manuale di tecniche di indagine

7 - le rappresentazioni grafiche di dati statistici

istat
istituto nazionale di statistica

note e relazioni
edizione 1989, n. 1

Autore del fascicolo:
Pierpaolo Napolitano

Editing di
Mario Nanni e Claudio Antonio Pajer

Supplemento all'Annuario Statistico Italiano

ISSN: 0535-9856

abete grafica s.p.a. - Roma - Contratto n. 104 del 7-12-1988 - copie 3.000

INDICE

APPENDICE

PRESENTAZIONE

Il *Manuale di tecniche di indagine,* la cui preparazione è stata curata dal Reparto Studi dell'Istituto, si configura come guida per la razionalizzazione delle operazioni di rilevazione ed è stato pure concepito quale strumento didattico da utilizzare ai fini della formazione dei funzionari dell'Istat. Poiché nell'effettuazione di indagini statistiche sono impegnati molti altri organismi pubblici e privati, si ritiene che esso possa costituire uno strumento utile anche per l'attività di questi organismi, in particolare di quelli che hanno un qualche ruolo nel sistema informativo socio-economico del Paese.

Il *Manuale* prende in esame i vari segmenti del *ciclo produttivo* nei quali si sviluppa normalmente ogni indagine statistica cogliendo aspetti che vanno dalla costruzione del disegno campionario al controllo della qualità dei dati, dall'analisi delle caratteristiche delle varie tecniche di indagine alla definizione di criteri standardizzati per la presentazione dei risultati. Pensato inizialmente per le indagini condotte con il metodo del campione, in particolare per quelle sulle famiglie, nella sua definitiva articolazione esso detta norme valide per fasi di lavoro riscontrabili nelle rilevazioni totali ed allarga pertanto il suo campo di applicazione che finisce per comprendere la generalità delle indagini.

La sua impostazione riflette il desiderio di colmare il divario fra il *libro di testo* ed il *manuale operativo.* Se da un lato infatti non si rinuncia al rigore della formalizzazione e si introducono spunti di innovazione sul piano metodologico, dall'altro si tengono ben presenti le esigenze del lavoro sul campo e risulta quindi ampio lo spazio riservato alle esemplificazioni.

Il *Manuale* consta dei seguenti fascicoli:

1. Pianificazione della produzione di dati
2. Il questionario: progettazione, redazione, verifica
3. Tecniche di somministrazione del questionario
4. Tecniche di campionamento: teoria e pratica
5. Tecniche di stima della varianza campionaria
6. Il sistema di controllo della qualità dei dati
7. Le rappresentazioni grafiche di dati statistici

In ogni caso va precisato che il *Manuale* non è da considerarsi completato in quanto è previsto che ai fascicoli programmati se ne aggiungano altri mano a mano che l'attività di ricerca avrà portato a termine l'esplorazione di aspetti per ora solo individuati.

PREMESSA

Il presente volume è il frutto del coordinamento di varie attività sviluppate presso l'Istat nel periodo 1985-90. Nel settembre 1985 fu istituito un gruppo di lavoro, formato da funzionari dell'Istat, con lo scopo di preparare gradualmente la standardizzazione della produzione di grafici statistici in Istat. Il gruppo si è avvalso della consulenza del Prof. R. Bachi dell'Università ebraica di Gerusalemme, in particolare per programmare metodi cartogrammatici e geostatistici adatti all'analisi dei dati prodotti dall'Istituto.

In questo periodo di tempo, per permettere l'informatizzazione del processo grafico, l'Istat ha incrementato la sua disponibilità di hardware e software necessari all'elaborazione grafica. Particolare sviluppo è stato dato alle tecniche di rappresentazioni grafiche e di analisi delle serie territoriali.

Tale attività si è ulteriormente concretata nella preparazione di questo fascicolo 7 del *Manuale di tecniche d'indagine*, relativo alla rappresentazione grafica delle serie statistiche.

Punti di riferimento fondamentali nella preparazione dei grafici statistici sono i seguenti:

a) il lettore del grafico ha il diritto di attendersi che esso dia in modo chiaro e corretto la rapida percezione dell'informazione contenuta nei dati;

b) la scelta del tipo di grafico e della scala utilizzata deve essere fatta tenendo conto del tipo di dati rappresentati, del tipo di lettore cui i grafici sono rivolti e, quindi, dello scopo della rappresentazione, in base a criteri scientifici;

c) in linea di massima, si deve tendere ad assicurare la consistenza fra le soluzioni grafiche adottate di volta in volta: i dati dello stesso tipo, destinati allo stesso tipo di lettore (pubblico generico, alunni delle scuole, ricercatori, pubblici amministratori, etc.) vanno rappresentati con lo stesso sistema grafico nelle pubblicazioni dell'Istat, al fine di consentire rapidi ed efficaci confronti fra i grafici.

Questo volume ha lo scopo di dare una risposta razionale a tali esigenze.

Il carattere interdisciplinare della materia, che richiede competenze statistiche, di psicologia della percezione, in campo grafico ed informatico, può spiegare la difficoltà dello sviluppo di una metodologia scientificamente basata, completamente soddisfacente e capace di dare le risposte ad ogni quesito. A tal fine lo sviluppo della sperimentazione empirica sulla percezione e risoluzione dei diversi simboli grafici, anche in funzione dello scopo della rappresentazione, può costituire la base su cui definire un sistema grafico razionale.

CAPITOLO 1 - CENNI STORICI SULLE RAPPRESENTAZIONI GRAFICHE IN CAMPO STATISTICO

1. UN BREVE PROFILO STORICO DELL'USO DEI METODI GRAFICI IN STATISTICA [1]

L'introduzione delle rappresentazioni grafiche nel campo della statistica generalmente la si attribuisce a William Playfair (1759-1823). Egli, in effetti, produsse con il suo *Commercial and Political Atlas* del 1786 il primo brillante esempio di come una sensibilità di tipo grafico si potesse integrare con l'esigenza di rendere di immediata comunicazione un insieme di dati, piuttosto complessi da decifrare.

Come ci ricorda il Funkhouser (1937), la prima edizione dell'Atlas consta di 44 diagrammi, che rappresentano, tutti ad eccezione di uno, delle serie storiche. Nell'Atlas è compreso anche un diagramma a barre affiancate, rappresentante l'importazione e l'esportazione della Scozia con vari paesi nel 1781. Insieme con una nuova edizione dell'Atlas, Playfair pubblicò nel 1801 uno *Statistical Breviary*, dove presentò nuovi tipi di grafici fra cui i noti e fortunati diagrammi a torta.

Con l'invenzione della geometria analitica da parte di Cartesio nel 1637, che si inserisce a sua volta in un periodo di ampio sviluppo dell'investigazione scientifica, l'uso delle rappresentazioni grafiche aveva trovato ampia diffusione, anche al fine di misurazione; per approfondire il tema, sicuramente interessante, si può far riferimento ai capitoli iniziali del testo di Boldrini (1968).

L'opera di Playfair trova perciò importanti precedenti e nelle scienze naturali e nei campi vicini alla ricerca di tipo statistico o a quelle discipline che sarebbero confluite nella statistica, come cercheremo brevemente di descrivere.

Dal lavoro fondamentale di J. Graunt *Osservazioni naturali e politiche fatte sui bollettini di mortalità* del 1662, che può considerarsi l'inizio della disciplina denominata da W. Petty *Aritmetica Politica*, lo scienziato C. Huygens, come ricorda E. Lombardo (1986), trasse lo spunto ed i dati per rappresentare graficamente le curve di sopravvivenza e la vita media al variare dell'età.

A quasi cento anni dal lavoro di Graunt, nel 1761, troviamo il contributo di J. L. D'Alembert che, in uno studio sul vaiolo, disegnò alcune curve di mortalità derivate da ipotesi teoriche. Sempre in Francia, J.B.J. Fourier, chiamato a predisporre le pubblicazioni relative al censimento della popolazione di Parigi del 1817,

(1) Per la stesura del seguente paragrafo si è fatto ampio riferimento al lavoro di Funkhouser (1937).

affiancò le tabelle con alcuni grafici; da essi egli poteva ricavare geometricamente la durata media e probabile di vita e l'età media della popolazione; nel lavoro egli disegnò le curve delle frequenze cumulate della popolazione distinta per età (Caselli, Lombardo 1988).

Val la pena di ricordare che il noto naturalista e geografo A. von Humboldt mostrò notevole apprezzamento per il lavoro di Playfair; egli stesso affiancò il saggio politico sulla «Nouvelle Espagne» con un Atlante, pubblicato nel 1811, dove fece ampio uso di diagrammi a barre, curve, quadrati sovrapposti e diagrammi a barre suddivise; questi ultimi due tipi di grafici possono considerarsi sue invenzioni.

Nel corso del 19° secolo l'importante lavoro del belga A. Quetelet ebbe il merito di ricondurre ad unità i filoni sparsi della scienza statistica. Egli mostrò di apprezzare il ricorso ai metodi grafici che avevano per lui soprattutto lo scopo di aiutare il ragionamento e l'analisi.

Suo noto contemporaneo fu A.M. Guerry, che ricordiamo per aver pubblicato nel 1833 l'*Essai sur la statistique morale de la France,* con numerose rappresentazioni grafiche. Fra queste anche alcuni istogrammi, che rappresentano le frequenze di alcuni crimini per categorie ordinate di una variabile continua come l'età. Giova ricordare che solo alla fine del secolo K. Pearson avrebbe coniato il termine tecnico di istogramma (Beniger e Robyn 1978).

Quando in Inghilterra, con l'influenza delle idee di Quetelet, sorse l'interesse per le rappresentazioni grafiche, l'aspetto che venne privilegiato fu quello rivolto all'analisi piuttosto che alla presentazione del dato. Vanno citati i nomi di W. Farr e F. Galton in campo statistico e quelli di A. Marshall e W.S. Jevons in campo economico. Nei suoi primi studi sulla correlazione, Galton ne determinò il coefficiente per via grafica; egli inoltre fu il primo in Europa, a detta del suo allievo Pearson, a realizzare carte di tipo meteorologico.

Una citazione va fatta anche del lavoro di Florence Nightingale, famosa per il suo impegno profuso nel miglioramento delle condizioni sanitarie dell'esercito britannico; numerose rappresentazioni grafiche, con uso anche del colore (Cox 1978), appaiono nel suo voluminoso rapporto del 1857.

Per quanto riguarda i cartogrammi la tradizione vuole (Funkhouser 1937) che la prima carta prodotta da uno statistico sia stata quella redatta da A.W. Crome, professore di politica economica e statistica all'Università di Giessen; la sua *Producten — Karte von Europa* fu pubblicata a Dessau nel 1782.

Una produzione significativa e brillante di tali carte si sviluppò, comunque, in Francia dove vennero inizialmente chiamate *«carte figurative»* e quindi *«cartogramme»*. Si usava distinguerle in quattro tipi fondamentali: 1) *cartogramme à teintes ou à hachures dégradées,* la cui invenzione si fa risalire a Dupin; 2) *cartogramme à bandes,* inventato da Minard, nel quale due località vengono unite da una striscia il cui spessore è proporzionale al flusso che le connette; 3) il terzo tipo prevede la sovrapposizione di diagrammi semplici nel centro delle zone in cui è stato diviso il territorio; 4) il quarto tipo consiste nelle carte con curve isometriche.

Nel 1843 L. Lalanne applicò il metodo delle curve isometriche a dati meteorologici classificati secondo il mese e l'ora; l'idea venne ripresa da Galton per lo studio grafico della correlazione.

L'applicazione del metodo a dati statistici distribuiti sul territorio si deve a L.L. Vauthier con la carta della densità di popolazione di Parigi, realizzata nel 1874.

Nel 1884, sempre in Francia, C.T. Minard produsse centinaia di cartogrammi del tipo a bande che rappresentavano flussi di passeggeri fra le diverse stazioni ferroviarie.

