai più importanti personaggi del suo tempo, e quasi tutte di rilevante contenuto politico e/o culturale, per cui costituiscono un documento essenziale anche per la ricostruzione degli eventi del periodo cui si riferiscono.

Nel caso di Flodoard[77] il numero totale delle località esplicitamente citate nel testo è 174, per complessive 800 citazioni. A queste vanno poi aggiunte 7 località citate solo implicitamente, ma perfettamente identificabili. Nel nostro conteggio abbiamo incluso da un lato anche gli edifici (chiese, abbazie, palazzi reali e imperiali), dall'altro i distretti (*pagi*), le cui limitate dimensioni consentono un'associazione diretta con la località più importante del distretto stesso. Abbiamo invece escluso per semplicità di calcolo fiumi e monti (19 nomi e 114 citazioni), regioni (o regni) e relativi etnonimi (34 nomi e 430 citazioni).
Riferendo l'analisi alle 181 località di cui sopra, abbiamo calcolato un raggio di percezione di 180 Km. Il numero delle località che si trovano entro un raggio di 50 Km da Reims è 36 (20% del totale dei nomi, e 40% delle citazioni), entro 100 Km se ne trovano 66 (36% dei nomi e 58% delle citazioni) e 89 (49% dei nomi, 69% delle citazioni) distano da Reims meno di 150 Km. Il raggio di percezione, con questa definizione, include 95 località (oltre il 50%) e 574 citazioni (oltre il 70%).
Questi dati evidenziano la fortissima focalizzazione di Flodoard sulla regione più direttamente influenzata dalle dinamiche relative alla città di Reims e alla sua chiesa. Sono citate moltissime località, anche minori, appartenenti alla diocesi di Reims, e numerosi castelli della provincia ecclesiastica, ma quando usciamo dai confini della provincia di Reims gli unici riferimenti geografici riguardano le città episcopali, e con l'aumentare della distanza restano soltanto le sedi metropolitane, a parte il lungo elenco di vescovadi tedeschi relativo al sinodo di Ingelheim:

[77] Fodoard, *Annali (919-966),* Introduzione, traduzione, note ed excursus di P. Rossi, PLUS, Pisa 2007:

evidentemente la circostanza fu percepita come epocale e di rilevanza continentale, e questo spiega e giustifica anche il momentaneo allargamento dell'orizzonte geografico.
L'analisi quantitativa permette anche di individuare un asse principale della percezione, che nel caso in esame tuttavia non fa che confermare un fatto intuitivamente prevedibile, trattandosi di un uomo di chiesa, ovvero una polarizzazione dell'attenzione verso ciò che avviene a Roma che è del tutto sproporzionata rispetto alla distanza fisica dell'Urbe dai luoghi su cui si concentra l'interesse "naturale" dell'autore. Ma è possibile ripetere l'analisi depurandola dai dati relativi alla città di Roma. Si trova allora che il raggio di percezione "crolla" a 123 Km (pur continuando a includere all'incirca i 2/3 delle citazioni residue) e l'asse principale si riorienta su una direzione nordest-sudovest che è quella che ci si potrebbe comunque aspettare per una regione, come la Champagne di Reims, che fa da cerniera tra la Germania renana di Aquisgrana e Colonia e quella *Francia* robertingia che presto troverà il suo centro di gravità a Parigi.
L'altro asse principale, matematicamente ortogonale al precedente, è anch'esso non affatto casuale, perché corrisponde alla direttrice che va da Laon a Châlons, a sua volta immediatamente riconoscibile come il tratto *champenois* della *Via Francigena*, il percorso dei pellegrini (in particolare Angli, più volte citati da Flodoard) diretti verso Roma. È interessante notare il numero delle *stationes* della *Francigena*, anche lontane da Reims, citate da Flodoard (da Pavia a Vercelli, da Saint-Maurice a Brienne, da Châlons a Reims, da Corbeny a Laon, da Thérouanne a Guines),