Nel 1877 apparve sul testo di statistica di G. von Mayr una parte completamente dedicata alle rappresentazioni grafiche. Egli presentò i diversi tipi di grafici classificandoli in diagrammi e cartogrammi e quindi introdusse una ulteriore specificazione in funzione della dimensione delle figure geometriche utilizzate nel grafico, dal punto fino al volume. Per quanto riguarda l'Italia, si possono ricordare i contributi di L. Perozzo apparsi sugli Annali di Statistica del 1880-81, dove egli proponeva complessi grafici statistici tridimensionali, definiti stereogrammi; va anche ricordata l'opera di A. Gabaglio con il suo testo del 1888, che contiene un'ampia parte dedicata alle rappresentazioni grafiche, fra cui egli presenta anche alcuni diagrammi polari.

Sono soprattutto da ricordare uno scritto del Benini del 1905 sul Giornale degli Economisti, teso a divulgare l'uso dei diagrammi logaritmici e un articolo del Gini apparso sulla stessa rivista nel 1914, testimonianza del suo vivo interesse per l'argomento. Quest'ultimo ritornerà più volte sulla questione apportando interessanti osservazioni sia di metodo che di sostanza.

2. CENNI STORICI SUL PROBLEMA DELLA STANDARDIZZAZIONE IN CAMPO GRAFICO [2]

Il problema della standardizzazione ed il suo significato erano ben presenti fin dall'inizio della storia delle rappresentazioni grafiche in campo statistico. Nei congressi internazionali di statistica, che si tennero in Europa dal 1853 al 1876, largamente dovuti alla iniziativa di Quetelet, il tema fu più volte trattato.

(2) Il materiale di questo paragrafo fa diretto riferimento ai lavori di Funkhouser (1937) e Schmidt (1978).

Al terzo congresso tenuto a Vienna nel 1857 fu presentato un rapporto sulla cartografia e i metodi grafici in generale, che tenteva di classificare i fenomeni in base a quello che si riteneva fossero le quattro forme base del conoscere umano: il che cosa, il quanto, il dove ed il quando. Si arrivò financo a definire un sistema di associazione fra tipo di dato e grafico. Vi furono accese discussioni, cui non seguì nessuna decisione unanime ed effettuale.

Nel congresso di The Hague del 1869 vennero adottate due risoluzioni; la prima, presentata da Mayr, raccomandava l'uso delle rappresentazioni grafiche nelle pubblicazioni statistiche ufficiali; la seconda chiedeva la presentazione al congresso successivo di una memoria sui metodi grafici in uso prevalente e sui possibili criteri per renderli uniformi e compatibili. Si vuol ricordare che dal 1872 l'ufficio del Censimento degli U.S.A. iniziò la pubblicazione di un Atlante concepito con lo scopo di divulgare l'informazione statistica, sollecitato in questo dagli stessi organi confederali.

Nel successivo congresso di S. Pietroburgo del 1872 vennero presentate due memorie distinte: una sui diagrammi, l'altra sui cartogrammi. La prima costituisce un'accurata rassegna dei metodi grafici in uso. Nella seconda A. Ficker sosteneva che per fare una corretta scelta delle classi di valori per la rappresentazione cartogrammatica, era opportuno utilizzare come criterio quello dei raggruppamenti naturali; egli intendeva che la scelta delle classi dovesse di volta in volta basarsi sul tipo di dati da rappresentare, allo scopo che quelle fossero effettivamente omogenee al loro interno. La sua tesi si contrapponeva a quella di von Mayr, che sosteneva l'opportunità di determinare i valori limiti delle classi partendo dalla escursione fra minimo e massimo della variabile da rappresentare e dividendo questa per il numero delle classi dèsiderate. In sede di conclusioni si dichiarò, ancora una volta, che non erano maturi i tempi per concepire una soddisfacente standardizzazione.

Nel 1885, anno in cui si celebrò il giubileo della London Statistical Society, venne fondato l'International Statistical Institute (ISI). Nelle periodiche sessioni tenute da tale istituto furono fatti ulteriori tentativi per la standardizzazione della produzione dei grafici, senza tuttavia alcun risultato apprezzabile. Si vuole anche ricordare che in quella occasione A. Marshall sollevò il delicato problema della scelta della scala temporale per la rappresentazione delle serie storiche.

A Parigi nel 1909 fu creato un comitato composto, fra gli altri, da R. Benini, J. Bertillon, Von Bortkiewicz, W. Lexis e F.Y. Edgeworth. Questo comitato presentò alla sessione di The Hague del 1911 due raccomandazioni; la prima concerneva la rappresentazione delle serie storiche e stabiliva una relazione fra il valore medio del fenomeno da rappresentare nel periodo 1901-10 e l'intervallo

di 30 anni della scala dei tempi; la seconda verteva sulle rappresentazioni grafiche delle curve di frequenza. Esse, questa volta, in quanto raccomandazioni, furono adottate.

Nel 1914 negli Stati Uniti, la American Society of Mechanical Engineers, lanciò un invito alle altre società scientifiche americane affinché si riunissero in un comitato congiunto avente lo scopo di definire gli standard per le rappresentazioni grafiche. Il Joint Commitee risultò composto da 17 associazioni ed agenzie, fra cui l'American Statistical Association: questa volta, forse anche poiché la presenza di più nazioni non costituiva elemento di ostacolo ai lavori, si ottenne il consenso intorno ad alcune parti essenziali.

I risultati, pubblicati nel 1915, uscirono sotto forma di 17 semplici regole, la maggior parte delle quali riguardanti il modo corretto di rappresentare le serie storiche.

Di queste ne ricordiamo alcune, così come vengono riprese da Luzzatto-Fegiz (1934) e Salvemini, Girone (1981):

a) ogni grafico deve contenere in sè tutte le indicazioni necessarie per la sua esatta interpretazione, indipendentemente dal testo; quindi titolo chiaro dell'oggetto delle rappresentazioni, l'epoca a cui si riferiscono i dati, l'ambito territoriale, la fonte e le scale di misura adottate;

b) il grafico deve essere riprodotto ed usato in modo autonomo dal testo originario; quando è possibile occorre accompagnarlo con i dati che esso rappresenta;

c) i numeri e le parole scritte sui grafici devono essere leggibili senza girare il foglio;

d) se si vogliono rappresentare più diagrammi nello stesso grafico conviene scegliere: 1) un segno diverso per ogni diagramma; 2) indicare accanto ad ogni curva, il fenomeno a cui essa si riferisce; 3) segnare ai margini del grafico le scale adottate (una o più secondo il caso); 4) si deve evitare che la rappresentazione risulti aggrovigliata o confusa; si consiglia di ricorrere eventualmente a più grafici paralleli, con scale spostate;

e) nei grafici cartesiani è opportuno disegnare un reticolato a linee spezzate in modo da agevolare l'occhio nella lettura;

f) scegliere giudiziosamente il metodo di rappresentazione, in modo che sia il più adatto al tipo di tabella data; quando si possono applicare correttamente più metodi, dare la preferenza a quello più semplice;

g) eseguire la revisione ed il controllo per evitare errori materiali di riproduzione.

Nell'ambito delle sessioni dell'ISI vi furono altre iniziative di cui

ricordiamo quelle che videro promotore nel 1933 e nel 1949 lo statistico italiano C. Gini.

Nella riunione del 1933 K. Winkler presentò la proposta, che venne approvata, di istituire una commissione per la standardizzazione.

Nell'ambito della sessione dell'International Statistical Institute del 1975, Bachi (1975) presentò la sua proposta di standardizzazione congiunta degli assi x e y nei diagrammi cartesiani.

Vogliamo brevemente ricordare il contributo di Cox (1978), che in occasione della Conferenza di Sheffield sui metodi grafici in statistica presentò una memoria, dove sottolineava, fra l'altro, la necessità di arrivare a definire una teoria dei metodi grafici e descriveva alcuni possibili regole che potrebbero presiedere ad una accurata realizzazione dei grafici.

Fra queste ricordiamo:

a) sugli assi occorre sempre indicare chiaramente il significato delle variabili e le unità di misura;

b) le amputazioni della scala devono essere indicate con interruzione degli assi;

c) rendere agevole il confronto dei grafici tra loro collegati affiancandoli ed utilizzando le stesse scale;

d) fissare le scale in maniera che le relazioni approssimativamente lineari forniscano un angolo di 45 gradi con l'asse delle ascisse;

e) la tecnica di rappresentazione non deve influenzare il lettore del grafico.

Nel 1976 l'American Statistical Association (ASA) creò una Commissione ad hoc per la grafica statistica, affinché valutasse la produzione grafica nelle riviste scientifiche e nella stampa d'ampia diffusione, indagasse sui rapporti fra grafica statistica ed informatica, valutasse i settori dove la conoscenza era adeguata e, laddove era necessario, fornisse delle proposte per lo sviluppo della ricerca, l'insegnamento e l'educazione più ampia in campo grafico. Di tale comitato facevano parte, fra gli altri, Bachi, Biderman, Schmid, Cleveland e Tufte.

Nel rapporto finale della Commissione ad hoc si sollecitava, fra l'altro, l'ASA ad avviare più stretti contatti con l'istituto americano degli standard (ANSI), per promuovere più ampi criteri di standardizzazione [3]. La Commissione ad hoc riteneva necessario estendere la standardizzazione dalle serie storiche anche alle serie territoriali.

La Commissione inoltre sollecitò la creazione di una sezione permanente dell'ASA di Statistical Graphics allo scopo di migliorare la qualità della produzione grafica, di promuovere la ricerca secondo i filoni già individuati, con interventi anche nel campo della Computer Graphics, di promuovere l'educazione alla corretta lettura dei grafici statistici.

(3) Esiste negli U.S.A. un comitato denominato *American National Standard Committee on Preferred Practice for the Preparation of Graphs, Charts and Other Technical Illustrations;* esso è designato con la sigla Y15 all'interno dell'ANSI, l'istituto americano per gli standard. Esso è composto da 16 membri effettivi, di cui 12 in rappresentanza di associazioni professionali e di commercio, 3 esperti, 1 membro del settore industriale. La presenza dell'ASA è stata garantita in modo praticamente continuo. Nel 1979 il comitato Y15 ha provveduto alla ristampa aggiornata degli standard in materie di serie storiche, usciti la prima volta nel 1915.

3. ALCUNE CONSIDERAZIONI SUL LAVORO GRAFICO ED UNA BREVE RASSEGNA DELLA PASSATA PRODUZIONE DELL'ISTAT

L'uso dell'informatica e degli elaboratori elettronici in alcune funzioni dei processi lavorativi ha importanti conseguenze sul modo di concepire e realizzare i prodotti; questa interviene direttamente sull'organizzazione dei processi di produzione anche attraverso la introduzione di elementi di standardizzazione. Il linguaggio e la logica dell'informatica richiedono rigore definitorio, chiarezza sugli obiettivi e completezza nella loro specificazione. L'uso degli elaboratori richiede competenze tecniche talvolta complesse nella gestione degli strumenti hardware e software. Solo da una riuscita interazione di questi diversi elementi il processo di rinnovamento può aspirare al successo.

L'informatica tende, per questa via, ad investire vari campi dell'attività umana e, specialmente, come già detto, i processi organizzativi e decisionali.

Una parte della letteratura sull'argomento sottolinea come l'avvio di processi di informatizzazione comporta il rischio di un adeguamento a soluzioni che non paiono sensibili ad esigenze di rigore metodologico che sono proprie dell'ambito culturale che conosce il problema nella sua complessità, (Schmid 1983); in generale infatti la logica dell'informatica non può esaurire la specificità e la ricchezza di una data tematica.

La valorizzazione delle competenze interne ed il riferimento alle esperienze accumulate nel tempo sono certamente, nel contesto delineato, operazioni proficue; esse hanno anche il senso di un confronto con soluzioni e proposte del passato più o meno recente e con la filosofia generale che le ha ispirate.

In effetti per rendere efficace e calare in modo operativo un processo di rinnovamento si può affermare che vi sono da realizzare inizialmente le due operazioni seguenti: a) una ricognizione dall'ampio respiro metodologico del problema da affrontare per

predisporre le soluzioni più adeguate; b) un confronto costruttivo e critico con precedenti esperienze nel settore.