Nel caso di Richer[78] il numero totale degli etnonimi e toponimi distinti è circa 230 (più una ventina di nomi di chiese), con un numero complessivo di citazioni pari a circa 1300. In quest'ambito, e con qualche piccola e inevitabile arbitrarietà nella classificazione, abbiamo contato 143 nomi di località (*urbs, civitas, castrum, oppidum, palatium, abbatia, monasterium*), cui corrispondono 621 citazioni, 19 nomi di fiumi (con 74 citazioni), 25 nomi di regioni o territori (*regio, pagus, terra, ager*) (con 298 citazioni), 27 nomi di popoli (con 262 citazioni). La restante quarantina di citazioni corrisponde ai nomi di monti, mari e continenti che compaiono nell'introduzione geografica [I.1-2].
Esiste poi un centro naturale di percezione, che può quasi sempre essere identificato mediante l'addensarsi dei riferimenti geografici, e che in Richer è ovviamente la città di Reims (citata ben 113 volte). Possiamo quindi ordinare i riferimenti geografici, oltre che secondo la loro frequenza, sulla base della loro distanza da Reims. Scopriamo allora che il 50% delle citazioni di località si trova entro un raggio di 105 Km da Reims, mentre i 2/3 si trovano entro un raggio di 200 Km.
Un'analisi tecnicamente più sofisticata, che permette di includere anche gli altri riferimenti geografici, porta comunque a definire un raggio di percezione valutabile intorno ai 200 Km (che si ridurrebbero a poco più di 150 Km se escludessimo dal calcolo il nutrito gruppo dei richiami alla città di Roma, che non sono certo riconducibili a una particolare attenzione da parte di Richer per tutto ciò che avveniva nell'ampio spazio geografico compreso tra Reims e la sede del Papato. È interessante notare che si tratta di un valore di poco ma significativamente superiore a quello di Flodoard, e che le più significative differenze tra i due testi sono quasi tutte riconducibili all'insistente attenzione di Reims verso la regione di Liegi, che conferma la nostra interpretazione di questo indizio come possibile spia dell'origine geografica della famiglia di Richer
I valori succitati ci offrono una misura quantitativa, e anche abbastanza accurata, dello spazio percettivo di Richer, e ci mostrano chiaramente che la sua *Gallia* è in realtà poco più ampia della provincia ecclesiastica di Reims. Questo spazio ha a sua volta due direzioni privilegiate (tecnicamente gli assi principali della distribuzione delle citazioni), che sono approssimativamente perpendicolari tra loro, e naturalmente si incrociano a Reims. Uno degli assi corrisponde a una direttrice NE-SO che va da Liegi a Chartres e rappresenta in qualche modo anche il percorso esistenziale di Richer. Il secondo asse invece, sulla direttrice NO-SE, coincide sostanzialmente con la tratta francese della *via Francigena*, dal Passo di Calais alle Alpi Pennine (Passo del Gran San Bernardo), con un ovvio prolungamento verso Pavia e Roma.

[78] Richer di Saint-Remi, *I quattro libri delle Storie (888-998)*, Introduzione, traduzione e note di P. Rossi, PLUS, Pisa 2008

Proprio all'epoca di Richer (e più precisamente a una data compresa tra il 990 e il 994) risale la prima testimonianza documentaria della *Francigena*, ovvero l'itinerario di Sigeric arcivescovo di Canterbury. E dall'analisi del testo di Sigeric possiamo ricavare un altro dato interessante: la lunghezza tipica di una tappa giornaliera lungo la *Francigena* è di circa 27 Km, se lo si calcola nel tratto tra il Giura e il passo di Calais, che comprende 22 *stationes*, di cui 13 a sud di Reims. Ciò significa che, tradotto in giornate di cammino, lo spazio percettivo di Richer corrisponde all'incirca a una settimana. Possiamo rileggere in questa chiave il famoso viaggio a Chartres [IV.50], notando che la distanza Reims-Chartres è di circa 200 Km, e quindi si tratta di un viaggio ai confini dello spazio percettivo, almeno sulla base di ciò che il testo ci lascia intendere sul suo autore.

Nel caso di Gerbert[79] il numero totale degli etnonimi e toponimi distinti è circa 160, con un numero complessivo di citazioni pari a poco circa 540. In quest'ambito, e con qualche piccola e inevitabile arbitrarietà nella classificazione, abbiamo contato 106 nomi di località (*urbs, civitas, castellum, curtis, abbatia, monasterium, cella*), cui corrispondono 324 citazioni, 28 nomi di regioni, territori e fiumi (con 83 citazioni), 23 nomi di popoli (con 139 citazioni).
Ma il conteggio cambia significativamente se eliminiamo tutti i riferimenti classici, biblici e religiosi (un esempio per tutti, l'aggettivo "Romano", che si riferisce quasi sempre al Pontefice o alla Chiesa, e che compare 34 volte): si tratta di due dozzine di nomi e di circa 100 citazioni. Se vogliamo studiare l'orizzonte cognitivo di Gerbert a livello sincronico dobbiamo quindi focalizzare l'attenzione soltanto su 92 nomi di località, per un totale di 295 citazioni, oltre che su 40 nomi di regioni, popoli e fiumi, corrispondenti ad altre 150 citazioni. Si tratta di un insieme di dati relativamente limitato, ma sufficiente per la nostra analisi.
Il centro di percezione è anche per Gerbert ovviamente la città di Reims (citata ben 69 volte). Possiamo quindi ordinare i riferimenti geografici, oltre che secondo la loro frequenza, sulla base della loro distanza da Reims. Scopriamo allora che il 50% delle citazioni di località si trova entro un raggio di 130 Km da Reims, mentre i 2/3 si trovano entro un raggio di 200 Km. Un'analisi tecnicamente più sofisticata, che permetterebbe di includere anche gli altri riferimenti geografici, porta comunque a definire un raggio di percezione valutabile intorno ai 280 Km. È interessante notare quanto più grandi sono qui i valori rispetto a quelli trovati per Flodoard e Richer. I dati così analizzati ci offrono una misura quantitativa, e anche abbastanza accurata, dello spazio percettivo di Gerbert, e ci mostrano chiaramente che esso è assai più ampio della provincia ecclesiastica di Reims, i cui centri più importanti si trovano tutti nel raggio di un centinaio di Km da Reims. Nella determinazione del raggio di percezione entrano chiaramente in modo non marginale gli altri poli d'attrazione che caratterizzarono l'esistenza di Gerbert, ovvero la Catalogna (6 citazioni), Roma (15 citazioni), l'area di Bobbio (10 citazioni), l'Alsazia (5 citazioni).
Anche in questo caso l'asse maggiore è una direttrice NO-SE che coincide sostanzialmente con la *via Francigena*, un itinerario che Gerbert percorse numerose volte. L'altro asse, come si è già detto, corrisponde a una direttrice NE-SO che va dalla Germania renana alla valle della Loira.

Per la cartografia rilevante alle analisi sopra presentate e per altre informazioni sui testi e gli autori si rimanda al materiale contenuto alla pagina *web* http://www.df.unipi.it/~rossi/storia.html

[79] : Gerbert d'Aurillac, *Lettere (883-897)*, Introduzione, traduzione e note di P. Rossi, PLUS, Pisa 2009

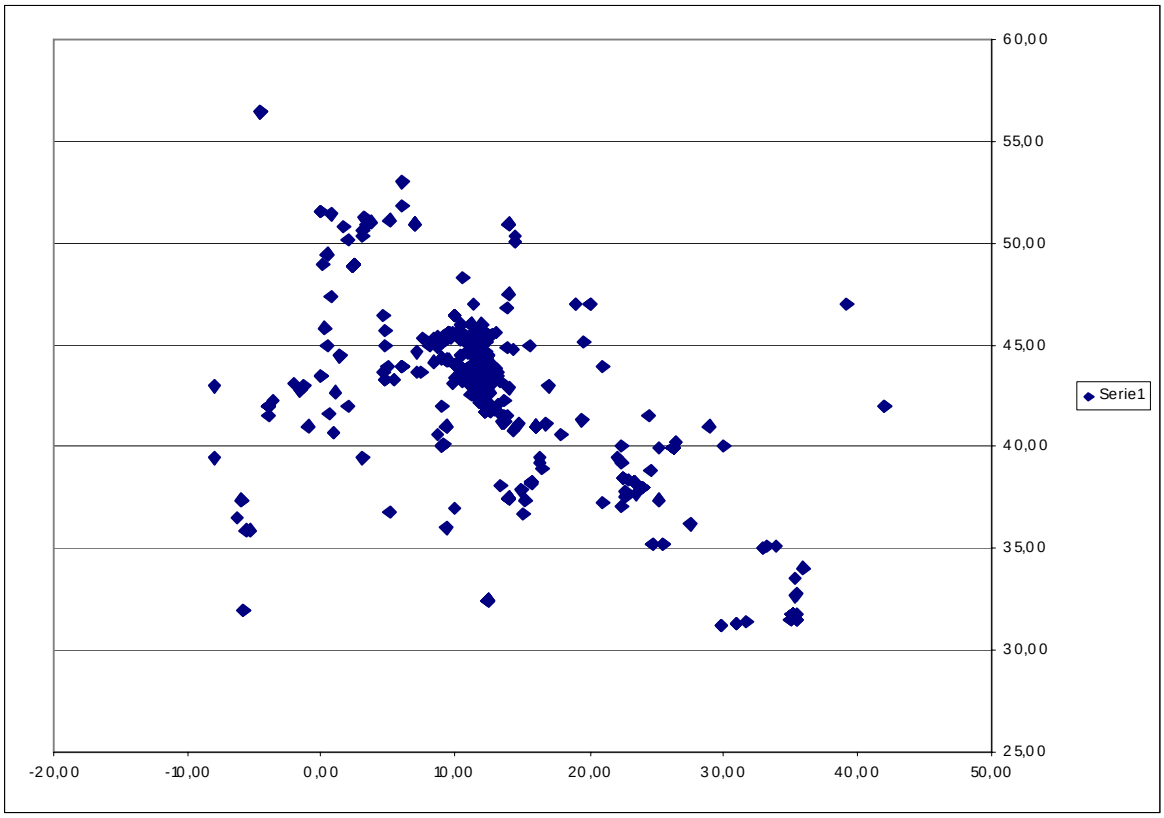
19. L'orizzonte cognitivo dei testi danteschi