Il campo specifico della rappresentazione grafica dei dati statistici riveste un ruolo importante per la divulgazione della informazione e della diffusione di cultura statistica. Se si riflette sulla quantità enorme di dati destinata a rimanere sostanzialmente inutilizzata, o il cui uso è relegato a pochi specialisti già sensibili a problemi determinati, può emergere la importanza che un uso diffuso della rappresentazione grafica può svolgere financo nella individuazione di problematiche. Riferendoci in particolare alle rappresentazioni di dati territoriali appare più evidente l'importanza dell'uso del grafico al fine di una analisi di tipo esplorativo delle distribuzioni territoriali.

D'altra parte l'aspirazione ad una corretta divulgazione della informazione statistica vede la rappresentazione grafica come strumento privilegiato di presentazione, allorquando questa abbia proprietà tali da preservare le caratteristiche del dato.

L'attenzione che giustamente si rivolge al miglioramento della qualità dei dati verrebbe paradossalmente a vanificarsi nel momento della presentazione di questi al pubblico, momento finale e punto di contatto, il più immediato e generale, con la realtà esterna, attraverso l'uso di metodi grafici che rendono l'informazione in modo distorto.

Gli effetti moltiplicativi, legati all'espandersi dei settori coinvolti dall'informazione statistica e dalla crescita culturale e di capacità critica necessariamente connessa ad esso, rischierebbero di ridursi notevolmente, se il grafico si limitasse a catturare l'attenzione invece di sollecitare la lettura e l'interpretazione critica.

Concludiamo questo paragrafo con una breve rassegna critica della passata produzione grafica dell'Istat, in particolare del *Compendio* e dell'*Annuario Statistico Italiano*. Ciò sembra essere utile ad una ricognizione ed un recupero critico della propria esperienza passata più o meno recente [4].

Un significativo impegno è stato rivolto in Istat alla rappresentazione della distribuzione della popolazione secondo classi di età ed il sesso attraverso l'uso della piramide delle età. Nella figura 1.1a viene riportato un esempio realizzato per il Compendio Statistico del 1929, che dà la distribuzione della popolazione ai censimenti del 1911 e 1921 per età, sesso e stato civile.

Nel Compendio del '38 venne riproposta da E. Gradara la rappresentazione del baricentro della stessa piramide come un indice

(4) Al fine di una valutazione critica più ampia della passata produzione grafica dell'Istat si può far riferimento alle considerazioni e le tabelle presentate nell'appendice I dell'articolo di Napolitano (1987).

sintetico di due importanti caratteristiche della popolazione; è facile vedere che il valore dell'ordinata del baricentro misura la età media della intera popolazione mentre il valore della sua ascissa dà una misura della sua composizione percentuale rispetto al sesso.

Nella figura 1.1b è rappresentata la piramide dell'età relativa al censimento speciale del 1936 con il rispettivo baricentro.

Nella figura 1.1c sono riportati i punti rappresentanti i baricentri delle piramidi delle età relative ai censimenti dal 1881 al 1936. La scala molto fine delle ascisse permette di risolvere le differenze di composizione percentuale rispetto al sesso ai vari censimenti.

Nella figura 1.1d viene presentato il grafico riportante i nati vivi, i morti e le eccedenze dei nati sui morti, su 1000 abitanti ed i matrimoni per 1.000 abitanti, dal 1872 al 1930. Nel grafico viene rappresentata la curva delle eccedenze dei nati vivi sui morti; ciò aiuta la valutazione dell'andamento dell'incremento naturale.

Nella figura 1.1e è riportato in un diagramma a barre suddivise il numero degli esaminati ed approvati, per sesso e per alcuni tipi di scuola; il grafico fornisce una gran quantità di informazione ma non è certamente di lettura immediata. Una serie di grafici a barre affiancate potrebbe rendere più agevole l'interpretazione dei dati.

Nella figura 1.1f vengono riportate alcune serie storiche relative all'andamento medio annuale di alcune produzioni agrarie. Sui grafici sono riportate due scale: la prima dà i rendimenti medi annuali per ettaro in quintali; la seconda dà le variazioni percentuali sulla media del quinquennio 1909-1913. Per tale grafico c'è da osservare che la quantità di inchiostro utilizzata per rappresentare la griglia di riferimento distoglie l'attenzione dall'informazione principale costituita dal dato; maggiormente marcata avrebbe dovuto essere la linea dello zero e l'uso del colore per i livelli positivi e negativi avrebbe facilitato la lettura. L'informazione del quinquennio di riferimento sarebbe stata opportunamente posta nel titolo.

Terminiamo questa breve rassegna con alcuni grafici dell'Annuario di un periodo successivo. Nella figura 1.1g sono riportate delle serie storiche relative ad alcuni indici di prezzi all'ingrosso. Il trend dei vari indici presenta una crescita lieve; ciò consente rappresentazioni grafiche che utilizzano una sola scala di tipo lineare e dà la possibilità di confronti semplici e coerenti. Un'osservazione critica va rivolta al fatto che il valore di partenza dell'asse delle ordinate è assunto arbitrariamente e non è stato posto uguale a zero, segnalando al lettore la interruzione della scala con un simbolo grafico opportuno.

Nella figura 1.1h è riportata una sequenza di grafici a barre relativi alla percentuale dei morti per classe di età e per gruppi di cause di morte. Il grafico dà la possibilità di confrontare facilmente

l'andamento della mortalità per classe di età rispetto alle varie cause.

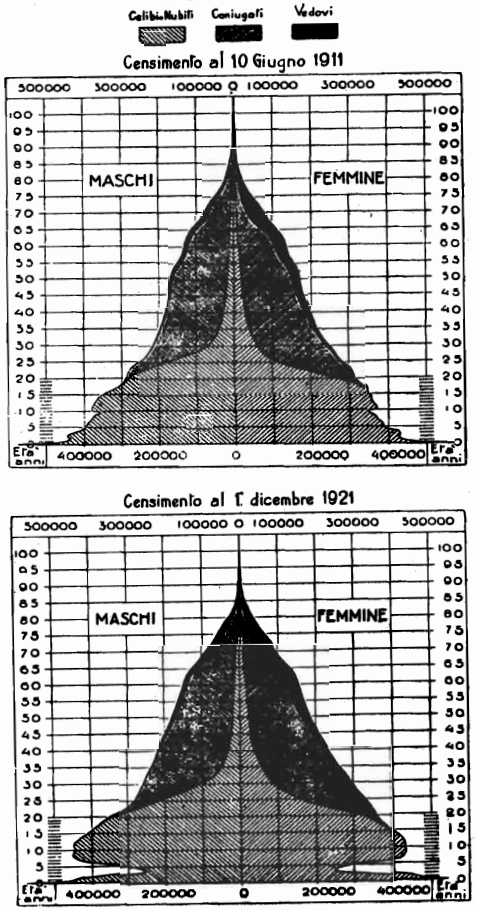
Nel diagramma relativo alla causa di morte «altri stati morbosi», per la classe di età fino a 5 anni si ha un valore particolarmente elevato e quindi una presentazione completa della barra relativa avrebbe appiattito il resto dei dati. La colonna è stata allora troncata nella sua parte alta, riportando accanto ad essa il valore corrispondente, a sottolineare che il dato non è comunque rappresentabile con la scala prescelta.

Infine nella figura 1.1i vengono dati esempi di serie territoriali che rappresentano separatamente la produzione di frumento, riso e granturco relativamente al 1961. Viene utilizzata un rappresentazione a punti, che preserva il carattere quantitativo dell'informazione ed il cartogramma è accompagnato dalla serie storica dell'andamento della produzione dal 1956 al 1961.

Figura 1.1 — Selezione di alcuni grafici del Compendio e dell'Annuario Statistico Italiano di vari anni rappresentativi della passata produzione dell'Istat

(a)

VI. - Distribuzione della popolazione per età, sesso e stato civile.

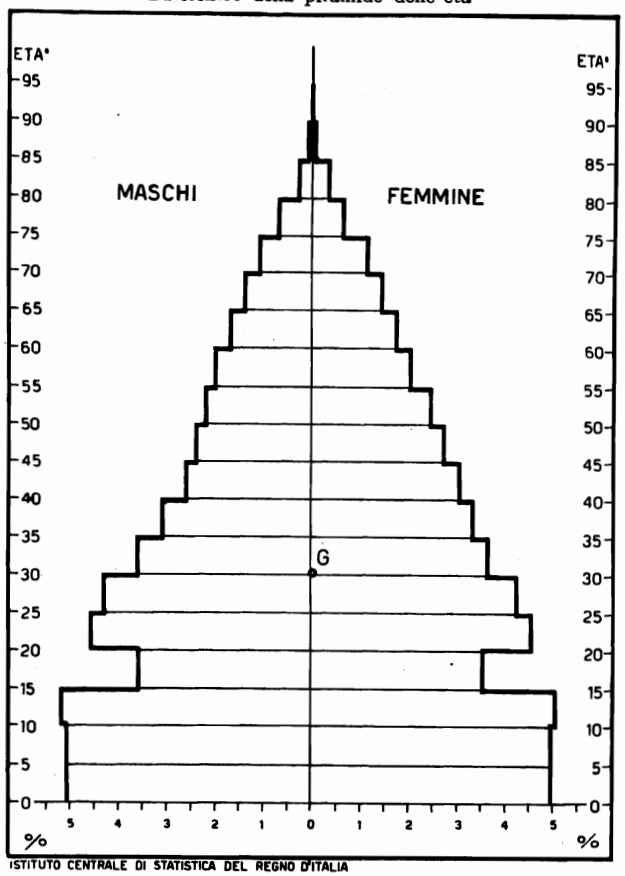


Fonte: Compendio Statistico Italiano (1929)

Figura 1.1 — Selezione di alcuni grafici del Compendio e dell'Annuario Statistico Italiano di vari anni rappresentativi della passata produzione dell'Istat

(b)

37. — Popolazione speciale del Regno al Censimento 1936
Baricentro della piramide delle età

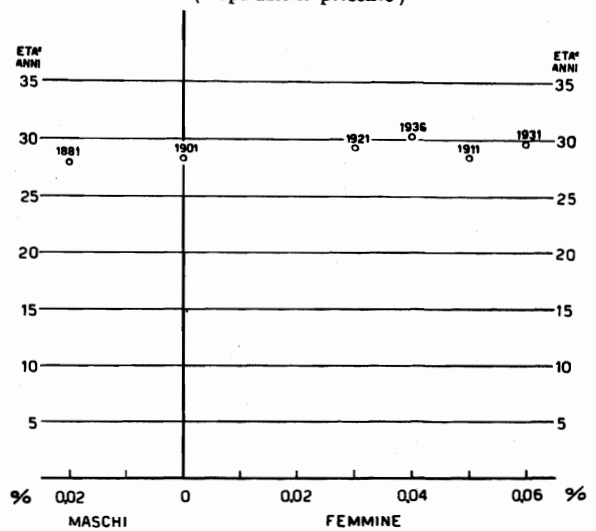


Fonte: Compendio Statistico Italiano (1938)

Figura 1.1 — Selezione di alcuni grafici del Compendio e dell'Annuario Statistico Italiano di vari anni rappresentativi della passata produzione dell'Istat

(c)

41. — Posizione dei baricentri delle piramidi delle età del Regno ai censimenti indicati
(Popolazione presente)



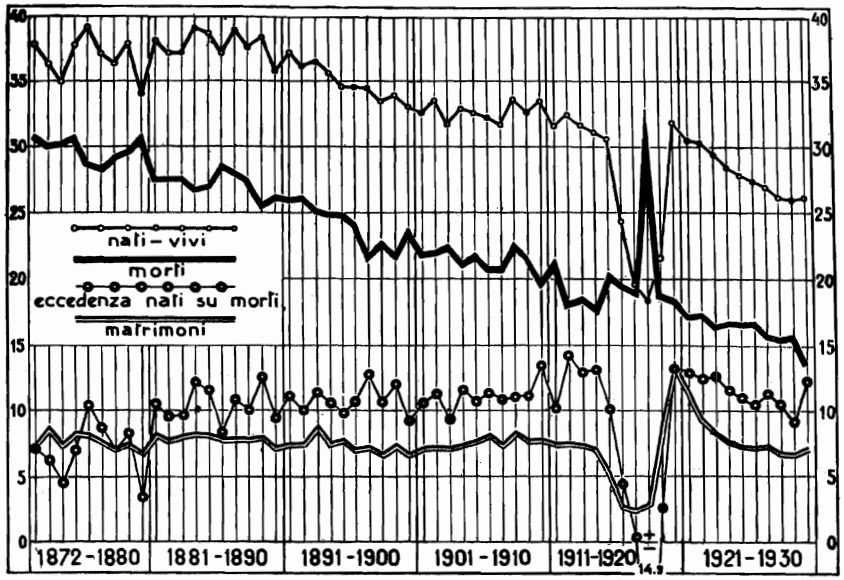
7. – Baricentri delle piramidi delle età per le popolazioni dei Compartimenti ai Censimenti del 1901 *e* 1936.