Tra le molteplici letture che si possono dare, e che sono state date, dei testi danteschi non è del tutto priva di interesse anche un'analisi della "geografia" di Dante che, sviluppata secondo i paradigmi che abbiamo introdotto, può offrirci qualche nuova informazione, da un lato sulla vastità dell'orizzonte cognitivo (e culturale) dell'Alighieri, che emerge chiaramente, come vedremo, da una ricognizione del testo della *Commedia*, e dall'altro sul particolare "punto di vista" da questi assunto in merito al problema della lingua, tema discusso in particolare nel *De Vulgari Eloquentia*.

Una ricognizione della *Commedia* alla ricerca dei riferimenti geografici è oggi certamente facilitata dalla disponibilità, anche in rete, di accurati repertori dei vocaboli usati da Dante nella sua massima opera, ma non può comunque prescindere da un'ispezione diretta del testo, in quanto non sono rare nella *Commedia* le indicazioni geografiche espresse mediante perifrasi nelle quali può anche non comparire alcun nome proprio di località, o menzionando edifici che solo implicitamente rimandano alla località in cui si trovano.
Compiendo quest'operazione abbiamo individuato nel testo (con un piccolo margine di arbitrarietà nella interpretazione del concetto di "indicazione geografica") 743 citazioni di toponimi o etnonimi, riconducibili a 354 distinte località, 253 delle quali sono citate una sola volta, mentre 43 sono citate due volte e 21 sono citate tre volte. Abbiamo poi riportato in una tabella, ordinate per frequenza decrescente (e distanza crescente da Firenze), le 37 località citate almeno quattro volte.

Firenze	53	Lombardia	9	Fiesole	5	Casentino	4
Roma	37	Appennino	8	Arezzo	5	Pisa	4
Italia	24	Lucca	8	Alpe	5	Padova	4
Toscana	16	Romagna	8	Puglia	5	Provenza	4
Troia	15	Francia	8	Sardegna	5	Germania	4
Tebe	11	Gerusalemme	7	Ebrei	5	Parnaso	4
Arno	10	Bologna	6	Giudei	5	Atene	4
Grecia	10	Po	6			Mar Rosso	4
Siena	9	Sicilia	6			Etiopia	4
Mantova	9	Spagna	6			India	4

Già una superficiale analisi qualitativa dei dati mostra l'ampiezza dell'orizzonte cognitivo dantesco, e forse soprattutto l'ampiezza dell'orizzonte culturale, vista l'abbondanza dei riferimenti classici e biblici. Ed è già abbastanza evidente che il centro di percezione si colloca a Firenze, o comunque in Toscana. La formula matematica, in effetti, conferma quest'ovvia constatazione indicando per il centro le coordinate geografiche di un punto situato pochi Km a est del capoluogo toscano.
Il calcolo formale del raggio di percezione produce il valore di 673 Km, abbastanza strabiliante se lo confrontiamo con i valori altomedievali ricavati in precedenza, ma anche in valore assoluto molto elevato se ragioniamo sui tempi che, all'epoca di Dante, dovevano essere in media necessari per percorrere quelle distanze. Anche in tal senso l'orizzonte cognitivo di Dante è in primo luogo un orizzonte culturale, e denota un interesse per il "mondo" che va ben al di là dell'insieme dei luoghi più o meno facilmente raggiungibili. Nessuna particolare sorpresa ci viene invece dall'individuazione dell'asse principale,, che si dispone da nord-ovest a sud-est seguendo approssimativamente l'orientamento della penisola italiana.
Non è invece privo di interesse l'esercizio consistente nel disporre i luoghi citati, sulla base delle loro coordinate geografiche, su un piano cartesiano. Vediamo così, grazie all'addensarsi dei punti, comparire un'immagine abbastanza chiara dell'Italia, mentre per il resto del Mediterraneo e dell'Europa sono di fatto evidenziati soltanto i contorni delle coste. Questa rappresentazione rende graficamente evidente il ruolo prevalente delle comunicazioni marittime anche nella trasmissione della conoscenza dei luoghi e di conseguenza nella loro percezione.



L'analisi del *De Vulgari Eloquentia* (DVE) può essere condotta con la stessa metodologia, ma occorre avere ben chiaro che l'obiettivo dell'indagine è parzialmente differente, e che la definizione stessa dei "luoghi" che devono essere presi in esame è suscettibile di modificarsi sulla base delle differenti finalità che s'intende dare allo studio. DVE, in effetti, si compone di parti differenti, delle quali la prima, finalizzata a un discorso generale sulle lingue, si muove in un contesto ampio, che si potrebbe anche definire "europeo", ricco di riferimenti geografici e in qualche misura assimilabile al contesto della *Commedia,* per quanto con un più evidente sbilanciamento verso il settentrione. Le altre due parti si concentrano invece sui dialetti della Penisola e sul loro uso. L'orizzonte cognitivo si restringe quindi deliberatamente, e il punto di vista si focalizza maggiormente sull'Italia centrale, con un cambio di prospettiva anche quantitativamente abbastanza significativo quando l'attenzione si sposta dagli usi "generici" della lingua a quelli "poetici". In quest'ultima parte i riferimenti geografici sono assai spesso risultanti dall'identificazione personale degli autori citati, e quindi sono condizionati dalla dinamica della produzione poetica duecentesca.
Siamo comunque sempre in presenza di un raggio di percezione non inferiore ai 200 Km, con un centro che certamente si situa in Toscana ma non è necessariamente coincidente con Firenze, e tende anzi a spostarsi verso lo spartiacque appenninico (come del resto già ipotizzato, con l'uso di tutt'altri strumenti concettuali, da Umberto Carpi[80]), e con un asse principale che, anche in questo caso, non può discostarsi più che tanto dall'asse diretto da nord-ovest a sud-est che definisce l'orientamento della penisola italiana.