Fonte: Compendio Statistico Italiano (1938)

Figura 1.1 — Selezione di alcuni grafici del Compendio e dell'Annuario Statistico Italiano di vari anni rappresentativi della passata produzione dell'Istat

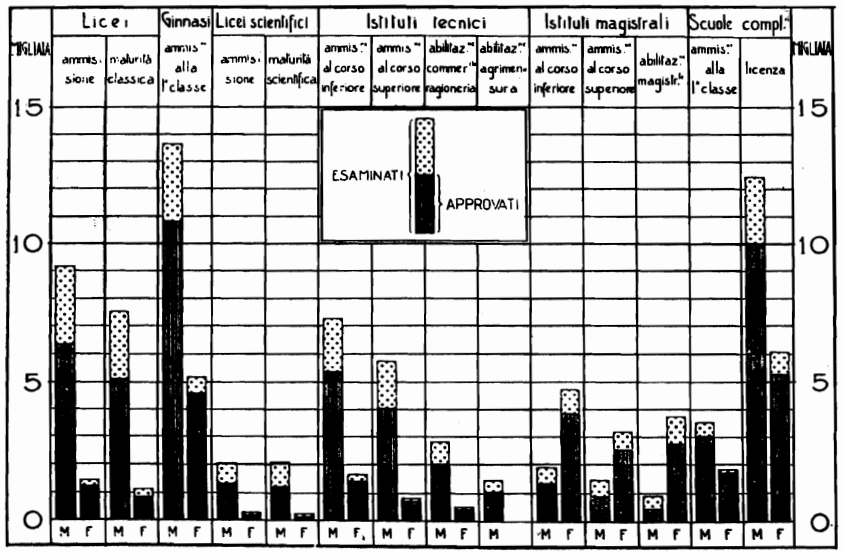
(d)

2. — Matrimoni, nati vivi, morti ed eccedenza dei nati sui morti per 1000 abitanti dal 1872 al 1930



(e)

7. — Istruzione media nell'anno 1928-29

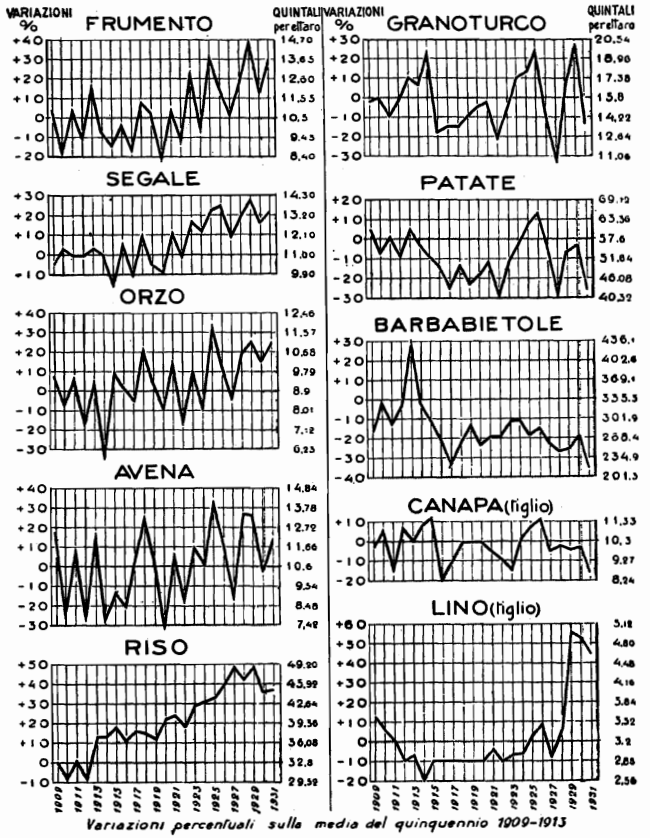


Fonte: Compendio Statistico Italiano (1931)

Figura 1.1 — Selezione di alcuni grafici del Compendio e dell'Annuario Statistico Italiano di vari anni rappresentativi della passata produzione dell'Istat

(f)

10. — Rendimenti medi annuali per ettaro delle principali produzioni agrarie

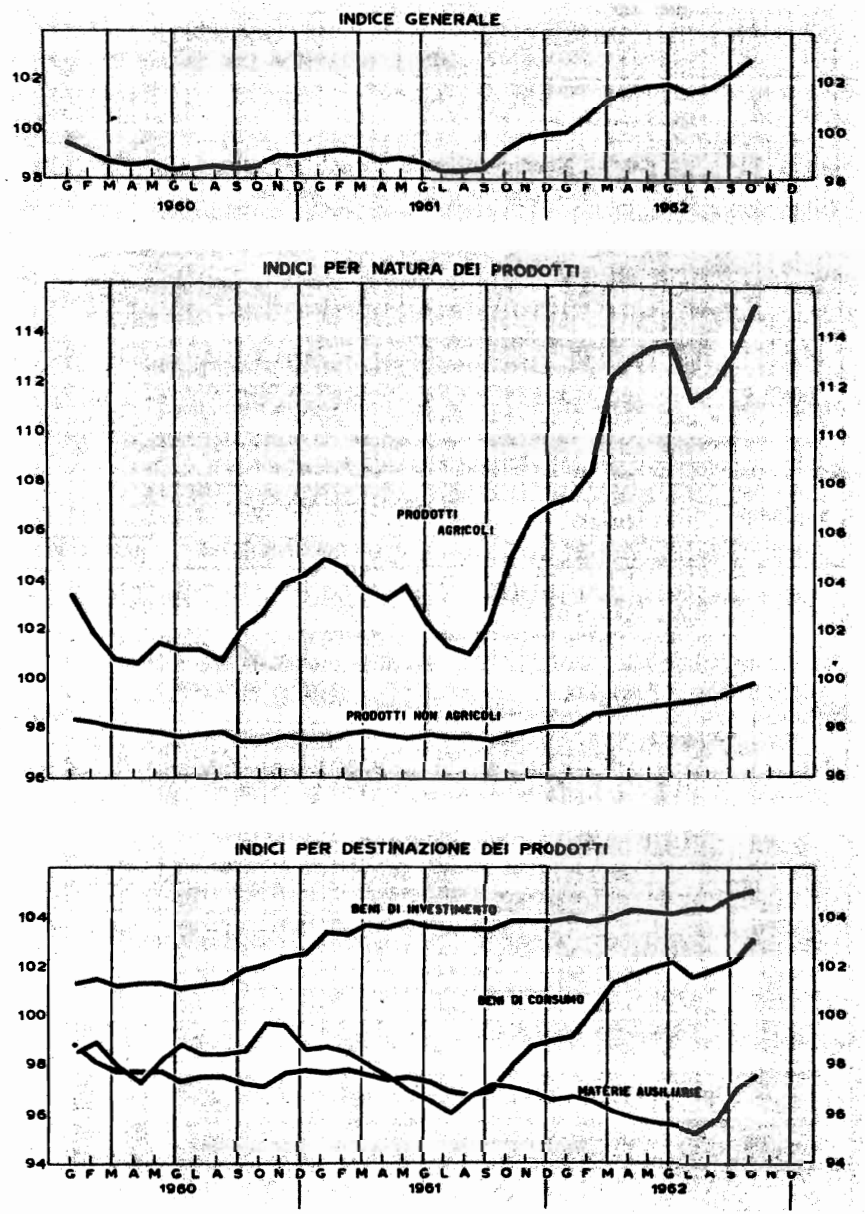


Fonte: Compendio Statistico Italiano (1931)

Figura 1.1 — Selezione di alcuni grafici del Compendio e dell'Annuario Statistico Italiano di vari anni rappresentativi della passata produzione dell'Istat

(g)

NUMERI INDICI DEI PREZZI ALL'INGROSSO
Base: 1953=100

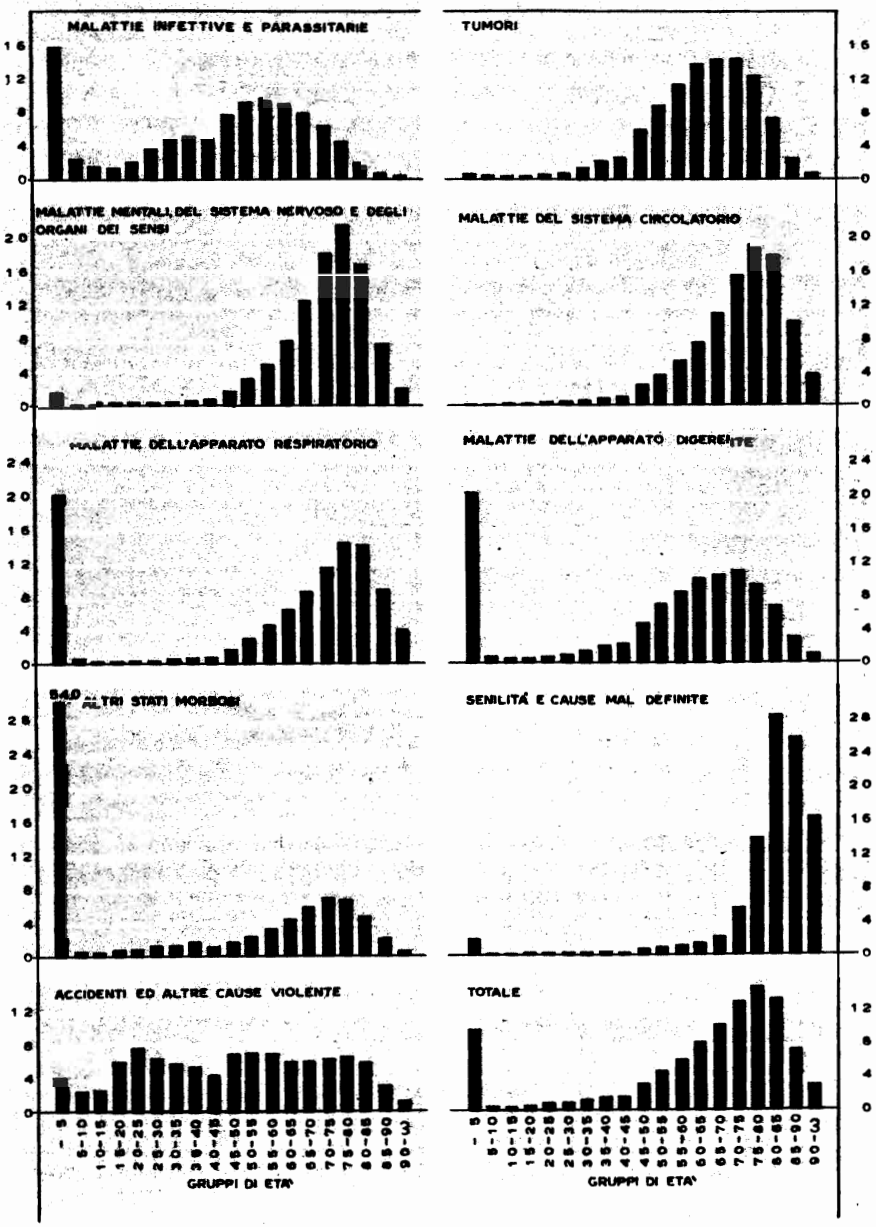


Fonte: Annuario Statistico Italiano (1962)

Figura 1.1 — Selezione di alcuni grafici del Compendio e dell'Annuario Statistico Italiano di vari anni rappresentativi della passata produzione dell'Istat

(h)

MORTI PER ETÀ SECONDO LA CAUSA
Percentuali per età dei morti per ciascun gruppo di cause
Anno 1960