[80] U. Carpi, *La nobiltà di Dante*, Polistampa, Firenze 2004

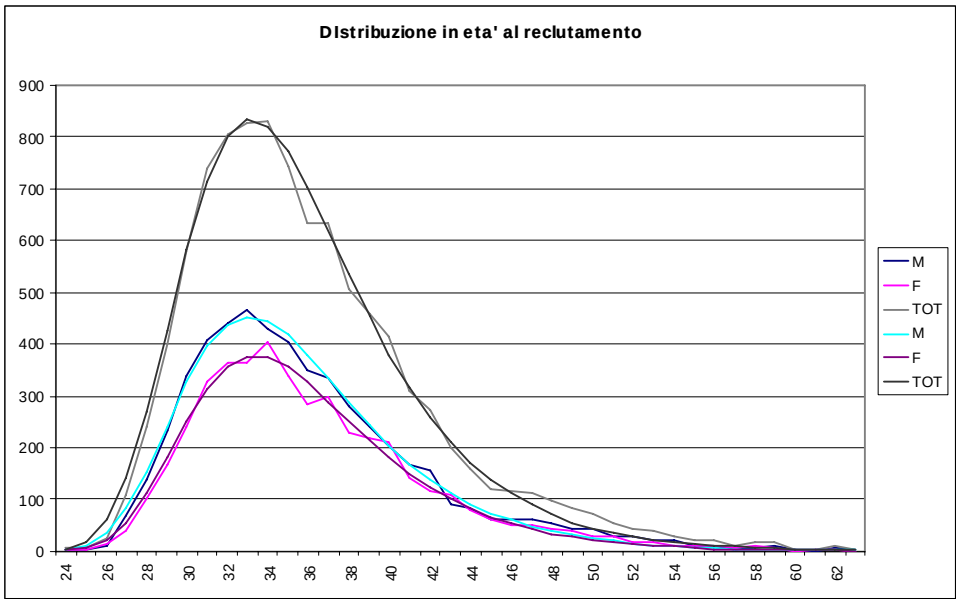
20. Il reclutamento universitario

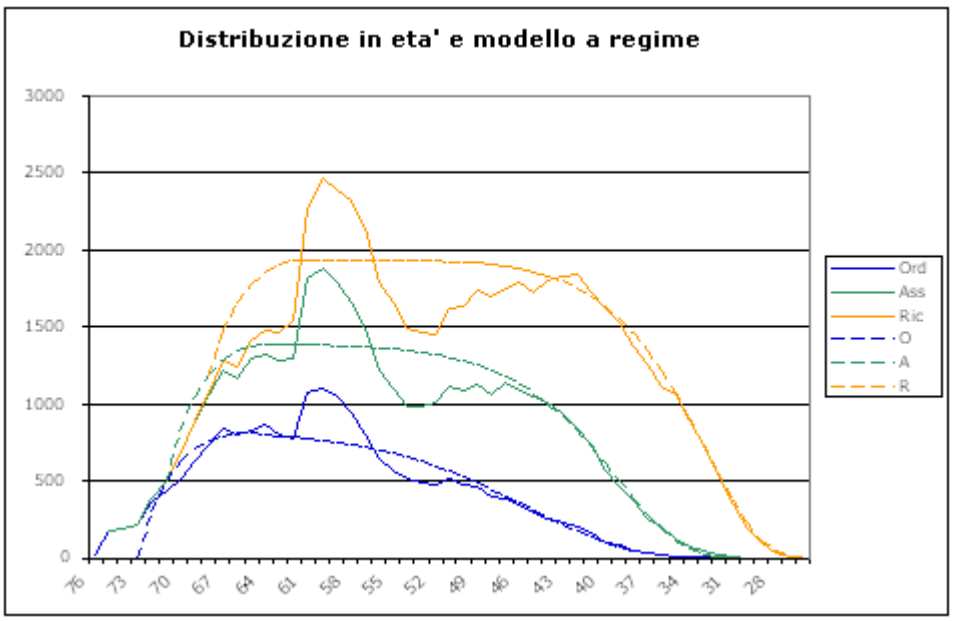
Le dinamiche sociali appaiono quasi sempre come processi estremamente complessi, condizionati da un grande numero di concause, e per i quali risulta pertanto difficile effettuare predizioni anche solo qualitativamente attendibili sulla base delle informazioni a disposizione in un determinato momento. A maggior ragione quindi non ci si aspetta di solito di poter formulare previsioni che siano quantitativamente accurate. Un tipico esempio di questa situazione sono le previsioni economiche di medio periodo che, in perfetta analogia con quelle meteorologiche, tendono a oscillare tra l'assoluta banalità ("dopo il freddo viene il caldo, dopo la crisi viene la ripresa") e la totale inattendibilità. Abbiamo anche visto che, almeno nel caso della meteorologia e probabilmente anche per l'economia, ci sono anche motivi teorici per postulare l'impossibilità di certi tipi di previsioni. Esistono tuttavia anche processi sociali governati da leggi più semplici, e ciò avviene in particolare quando un qualche elemento di rigidità strutturale interviene a ridurre la possibilità di fluttuazioni, come le anse di un fiume, che in pianura possono cambiare continuamente, ma sempre restando dentro una regione limitata dal fatto che l'acqua si muove verso il basso.
Molti fenomeni demografici appartengono a questa tipologia: difficilmente possono verificarsi brusche variazioni degli indici di natalità e di mortalità, e poiché l'andamento complessivo della popolazione dipende da questi due indici, che a loro volta sono numeri percentualmente piccoli, possiamo aspettarci di poterlo modellare in modo relativamente semplice e di poter effettuare previsioni su un arco di tempo non troppo limitato. Le equazioni che governano la dinamica di una popolazione dipendono in modo lineare dai coefficienti summenzionati, e di conseguenza esse risultano numericamente "integrabili" (ossia risolubili) e la bontà dell'estrapolazione è condizionata soltanto dal grado di stabilità temporale degli indici, o comunque dal nostro grado di conoscenza della loro evoluzione tendenziale.
A prima vista pochi scommetterebbero sul fatto che un'analisi di questo genere possa essere ripetuta utilmente per un processo apparentemente casuale e incontrollato come il reclutamento dei docenti universitari. Un'analisi dei dati empirici estesa a un periodo sufficientemente lungo rivela invece l'esistenza di alcuni elementi di stabilità strutturale, di cui è interessante anche indagare l'origine, tali per cui, sotto ipotesi abbastanza generali, è possibile effettuare previsioni sull'evoluzione del sistema quando esso non sia sottoposto a interventi esterni pesanti e traumatici (come ad esempio sostanziali modifiche della legislazione o sostanziali riduzioni del finanziamento).
I motivi teorici (riconoscibili *a posteriori*) per tale stabilità possono essere fatti risalire alla considerazione che, in una determinata fase dello sviluppo demografico ed economico-culturale del Paese, il numero di individui "adatti" e "disposti" alla carriera universitaria nati in ciascuna classe d'età non può che essere pressappoco costante. Di conseguenza, anche se è vero che il numero dei reclutati per anno solare è soggetto a continui sbalzi, il sistema universitario tenderà ad assorbire sempre lo stesso numero di individui per ciascun anno di nascita, mentre ciò che cambia è il tempo di attesa prima del reclutamento, che è una variabile molto elastica (esistono vari meccanismi sociali di "compensazione", dal sostegno familiare al precariato) e può quindi dilatarsi anche in misura considerevole.
Trattandosi evidentemente di un sistema dotato di una scala temporale intrinseca (legata alla durata media della vita umana) non ci dobbiamo aspettare andamenti a legge di potenza, ma piuttosto curve di tipo "demografico", la cui evoluzione temporale e il cui eventuale profilo "a regime" possono essere valutati abbastanza agevolmente a partire dalla conoscenza di alcune variabili empiricamente determinabili.
Senza entrare nel dettaglio della formulazione matematica del problema, per la quale si rinvia agli articoli originali[81], vogliamo soltanto qui ricordare i principali risultati.

---

[81] P.Rossi, *Un modello formale per la programmazione degli accessi e delle carriere negli E.P.R.,* Nuovo Saggiatore 10; 2 (1994) 33-38; *Le dinamiche di reclutamento e di carriera dei fisici* Nuovo Saggiatore 23; 3-4 (2007) 3-13

In primo luogo si conferma che la popolazione complessiva e la sua composizione anagrafica e per fasce e la loro variazione nel tempo sono determinate una volta che si conosca la distribuzione che rappresenta il numero medio di individui reclutati per anno come funzione dell'età al reclutamento (e della fascia docente), oltre che ovviamente l'età media di pensionamento
A sua volta la distribuzione in età al reclutamento ha una forma universale: si tratta di una distribuzione di Gompertz (che storicamente fu introdotta per rappresentare la dipendenza della probabilità di morte dall'età anagrafica). La distribuzione di Gompertz dipende da due parametri, riconducibili alla media e alla varianza. La varianza risulta essere indipendente dal tempo, mentre l'età media al reclutamento varia, con una precisa relazione matematica (di tipo lineare), al variare del numero dei reclutati per anno, e tende necessariamente a innalzarsi tanto più quanto più piccolo è il rapporto tra il numero dei reclutati e quello dei "reclutabili" (nel senso che abbiamo specificato in precedenza).





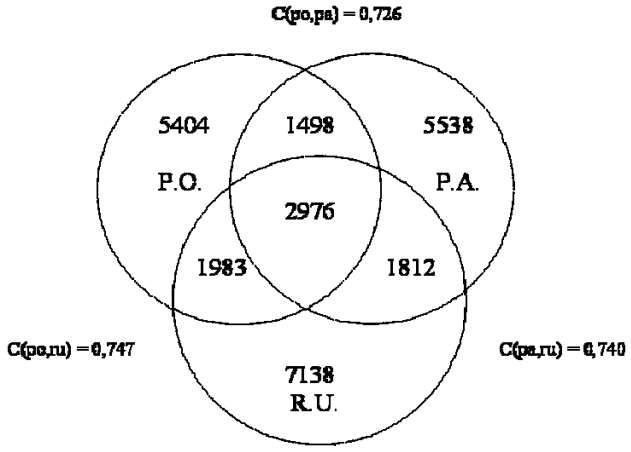
Lo studio della popolazione dei docenti universitari italiani, della quale esiste una base di dati facilmente accessibile in rete, ci permette anche di esemplificare l'utilizzo di alcuni concetti che sono stati introdotti[82] nel contesto dello studio delle distribuzioni in frequenza dei campioni di una popolazione. Abbiamo già osservato che le frequenze misurate sulle distribuzioni dei campioni sono connesse, con relazioni determinate dal calcolo delle probabilità, alle frequenze presenti nell'intera popolazione, ma dovrebbe essere evidente che, quando confrontiamo due campioni differenti, non troveremo in generale, soprattutto per piccoli campioni o per specie scarsamente presenti, che una particolare specie compare con la stessa frequenza nei due campioni.
In particolare, data una popolazione di N elementi e campioni di n elementi ciascuno, e usando la notazione $1/\alpha$ per il secondo momento invariante (che come abbiamo detto assume in media lo stesso valore per tutti i campioni indipendentemente dalla loro dimensione), possiamo calcolare la correlazione tra due campioni e, se la formazione dei campioni fosse realmente casuale, dovrebbe valere la relazione

$$C = n/N * (\alpha + N)/(\alpha + n)$$

Come si vede facilmente, la correlazione tende a 1, come è ovvio che sia, quando n tende a N, ossia quando i campioni coincidono con l'intera popolazione, ma tende a 0 al decrescere di n.