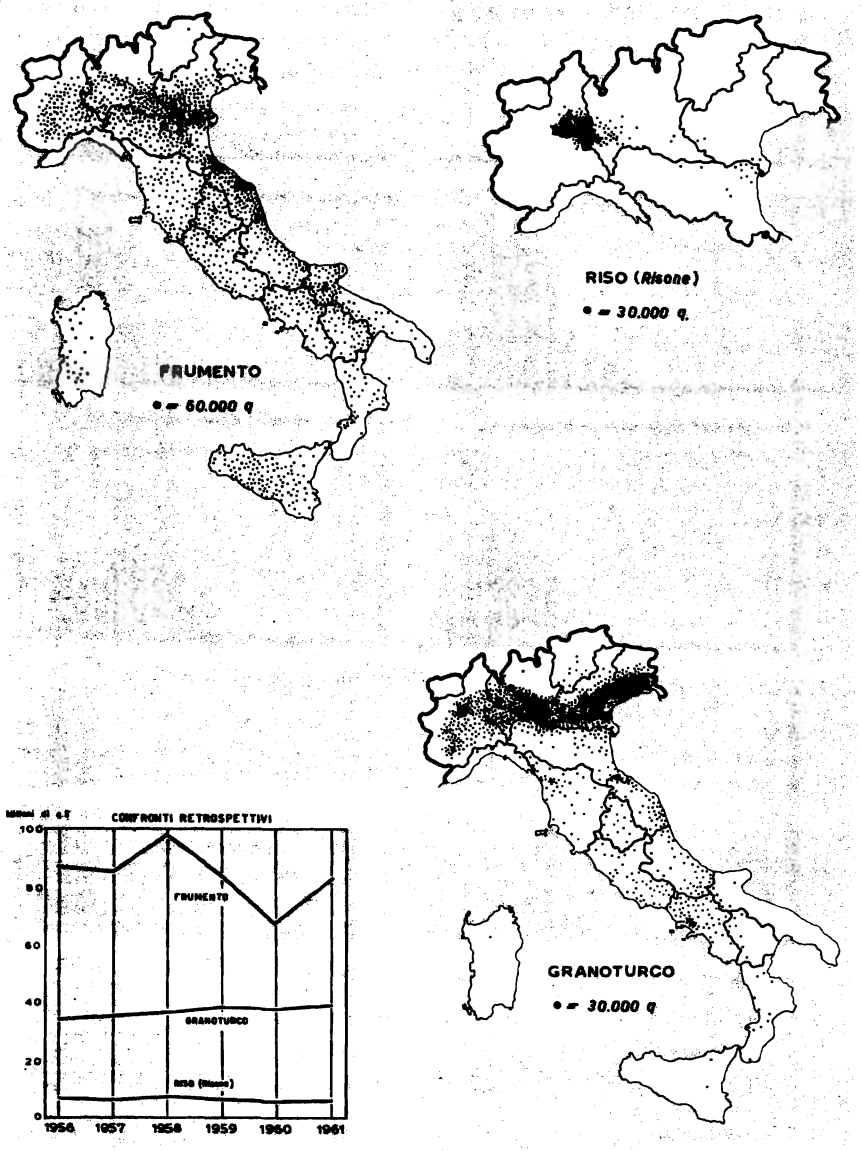


Fonte: Annuario Statistico Italiano (1962)

Figura 1.1 — Selezione di alcuni grafici del Compendio e dell'Annuario Statistico Italiano di vari anni rappresentativi della passata produzione dell'Istat

(i)

PRODUZIONE DEL FRUMENTO, DEL GRANOTURCO E DEL RISO
Anno 1961



ISTITUTO CENTRALE DI STATISTICA

Fonte: Annuario Statistico Italiano (1962)

CAPITOLO 2 - ALCUNI CRITERI PER UNA CORRETTA RAPPRESENTAZIONE GRAFICA DEI DATI STATISTICI

1. CONSIDERAZIONI GENERALI

Le rappresentazioni grafiche costituiscono spesso delle semplici illustrazioni che accompagnano un discorso. Il potere magico delle belle immagini colorate è davanti ai nostri occhi e quotidianamente sollecita la nostra fantasia. Sovente in televisione si mostra un grafico per rompere la monotonia di un discorso e dare un tocco brillante ad una noiosa citazione di cifre.

Lo scopo di questo fascicolo è riprendere alcuni possibili criteri per un sistema semplice e corretto di rappresentazioni grafiche, per dati statistici, nella convinzione che queste possano essere un effettivo supporto di informazione e svolgere un ruolo significativo nella comunicazione e nell'analisi dei dati statistici.

Si distingue generalmente fra due possibili usi dei grafici in statistica:

a) a scopo di analisi;

b) a scopo di presentazione.

La differenza degli obiettivi, che può diventare piuttosto sfumata, si riflette nelle diverse proprietà che vengono richieste ai grafici.

Per quelli destinati alla presentazione si chiede soprattutto che le caratteristiche e le relazioni presenti nell'informazione prescelta siano rese semplicemente ed accuratamente; la funzione della trasformazione dalla forma tabellare alla immagine grafica è di rendere immediatamente percepibili alcuni aspetti impliciti in quella, in modo da evidenziare il livello semantico prescelto.

I grafici destinati all'analisi dei dati vengono concepiti in modo che le ipotesi alternative, che possono aver generato i dati sperimentali, risultino visivamente ben differenziate. Proposte interessanti, come il box-plot ed alcune sue varianti, sono state formulate nell'ambito della disciplina *Analisi esplorativa dei dati* (Tukey 1977); alcune di queste richiedono la nostra attenzione, in quanto immediatamente convertibili ad un uso prettamente descrittivo.

Il nostro interesse si rivolge, infatti, essenzialmente, al problema di una corretta presentazione dei dati statistici. Per raggiungere questro scopo è opportuno, una volta selezionato il tema, procedere ad un'analisi dell'informazione da presentare, al fine di chiarire esattamente le relazioni che la compongono. Deter-

minare i caratteri e individuare con chiarezza le relazioni da rappresentare è indispensabile per scegliere correttamente il metodo grafico. Questa analisi riesce utile anche per determinare correttamente il titolo e la definizione degli elementi variabili del grafico.

Lo scopo dei prossimi paragrafi di questo capitolo è fornire delle indicazioni intorno alle questioni su accennate.

2. UNA CLASSIFICAZIONE DELLE SERIE SEMPLICI IN FUNZIONE DELLA LORO RAPPRESENTAZIONE GRAFICA — INDIVIDUAZIONE DELLE COMPONENTI DA RENDERE GRAFICAMENTE

Si distingue, inizialmente, fra dati elementari e dati aggregati. Nel caso di dati elementari, di tipo univariato, l'informazione si presenta come un insieme di modalità di un certo carattere, associato ad un insieme di osservazioni individuali, quali ad esempio l'insieme dei voti degli alunni di un corso di statistica oppure l'insieme che dà il colore dei capelli per ciascuno studente.

Se il numero degli individui non è grande, può essere utile rappresentare l'insieme delle modalità con un metodo grafico appropriato. Nel caso di dati quantitativi (come altezza, peso, voti) l'uso di una rappresentazione grafica come lo *stem-and-leaf,* oppure il *box-plot,* può fornire indicazioni piuttosto interessanti sulla forma della distribuzione dei dati.

Queste sono rappresentazioni statistiche proposte da Tukey nell'ambito della disciplina *Analisi Esplorativa dei dati.* Lo *stem-and-leaf,* che in italiano si può tradurre *diagramma a ramo-foglia,* si basa sull'idea di utilizzare nella rappresentazione le stesse cifre che compongono i valori delle misure. Le cifre più significative, ordinate verticalmente, vanno a costituire il ramo del grafico, la sua struttura principale; quelle meno significative le foglie del diagramma.

Il *box-plot,* o *diagramma a scatola,* si costituisce rappresentando su un asse orizzontale la posizione della mediana e dei due quartili, inferiore e superiore, della distribuzione e completando il disegno con la eventuale rappresentazione di valori anomali.

Per una trattazione più completa ed approfondita si può far riferimento al paragrafo 5 del Capitolo 3.

Il processo di aggregazione dei dati elementari genera le serie statistiche, che possono essere semplici o multiple, in funzione del numero dei caratteri coinvolti nell'aggregazione. Ci si limiterà, comunque, a considerare organicamente le serie statistiche semplici.

Al fine della scelta del metodo grafico più adatto per rappresentarle, ci pare sufficiente la seguente classificazione delle serie statistiche semplici:

a) distribuzioni di frequenze, ammontari o quantità derivate, associate a caratteri qualitativi non ordinabili, che costituiscono le serie qualitative sconnesse;

b) distribuzioni di tipo lineare (Bachi, 1968) che distinguiamo in:

1) distribuzioni di frequenze, ammontari o quantità derivate associate a caratteri qualitativi ordinabili linearmente;

2) distribuzioni di frequenze, ammontari o quantità derivate associate a caratteri quantitativi, distinguibili in continui e discreti;

3) serie storiche che possono essere distinte a seconda che si riferiscono a dati di stato o di flusso;

c) serie territoriali, che risultano composte da un elemento geografico e da un elemento descrittivo; di esse si tratterà estesamente nel Capitolo 4.

Trascuriamo di considerare le serie cicliche, dal momento che esse costituiscono una quantità esigua dei casi possibili e per il fatto che, tranne alcuni esempi eccezionali, possono essere adeguatamente rappresentate come serie di tipo lineare.

Per affrontare correttamente il problema della resa grafica di un insieme di dati statistici, elementari ed aggregati, è conveniente (Bertin, 1983) individuare preliminarmente le diverse componenti dell'informazione, le cui relazioni interessa presentare graficamente. Esplicitiamo questo punto con le seguenti esemplificazioni.

Se si vuole rendere graficamente come si distribuisce un insieme di individui rispetto ad un certo carattere, le due componenti da rappresentare sono, da una parte, il carattere con le sue modalità e, dall'altra, i singoli individui. Il grafico è la rappresentazione delle relazioni esistenti fra queste due componenti dell'informazione.

Se si vuole mantenere distinti i singoli individui si è obbligati ad attribuire ad ognuno di essi un simbolo differente; nel caso del grafico stem-and-leaf, destinato alla rappresentazione di osservazioni individuali semplici, queste vengono identificate con i valori ad esse attribuiti nella misura.

Se si vuole presentare la distribuzione di frequenze, ammontari o quantità derivate, associate ad un carattere di natura arbitraria, le componenti dell'informazione, la cui relazione interessa evidenziare, sono, da un parte, il carattere con le sue modalità e, dall'altra, le frequenze, o altre quantità numeriche, a queste associate.

Per quanto riguarda le serie storiche, le componenti da porre in relazione sono il tempo e le quantità associate, di stato o di flusso; esse sono entrambe variabili quantitative.

Per quanto riguarda le serie territoriali, l'aspetto principale è la presenza, accanto all'elemento di natura descrittiva, di un elemento geografico, che, preso come una componente da rendere graficamente, di per sé occupa entrambe le dimensioni del piano. Una visione d'insieme corretta delle serie territoriali, relative a caratteri quantitativi, con le tecniche attualmente disponibili, appare consentita solo se rappresentiamo, insieme alla componente geografica, una sola componente descrittiva, che viene a coincidere con le frequenze od altre quantità, relative ad una singola modalità del carattere in studio. Le difficoltà tecniche di rappresentazione congiunta dei dati quantitativi di due modalità associate alla stessa zona possono essere parzialmente superate tramite la rappresentazione di quantità derivate dal confronto fra le due distribuzioni territoriali, come si vedrà nel Capitolo 4.

In sintesi, si può dire che, nel campo delle rappresentazioni grafiche in statistica, le componenti dell'informazione sono almeno due e che, nel caso di serie statistiche non territoriali (casi a, b), la prima di esse è costituita da frequenze, ammontari o quantità derivate, mentre l'altra coincide con il carattere e le sue modalità. Nel caso delle serie territoriali (caso c), scegliendo di rappresentare la componente geografica, la modalità rappresentabile sul grafico è generalmente unica; al fine di preservare una corretta ed immediata visione d'insieme della distribuzione del fenomeno sul territorio. Possono fare eccezione le rappresentazioni relative a caratteri qualitativi le cui modalità siano univocamente associate al sistema di zone rappresentato.

In statistica si usa distinguere ulteriormente fra le variabili quantitative a seconda che possiedano una scala di tipo intervallo o di tipo rapporto. Per la prima si intende una scala che possiede uno zero, ovvero un punto di partenza dei valori, meramente convenzionale, come avviene, ad esempio, per le temperature misurate in gradi centigradi (°C) o per il tempo; per la seconda si intende una scala che possiede uno zero intrinseco come avviene, ad esempio, per ammontari o numeri indici.

Tale distinzione ha delle implicazioni nelle rappresentazioni grafiche delle serie statistiche, come vedremo nel paragrafo 2 del successivo capitolo.

3. CLASSIFICAZIONE DEI TIPI DI GRAFICI

Generalmente si distinguono quattro differenti tipi di rappresentazioni grafiche sul piano: diagrammi, cartogrammi, grafi e ideogrammi. Per quanto riguarda gli ideogrammi va ricordato l'interessante lavoro di O. Neurath (1936), che si poneva lo scopo di definire la struttura di un linguaggio visivo.

Va tuttavia osservato che la loro utilizzazione sistematica nelle rappresentazione grafiche di dati statistici comporta delle difficoltà già per articolare un'analisi statistica elementare.

In contesti diversi, comunque, essi costituiscono un mezzo per trasmettere in modo semplice ed immediato delle informazioni come un messaggio o una semplice indicazione, di tipo schematico, come si può vedere dagli esempi riportati dalla figura 2.1.

Questo strumento grafico svolge inoltre un ruolo importante in campo cartografico, dove al significato suggerito dall'ideogramma, che può o meno richiamarsi al fenomeno, si associa l'informazione di tipo spaziale.

Da un punto di vista meramente formale, si possono considerare degli ideogrammi gli stessi modelli geometrici utilizzati nello studio, nella progettazione e nel disegno di macchine, utensili e strutture spaziali le più diverse; nella sostanza essi richiedono un approccio completamente diverso e caratterizzato nel senso della modellistica matematica.

Per quanto riguarda i grafi, essi rappresentano relazioni fra gli elementi di una sola componente. Nella pratica sono sovente utilizzati per la rappresentazione di organigrammi e/o dei flussi procedurali, mentre nella statistica sono utilizzati in applicazioni molto specifiche come, ad esempio, l'analisi dei gruppi. La loro presenza è pervasiva ma compete a contesti diversi da quello qui trattato della rappresentazione grafica dei dati statistici, che concerne, per lo più, almeno due componenti, un carattere e delle intensità (o degli individui).

Il diagramma si può caratterizzare come quel tipo di grafico che mette in relazione due componenti distinte ed è certamente lo strumento più utilizzato in statistica come anche nei settori più diversi (scientifici, dell'organizzazione aziendale, gestionale, ecc.).

In campo statistico il problema più comune, come si è visto, è rappresentare la relazione fra una componente costituita dalle modalità di un carattere ed una componente quantitativa data da frequenze, ammontari o quantità derivate.

Per rappresentare graficamente delle quantità sul piano si possono utilizzare metodi diversi: a) posizione dei punti lungo una scala fissa; b) variazioni della lunghezza di segmenti; c) variazioni di aree o di volumi.

Sulla base di alcune ipotesi congetturali, sostenute dai risultati di ricerche empiriche nel campo della percezione umana, si evidenzia (Cleveland e McGill 1984) che la maggiore efficacia, nella rapidità di lettura, nella resa corretta della informazione e nella capacità di discriminazione fra grandezze diverse, sia da attribuire alla valutazione della posizione di punti lungo una scala e quindi alla valutazione ed il confronto della lunghezza di segmenti o barre.

Tali risultati, pur non potendosi considerare definitivi ed anzi suscettibili di ulteriori precisazioni, forniscono al momento attuale una base per scelte razionali.

Nel caso dei diagrammi appare ragionevole privilegiare come variabili visive l'uso della posizione dei punti lungo una scala ed il confronto di lunghezze, che danno, comunque, ampie possibilità di rappresentazioni grafiche[5].

Le variazioni tridimensionali delle rappresentazioni dei diagrammi a barre non aggiungono nulla all'informazione trasmessa, ed anzi talvolta ne complicano la lettura; la loro utilizzazione è giustificabile se si vuole rendere più attraente esteticamente il grafico in volumi destinati a non specialisti.

L'uso ottimale di tale modalità di presentazione e confronto richiede che le due componenti da rappresentare siano attribuite alle due direzioni ortogonali del piano, l'asse orizzontale delle ascisse e quello verticale delle ordinate.

In questo tipo di rappresentazioni sono compresi i diagrammi cartesiani, che hanno grande rilevanza in statistica, nelle diverse varianti in cui sono utilizzati, come i semilogaritmici e doppiologaritmici.

Per le serie storiche e le altre distribuzioni di tipo lineare appare piuttosto naturale utilizzare l'asse delle ascisse per rappresentare il carattere "tempo" nel suo sviluppo lineare e che l'asse delle ordinate venga utilizzato, con una scala opportunamente scelta, per rappresentare le frequenze od altre intensità.

Nel caso di distribuzioni di frequenze connesse ad un carattere qualitativo, in particolare di tipo sconnesso, può convenire generalmente di rappresentare le differenti modalità lungo l'asse delle ordinate; ciò sembra consentire una descrizione verbale più facile da leggere ed associare alle corrispondenti barre che rappresentano l'intensità del fenomeno[6].

(5) Nel caso di diagrammi di correlazione e di cartogrammi per serie territoriali riferite a zone, diventa naturale e corretto l'uso di simboli ad impianto areale; il problema è allora la scelta di un sistema di simboli basato su criteri scientifici.

(6) Per le serie relative ad un carattere ciclico si potrebbe pensare all'uso di diagrammi polari; tuttavia l'eseguità di tali serie e la possibilità di rappresentarle correttamente, tranne alcuni esempi particolari, come serie di tipo lineare, ci esime dal considerare esplicitamente questo caso ulteriore.

Per quanto riguarda i cartogrammi ci limitiamo al momento a presentare alcune considerazioni generali per riprendere l'argomento in modo più esteso nel Capitolo 4.

Il problema che si presenta nella costruzione di un cartogramma è che l'informazione di tipo descrittivo deve andare a sovrapporsi graficamente alla rappresentazione della componente geografica, che già da sè occupa entrambe le dimensioni del piano.

Un primo criterio da rispettare è che l'informazione di natura statistica va resa graficamente riproducendo, per quanto è possibile, le proprietà geometriche del dato, che si presenta associato a punti, linee o zone.

Accanto a tale aspetto, c'è il problema di render in modo adeguato e consistente la componente descrittiva del dato territoriale. Dal momento che la base geografica occupa di per sè entrambe le dimensioni del piano, occorre utilizzare altre variabili visive, che sovrapposte al disegno di questa, riescano a rendere adeguatamente l'informazione statistica.

Secondo Bertin (1983) la grafica mette a disposizione sei possibili variabili visive, oltre le due direzioni del piano: dimensione, valore, grana, colore, orientamento e forma. Nella figura 2.2 vengono riportate delle esemplificazioni dell'uso di tali variabili in relazione a dati areali; nel grafico 2.2a viene simboleggiato come la rappresentazione geografica di base occupa le due dimensioni del piano; vengono poi rappresentati nel grafico 2.2b i vari modi in cui può essere resa la variazione di dimensione; seguono gli esempi relativi a : la variazione di valore, legata alla percentuale di nero distribuito su un'area bianca; la variazione di grana che è basata sulla variazione del disegno stante la costanza di valore, nel senso che la quantità di nero distribuita su una data area bianca rimane fissa, mentre quel che varia è il modo in cui tale quantità di nero viene distribuita; la variazione di colore; la variazione di orientamento e infine di forma.

Sulla base di argomentazioni di ordine teorico e di carattere empirico, Bertin riesce a classificare le variabili visive in funzione delle loro proprietà percettive e di alcune operazioni logiche che esse consentono di effettuare.

Secondo Bertin la percezione di tipo quantitativo deve consentire di valutare in quale rapporto numerico si trovano due simboli. Il test che egli suggerisce per definire di tipo quantitativo una variabile visiva è chiedere al lettore il valore del simbolo più grande quando è stato definito il valore del simbolo minore. La variabile visiva in grado di superare tale test è la dimensione.

La variazione di valore, pur essendo legata alla percentuale di nero distribuito su un'area bianca, non possiede, secondo il Bertin, questa proprietà dal momento che l'occhio non è in grado di

ricostruire una scala quantitativa dalle varie gradazioni di grigio; essa si presta, perciò, a rappresentare correttamente delle relazioni di ordine fra le modalità di un carattere, quando queste non sono in numero eccessivo.

Si rimanda al testo di Bertin (1983) per una trattazione più ampia dell'argomento. Ci limitiamo ad aggiungere che il colore si presta compiutamente alla rappresentazione di differenze di qualità e consente anche associazioni visive immediate fra zone omogenee, ugualmente colorate; ma non si presta a riprodurre compiutamente relazioni d'ordine nè tanto meno di tipo quantitativo.

Nel seguito di questo fascicolo faremo riferimento alle tre variabili visive seguenti : dimensione, nel paragrafo 5, valore e colore nel paragrafo 6. Esse sono sufficienti a rendere in modo coerente, rispettivamente, le componenti descrittive di tipo quantitativo, qualitativo ordinabile e qualitativo sconnesso.

Si deve tener presente, comunque, che la variazione di grana può essere in grado di rendere efficacemente componenti di tipo qualitativo e va sicuramente considerata nella predisposizione dei grafici.

Un criterio che discende in modo naturale dalle argomentazioni precedenti è che, nella scelta di rappresentazioni grafiche basate su criteri scientifici, la variabile visiva deve essere adeguata alle caratteristiche della componente descrittiva per rendere appieno il contenuto dell'informazione.

Molte serie territoriali, nella pratica, si presentano come distribuzioni di frequenze, ammontari o quantità derivate, associate da una parte ad entità territoriali e dall'altra alle modalità di un carattere. Sul cartogramma non è possibile tuttavia rappresentare in modo semplice e corretto, preservando una visione d'insieme del fenomeno, più di una singola modalità del carattere. Quest'ultima viene a costituire la seconda componente, quella descrittiva, che il cartogramma può riuscire a rendere compiutamente.

Una particolare attenzione va perciò dedicata alla scelta di una simbologia adeguata alla rappresentazione di dati quantitativi; a tal proposito il sistema di simboli GRP, dall'inglese *Graphical Rational Patterns,* proposto da Bachi (1968), risulta rispondere al criterio individuato da Bertin e basato su principi costruttivi molto semplici, suscettibili di verifiche empiriche e, quindi, di possibili miglioramenti. Indichiamo in italiano tali simboli come «simboli grafici razionali» e li indichiamo in modo abbreviato come s.g.r..

Nel paragrafo 7 del prossimo capitolo si tratterà brevemente dei diagrammi areali, in particolare dei diagrammi a torta e dei diagrammi a barre suddivise, e verranno sottolineati i problemi che essi pongono per una visione rapida e corretta dell'informazione e gli ambiti limitati della loro possibile utilizzazione.

4. ALCUNE OSSERVAZIONI SULLA RAPPRESENTAZIONE GRAFICA DELLE SERIE MULTIPLE

Per quanto riguarda il caso delle serie statistiche multiple ci limitiamo ad alcune esemplificazioni relative al caso di distribuzioni di frequenze, od altre quantità, associate a due e tre caratteri.

Dal punto di vista grafico, per le serie statistiche multiple, sono possibili soluzioni grafiche diverse, in funzione di come vengono organizzate le dimensioni del piano. Si possono dedicare le due direzioni ortogonali alla rappresentazione dei due caratteri ed utilizzare una terza variabile visiva, ad esempio dei simboli grafici ad impianto areale adatti a rendere graficamente delle quantità, per rappresentare la componente quantitativa della informazione.