Nel caso dei docenti universitari possiamo studiare separatamente le distribuzioni dei cognomi di ricercatori, associati e ordinari, considerando ognuno dei tre gruppi come un campione casuale della popolazione italiana. Lo scostamento dei valori empirici di correlazione da quelli teorici sopra indicati potrà essere interpretato come una misura della non casualità dei campioni (nepotismo accademico). I risultati sono presentati nelle Tabelle, e possono essere interpretati notando che α è legati all'inverso dell'isonimia e vale circa 6.000, mentre n per ognumno dei tre campioni vale circa 20.000, per cui il valore atteso di C per campioni non correlati è prossimo a 0,75.

	ITALIA (elenchi tel.)	**UNIVERSITA' (modello)**	**UNIVERSITA' (dati)**
INDIVIDUI	**5.031.000**	**62.697**	**62.697**
COGNOMI	**215.623**	**26.408**	**26.562**
FREQUENZA 1	**72.450 (33,6%)**	**16.953 (64,2%)**	**16.304 (61,4%)**
FREQUENZA 2	**24.000 (11,1 %)**	**4.122 (15,6%)**	**4.613 (17,3%)**
FREQUENZA 3	**14.450 (6,7%)**	**1.743 (6,6%)**	**1.992 (7,5%)**
ISONIMIA	**0,00017**	**0,00017**	**0,00016**



[82] P. Rossi, *On sampling and parametrization of discrete frequency distributions* (in preparazione)

21. Metodologie della fisica sperimentale

Una trattazione delle applicazioni interdisciplinari della fisica nel contesto umanistico sarebbe particolarmente incompleta se non dedicasse almeno un sommario *excursus* ai numerosi utilizzi di tecniche, metodi e strumenti della fisica sperimentale che vengono fatti in diversi campi delle scienze umane.

Esiste in particolare un'intera disciplina di frontiera, che va sotto il nome di archeometria, e che si occupa specificamente di applicare la tecniche fisiche, chimiche e di scienza dei materiali al campo dell'archeologia.
Le principali aree dell'archeometria sono:

- lo sviluppo e l'applicazione di metodi di datazione assoluta e relativa di reperti e manufatti
- lo studio degli artefatti dal punto di vista dell'individuazione della provenienza, delle tecnologie di produzione e delle tecniche e destinazioni d'uso
- lo studio dei paleoambienti dal punto di vista climatico, biologico ed ecologico
- lo sviluppo di metodi matematici per la trattazione quantitativa dei dati
- lo sviluppo di tecniche non invasive per lo studio dei siti
- lo studio delle tecniche di conservazione e restauro
-

Applicazioni che più specificamente riguardano la fisica sono in particolare quelle relative alla datazione. Ricordiamo le tecniche basate sui fenomeni di decadimento radioattivo (Carbonio 14) per la datazione di materiale organico, le tecniche basta sulla termoluminescenza per materiali inorganici (ceramiche), la luminescenza otticamente stimolata (OSL) per la datazione di strati sepolti, la risonanza di spin elettronico (che permette ad esempio di datare i denti),la datazione a potassio-argon (adatta per gli ominidi fossili).
Le principali tecniche fisiche di analisi degli artefatti sono:

- la fluorescenza a raggi X (XRF)
- la spettroscopia di massa a plasma accoppiato induttivamente (ICP-MS)
- l'analisi mediante attivazione di neutroni (NAA)
- la microscopia elettronica (SEM)
- la spettroscopia del plasma indotto da laser (LIBS)
-

Un aspetto di grande importanza nelle analisi di tipo puramente fisico, rispetto alle analisi chimiche più tradizionali, è il loro carattere non invasivo o micro-invasivo, che permette di non danneggiare i reperti, spesso rari e "preziosi". In questo senso è particolarmente importante lo sviluppo di tecniche d'indagine applicabili anche alla produzione artistica e ai testi scritti (papiri, pergamene, carta, etc), avendo come obiettivi anche in questo caso la comprensione dei tempi, delle tecniche e delle modalità di produzione, la conservazione e ove possibile il restauro del bene culturale.[83]

[83] M. Martini, A. Castellano, E. Sibilia, *Elementi di archeometria. Metodi fisici per i beni culturali*, Egea, Milano 2007

INDICE