Un importante strumento di rappresentazione ed analisi dei dati statistici è il diagramma di correlazione, il quale consiste nell'indicare graficamente su un diagramma cartesiano l'intensità del fenomeno in corrispondenza a certi insiemi di valori delle modalità di due variabili. Tuttavia in questo lavoro non si discuterà di esso come anche di strumenti grafici similari, destinati alla rappresentazione di distribuzioni relative alle modalità incrociate di più caratteri [7].

Per semplicità si limiterà la trattazione alla rappresentazione di una sequenza di distribuzioni, ordinata in funzione della priorità assegnata ai caratteri.

È importante sottolineare che, nel selezionare le relazioni da rappresentare graficamente, la prima delle componenti resta comunque associata alle frequenze, ammontari o quantità derivate, lo studio delle cui variazioni costituisce il nostro precipuo interesse.

Per quanto riguarda le altre due componenti, conviene che esse vengano ordinate in funzione della priorità attribuita ai caratteri cui vengono associate. Questi possono essere distinti in caratteri di prima e di seconda specie (Salvemini, Girone 1981), al fine di operare una selezione fra le possibili diverse relazioni rappresentabili tra le varie componenti.

Ad esempio, nel caso della distribuzione della popolazione secondo l'età ed il sesso le componenti individuabili sono le seguenti:

a) ammontare della popolazione, componente quantitativa;
b) età, componente quantitativa;
c) sesso, componente qualitativa non ordinabile.

(7) Si ricorda la possibilità di utilizzare in questo caso le curve isometriche, descritte diffusamente dal Pedroni (1968).

Nel selezionare le relazioni da rappresentare, conviene considerare l'età come carattere di prima specie ed il sesso come carattere di seconda specie. Tale scelta dà luogo alla tradizionale piramide delle età, di cui viene fornito un esempio ripreso dall'*Annuario Statistico Italiano 1988* nella figura 2.3. È ovviamente legittima la scelta opposta che darebbe comunque luogo ad un diverso ed inusuale tipo di rappresentazione grafica.

Nel caso della distribuzione dei morti per classe di età e causa di morte, le tre componenti sono ovviamente le seguenti:

a) il numero dei morti, componente quantitativa;

b) le classi di età, componente ordinabile ottenuta dalle età per aggregazione;

c) le cause di morte, componente di tipo qualitativo non ordinabile.

La scelta delle classi di età come carattere di prima specie ci porta a selezionare un tipo di rappresentazione grafica piuttosto che un altro. Sarebbe certamente legittimo pensare ad una rappresentazione che per ogni classe di età riporta la distribuzione dei morti in funzione delle cause selezionate.

Nella figura 2.4 vengono forniti due esempi in cui il tempo è stato considerato rispettivamente come variabile di 1° e 2° specie; questa ultima scelta appare ragionevole solo se il numero dei dati di tempo è limitato.

Per un insieme di misure congiunte associate allo stesso carattere, che possiamo definire «serie parallele» e considerare come una variante delle serie multiple, come ad esempio la misura della superficie dei terreni dedicati a certi tipi di coltivazione e della corrispondente produzione, si può applicare l'analisi dell'informazione in questo modo. Si ha una componente quantitativa relativa all'ammontare delle grandezze e due altre componenti associate rispettivamente al carattere «tipo di coltivazione» ed alla coppia di variabili parallele «superficie, produzione». Il problema posto dalla diversità della unità di misura delle due variabili può essere risolto trasformando tali quantità in numeri adimensionali, dividendo ad esempio le singole misure per l'ammontare complessivo della corrispondente grandezza. Altrimenti la componente quantitativa richiederà per la sua rappresentazione l'utilizzazione di due scale distinte.

Al fine di chiarire i problemi su trattati, si fa riferimento ad un esempio ripreso da Bonin (1975), relativo all'emigrazione in sei paesi per quattro anni consecutivi, dal 1966 al 1969. I dati su cui si basa l'esempio sono dati nel seguente prospetto:

Prospetto 1 — Persone emigrate da alcuni paesi negli anni dal 1966 al 1969 (dati in migliaia di unità)

Anni	Portogallo	Spagna	Italia	Yugoslavia	Grecia	Turchia	Totale
1966	87	57	219	79	54	46	**542**
1967	64	55	164	83	17	54	**437**
1968	53	67	158	110	25	65	**478**
1969	42	80	140	243	63	160	**728**
Totale	**246**	**259**	**681**	**515**	**159**	**325**	**2.185**

Dividendo le quantità di ogni paese per il corrispondente valore totale dell'emigrazione in quattro anni, o quelle relative ad ogni anno per l'emigrazione complessiva dai sei paesi, si ottengono tabelle con le colonne e le righe, rispettivamente, ricondotte ad ammontari unitari; talvolta può essere proficuo rappresentare tali dati trasformati piuttosto che i valori iniziali.

Fissando la nostra attenzione sulla tabella data, le possibili sequenze di grafici rappresentabili, sulla base della procedura indicata, sono riportati nella figura 2.5, ripresa da Bonin (1975); se si sceglie come carattere di prima specie l'insieme dei paesi si ottiene il grafico 2.5b.

Entrambe le sequenze paiono statisticamente ammissibili, anche se, in questo caso, l'evoluzione temporale del fenomeno relativamente a ciascun paese sembra fornire un'informazione più intellegibile.

Nei grafici proposti nella figura 2.5 le due direzioni del piano, in ciascun diagramma, risultano assegnate ad una sola componente, evitando che l'omogeneità di queste venga interrotta, alternando le modalità di due caratteri differenti.

Nella figura 2.6a viene riportato un grafico relativo al medesimo prospetto su riportato, dove sull'asse delle x vengono rappresentati i vari paesi, mentre l'asse delle y è destinato a rappresentare contemporaneamente le quantità ed il tempo. Ciò rende indubbiamente più complicata la lettura dei dati e fa perdere semplicità ed efficacia al grafico, costringendo inoltre il lettore a confrontare delle altezze che partono da basi diverse.

Nella figura 2.6b vengono riportati due grafici per una coppia di ipotetiche serie storiche relative alla superficie destinata alla coltivazione del frumento e della relativa produzione; la diversità delle unità di misura costringe ad usare due scale diverse. Nel grafico a sinistra la rappresentazione delle due misure si alterna rompendo l'omogeneità della direzione orizzontale del piano; nel grafico a destra vengono rappresentate separatamente le due misure con gran vantaggio nella percezione dell'andamento delle due serie. In questo esempio il tempo è stato selezionato come carattere di

prima specie e la coppia di variabili superficie-produzione come carattere di seconda specie.

Riportiamo, in conclusione, nella figura 2.7 un ulteriore esempio, ripreso anch'esso dal Bonin (1975), per una serie relativa a tre caratteri:

a) il tempo,
b) tipo di coltivazione di cereali (frumento, avena, segale),
c) la coppia di misure (superficie, produzione).

Appare evidente come rendere separatamente le varie serie, semplificando la struttura dell'asse delle x ed organizzando in modo adeguato la sequenza dei diagrammi, facilita notevolmente la lettura del grafico.

Tornando al caso di due caratteri, le differenti distribuzioni, ciascuna ottenuta dallo sviluppo delle varie modalità del carattere di prima specie e parametrizzate dalle modalità del carattere di seconda specie, possono essere organizzate sul piano in una sequenza di distribuzioni giustapposte.

Le piramidi delle età forniscono un esempio di rappresentazione di questo tipo. Nel caso della distribuzione dei morti per età e causa di morte, dal momento che il carattere di seconda specie possiede un numero di modalità elevato, dopo aver selezionato le più significative, conviene rappresentare le varie distribuzioni in una serie di grafici fra loro paralleli, messi in corrispondenza della stessa suddivisione delle età lungo l'asse delle ascisse (figura 1.1h del Capitolo 1).

Un ulteriore elemento di accurata valutazione è se convenga rappresentare la distribuzione dei valori assoluti o, invece, proporzionare questi ad un ammontare complessivo costante.

Una rappresentazione grafica che può essere talvolta utile per i confronti fra distribuzioni parallele è il grafico delle differenze fra le intensità relative a modalità uguali del carattere di prima specie, che consente una più accurata valutazione di queste grandezze.

5. IL SISTEMA DEI «GRAPHICAL RATIONAL PATTERNS» (GRP)

Per rappresentare dati quantitativi, nel caso di serie territoriali associate a zone di area non nulla, è disponibile presso l'Istat il sistema dei «Graphical Rational Patterns» (GRP) o simboli grafici razionali, che indichiamo in modo abbreviato come s.g.r., basato su criteri scientifici e di semplice uso. Esso è stato proposto da

Bachi nel 1968 e successivamente nel 1978 in modo informatizzato.

Il sistema dei GRP si basa su tre criteri costruttivi principali. Il primo è la proporzionalità fra la quantità numerica da rappresentare e la quantità di inchiostro del simbolo che la rappresenta; il secondo criterio è la costanza di forma (per i simboli adottati in Istat, tale forma è il quadrato), che dovrebbe consentire la fusione dei simboli a livello di visione generale del cartogramma, per garantire una valutazione complessiva delle caratteristiche della distribuzione.

Per ovviare alla imperfezione della vista umana nella valutazione delle grandezze e all'elemento soggettivo comunque presente in tale stima, si utilizza come terzo criterio costruttivo dei simboli il sistema decimale. Quando si passa dalla visione globale a quella puntuale, la valutazione può essere resa tanto accurata quanto i limiti della approssimazione della scala prescelta lo consentono.

Nella scala che copre con passi unitari il campo dei valori interi compresi da 1 a 100, vengono definiti tre simboli base, l'unità, il dieci ed il cento; per essi la quantità di inchiostro utilizzato è nella proporzione definita dalla sequenza di questi valori. Dalla combinazione di tali simboli, opportunamente disposti in forme regolari e compatte, si può facilmente costruire l'intera sequenza degli interi da 1 fino a 90. Per mantenere costante lo spazio occupato dalla simbologia, i valori da 91 a 99 vengono rappresentati tramite una cornice quadrata, che copre uno spazio proporzionale al valore 90, al cui interno vengono riprodotti i valori da 1 a 9.

Si riporta nella figura 2.8 la sequenza dei valori da 1 a 100; nell'ultima riga si hanno i simboli che rappresentano i valori da 1 a 10; nell'ultima colonna si hanno i simboli che, partendo dal basso, rappresentano le decine da 10 a 100; qualsiasi valore intermedio si può ottenere nel punto di incrocio fra la decina e l'unità che definiscono il numero.

6. L'USO DELLA VARIAZIONE DI «VALORE» E DI COLORE

In questo paragrafo si riprendono alcune considerazioni sull'uso del colore e della variazione di «valore» nei grafici statistici. La variazione di «valore» è definibile come la progressione che l'occhio percepisce nell'intervallo dei grigi compreso fra il nero ed il bianco. Questa progressione non dipende dal colore; variazione di valore si può avere sulla base di colori diversi come il blu o il rosso.

Nel caso di assenza di colore la variazione di valore genera una scala di chiaroscuri acromatica, che si può ottenere mescolando il

nero e il bianco in proporzioni diverse o spalmando sempre più il nero sopra una data superficie bianca. In pratica si costruiscono scale da 4 a 12 passaggi tonali al più ed è ovvio che quanto più la scala è limitata tanto più le gradazioni appaiono diverse fra loro e meglio percepibili le differenze.

In termini rigorosi si può definire il valore di una superficie come il rapporto fra la quantità totale di nero e la sua area; si può associare al nero pieno il valore 1 e al bianco di base il valore 0. In letteratura vengono forniti dettagli per la costruzione di scale appropriate, per rendere ottimale la distinzione fra i differenti gradi (Bertin 1983). Nonostante la possibile definizione quantitativa del valore, l'occhio umano non pare che sia in grado di ricostruire da esso più di una scala di tipo ordinale (Bertin 1983).

Tale variabile può perciò rappresentare correttamente componenti associate a caratteri di tipo qualitativo ordinabile. Bertin non consiglia l'uso, nella grafica statistica, di un numero di gradazioni superiore a sei o sette, comprendenti i valori nero e bianco degli estremi.

La capacità espressiva del colore si apre al mondo della ricerca dei rapporti cromatici, armonici o di contrasto, con valenze di espressione artistica.

Esiste una vasta ed interessante letteratura sull'argomento; per padroneggiare la vasta gamma delle possibilità compositive che consente il colore, se ne è definita una grammatica ed una sintassi, sulla cui base è possibile individuare differenze sottili fra i colori ed articolarne il gioco dei possibili contrasti (Itten 1982; Luzzatto, Pompas 1980). Il colore è anche interpretabile simbolicamente.

Per quanto riguarda l'uso del colore è comunque necessario far presente alcune considerazioni preliminari; come già per il valore vanno distinti gli aspetti fisici e quelli percettivi del colore.

Come è noto la percezione del colore nasce dalla stimolazione dell'occhio da parte della luce, che è radiazione elettromagnetica caratterizzata da una ben determinata distribuzione di energia sulle frequenze che costituiscono lo spettro del visibile.

Dal punto di vista fisico, si può perciò definire sinteticamente il colore in base alla sua composizione spettrale.

Ciò che complica il processo della visione è l'intervento del complesso sistema della percezione e della sua elaborazione da parte del cervello umano; senza entrare nei dettagli della descrizione di tale sistema, ci limitiamo a ricordare che la retina è composta da due tipi di cellule sensibili alla luce: i coni e i bastoncelli. Questi ultimi hanno un'elevata sensibilità al buio, in presenza di scarsa quantità di luce, e non distinguono i colori. Al contrario sono i coni a svolgere la funzione specifica di percezione del colore ed esistono diverse teorie che tentano di spiegare gli aspetti fisiologici di tale processo; fra queste ricordiamo la prima e

più famosa, che risale a più di 100 anni fa, di Young ed Helmoltz, che distingue tre tipi di coni con sensibilità diverse alle diverse frequenze; in questa teoria si distinguono coni sensibili al verde, al rosso e al blu; dall'attività combinata di questi, vengono percepiti tutti gli altri colori.

Sulla base di tale teoria diventa necessario distinguere fra colori metamerici e non metamerici. Viene definito metamerico quel colore che nasce dalla combinazione, secondo percentuali determinate, dei tre colori fondamentali e si distingue dal non metamerico soltanto per la sua composizione spettrale, dal punto di vista fisico e non percettivo. Ad esempio il verde può essere visto sia come colore puro, caratterizzato da una composizione spettrale che si accentra intorno ad una ben determinata frequenza, sia come la sovrapposizione di blu e giallo, che insieme danno un composizione spettrale completamente diversa dalla prima. Colori metamerici e non possono, comunque, dar luogo ad impressioni visive molto diverse a seconda del grado e tipo di illuminazione dei due colori.

Sulla base degli sviluppi di tali teorie, sono stati costruiti diversi modelli per la classificazione dei colori, i quali possono essere distinti in funzione della scomposizione che utilizzano: di tipo fisico o di tipo percettivo.

Fra i primi vanno ricordati i modelli RGB e YMC ed il modello CIE nei suoi diversi aggiornamenti. Il modello RGB fa riferimento alla proprietà prima descritta dei colori di poter essere ottenuti dalla combinazione, secondo percentuali date, di rosso (Red), verde (Green) e Blu. È noto che tale modello fa riferimento alla sintesi additiva del colore, quella sintesi che partendo dal nero, per aggiunta di luce di colore diverso genera i vari colori e si presta perciò molto bene alla individuazione di questi per i terminali grafici; una esemplificazione di tale processo è visibile nella figura 2.9.

Al contrario il modello YMC fa riferimento alle proprietà della sintesi sottrattiva, di cui si fa uso nella determinazione dei colori nei processi di stampa; si parte infatti in questo caso dal bianco e per successive aggiunte di strati dei colori fondamentali, secondo percentuali fissate, si generano i vari toni; tali colori fondamentali sono il giallo (Yellow), Magenta e Ciano. Si aggiunge talora il nero come ulteriore colore base per arricchire la gamma delle possibili sfumature.

Esistono delle formule di trasformazione da una composizione all'altra; tuttavia, anche perchè l'associazione coinvolge necessariamente un colore metamerico, l'aspetto dei due colori può dipendere notevolmente dal contesto in cui vengono percepiti.

In campo internazionale, fin dal 1931, la Commission Internationale d'Eclairage (CIE) definì un modello, scientificamente basato, per la univoca definizione di ogni possibile colore, elimi-

nando alcune difficoltà concettuali cui dà adito il modello RGB, in particolare l'apparire per la definizione di alcuni colori di pesi negativi. Lo schema del CIE, che ha avuto successivamente degli aggiornamenti, si basa sulla combinazione di tre colori X,Y,Z definiti teoricamente, da cui può ottenersi qualsiasi altro colore.

La figura 2.10, ripresa dal testo di Frova (1984), è ricavata in base alle ipotesi del modello. Lungo il perimetro curvo sono posizionati i colori puri, ciascuno caratterizzato da una ben precisa lunghezza d'onda, riportata sul grafico in nanometri (= 10^{-9} metri); al suo interno sono posizionate tutte le sfumature ottenibili dalla sovrapposizione di almeno due colori puri del bordo.

L'altro tipo di modelli in uso nella pratica è quello che fa riferimento alle proprietà percettive dei colori. Fra questi ricordiamo il sistema HLS, usato dalla Tektronix e basato a sua volta sul modello di Ostwald, e il modello di Munsell. Per una panoramica dei diversi modelli e della convertibilità reciproca, da un modello all'altro, si può far riferimento al testo di Foley e Van Dam (1983).

Attorno al 1915 W. Ostwald definì un doppio cono sulla cui circonferenza più esterna collocò il cerchio di 24 tinte pure che schiariscono verso il vertice bianco nella parte superiore del solido e scuriscono in quello inferiore verso il nero, ottenendo così delle scale chiaroscurali cromatiche; poi sezionò il doppio cono in 24 parti, corrispondenti a dei triangoli in cui la variazione di tinta viene rappresentata nelle diverse gradazioni di chiarezza secondo uno schema preciso. Nella direzione orizzontale ogni tinta si mescola con il colore complementare, perdendo in purezza e brillantezza. Tale sistema è stato ripreso dalla Tektronix che ha fissato tre variabili per definire la posizione dei colori all'interno del cono. H sta per hue, colore puro o tinta, e può assumere i valori: rosso, giallo, verde, ciano, blu, magenta e gli altri colori puri previsti che si susseguono secondo l'ordine antiorario del diagramma del sistema CIE; essi sono ordinati in modo che a ciascun colore corrisponde nel punto diametralmente opposto sulla medesima sezione il colore puro complementare, mescolandosi col quale si genera il grigio in una delle sue possibili gradazioni. L sta per brillantezza e misura la quantità di bianco o di nero aggiunta alla tinta pura; essa vale 0 per il nero ed 1 per il bianco. S sta per il grado di saturazione del colore; per S=0 si ha la minima saturazione ed il colore è solo una gradazione di grigio; per S=1 si ha il colore puro in corrispondenza della tinta prescelta. Il valore di H, del colore puro, determina nel cono la sezione trasversale corrispondente; il valore di L determina la quota a cui si trova il colore prescelto; infine S fissa sul segmento che unisce i due colori complementari di pari luminosità il tono desiderato.

Degno di menzione come sistema di classificazione dei colori è il cosiddetto albero di A.H. Munsell rappresentato schematicamente nella figura 2.11 (Wyszecki, Stiles 1982); questo è organizzato intorno ad un asse centrale acromatico che ne costituisce il tronco, attorno a cui sono disposti i colori puri, posti ad una distanza variabile dal tronco in relazione al grado di saturazione o purezza e collocati più o meno in alto, ed è questa la novità rispetto al cono di Ostwald, in relazione alla loro chiarezza o brillantezza; si propone, in sostanza, una rottura della simmetria a favore della evidenziazione di una qualità specifica del colore, la sua chiarezza o luminosità o valore. In questo solido irregolare un colore molto luminoso come il giallo è collocato in alto rispetto alle altre tinte.

Dopo questa breve digressione sulla teoria del colore torniamo al problema del loro uso in statistica. Nelle rappresentazioni grafiche statistiche l'uso del colore ha la sua massima efficacia nella rappresentazione delle modalità di un carattere qualitativo.

In questa scelta bisognerebbe comunque tener conto della considerazione di Bertin (1983) concernente la confusione che spesso viene fatta fra colore e valore. La serie di dodici colori puri (viola, blu-viola, blu, blu-verde, verde, giallo-verde, giallo, giallo-arancio, arancio, rosso-arancio, rosso, rosso-viola), che non contengono, dal punto di vista cromatico, alcuna quantità di bianco né di nero, e che vengono definiti anche colori saturi, come ben evidenziato dall'albero di Munsell, non ha una luminosità costante per l'occhio umano.

I dodici colori puri sono ordinati in due scale: la prima, la cosiddetta serie di colori freddi, di luminosità crescente, dal viola al giallo; la seconda, la serie dei colori caldi, di luminosità decrescente dal giallo al viola.

Si riprende nella figura 2.12 una pagina del testo di Bertin che riporta in una versione ridotta, nell'ultima sequenza di colori, indicata con il numero "3", la serie dei colori puri. Nella prima e settima striscia si hanno rispettivamente il nero ed il bianco; nella prima ed ultima colonna si hanno quindi dal basso verso l'alto gradazioni crescenti di grigio. I colori puri, indicati con un puntino bianco, sono disposti al livello corrispondente di luminosità. I restanti colori sono ottenuti da quelli puri aggiungendo del bianco o del nero al fine di variarne la luminosità.

Si ha così una rappresentazione a due dimensioni; lungo la direzione verticale varia il valore, cioè la luminosità dei colori, lungo la direzione orizzontale varia il tipo di colore puro, definito anche tinta; dalla combinazione di tali due variabili vengono determinati i vari possibili colori detti anche toni.

In basso la penultima sequenza, sovraindicata con il numero "2", è una serie di toni aventi il medesimo valore; tale proprietà

consente, nella rappresentazione di modalità sconnesse, di dare la medesima importanza a ciascuna di esse.

Al fine di una utilizzazione razionale del colore è necessario tener conto delle sue complesse proprietà.

7. IL LEGAME FRA I SIMBOLI ED I DATI STATISTICI DA RAPPRESENTARE

L'individuazione delle componenti e delle relazioni da presentare graficamente agevola il compito di trovare gli adeguati collegamenti fra gli elementi grafici e le componenti stesse, come di individuare il tema che caratterizza il grafico nel suo complesso.

Questo ultimo, reso sinteticamente, viene a costituire il titolo, da porre preferibilmente in testa al grafico; esso è l'elemento unificante delle parti variabili in esso presenti. Nel titolo vanno specificati, necessariamente se non sono riportati altrove nel grafico, il tempo e lo spazio cui le grandezze e le relazioni presentate si riferiscono. In esso è implicito un breve elenco delle componenti presenti nel grafico che vanno citate per ordine di importanza.

Ciascuna componente va correttamente identificata, dal punto di vista semantico e rispetto agli elementi grafici associati. Deve essere di semplice percezione come è stata resa graficamente la variabilità di ciascuna di esse. Ciò può essere specificato assegnando una scala ad un asse di riferimento od associando una descrizione verbale ai diversi elementi delle componenti presenti nel grafico, anche attraverso l'uso di una appropriata legenda.

Insieme con la scelta del tipo di metodo grafico, di cui si è discusso nel paragrafo 3, va scelto il legame fra i dati da rappresentare ed i simboli usati nella rappresentazione, in riferimento alle trasformazioni numeriche alle quali quelli possono essere sottoposti, al fine di mettere in più chiara evidenza le caratteristiche della distribuzione.

Una scelta possibile è di rappresentare sul grafico una scala che varia dal massimo al minimo dei valori assunti dal fenomeno, al fine di mettere nella massima evidenza le sue variazioni. In certi casi si può essere interessati, volendo confrontare fenomeni diversi, ad eliminare dai valori l'ordine di grandezza del fenomeno, come viene misurato, ad esempio, dal suo valor medio. Tale eliminazione si può effettuare per differenza o per rapporto in funzione delle caratteristiche del fenomeno.

In altri casi può essere necessario eliminare l'effetto dovuto alla diversa variabilità dei dati e rappresentare i valori standardizzati.

In ogni caso sul grafico vanno riprodotti i valori effettivamente

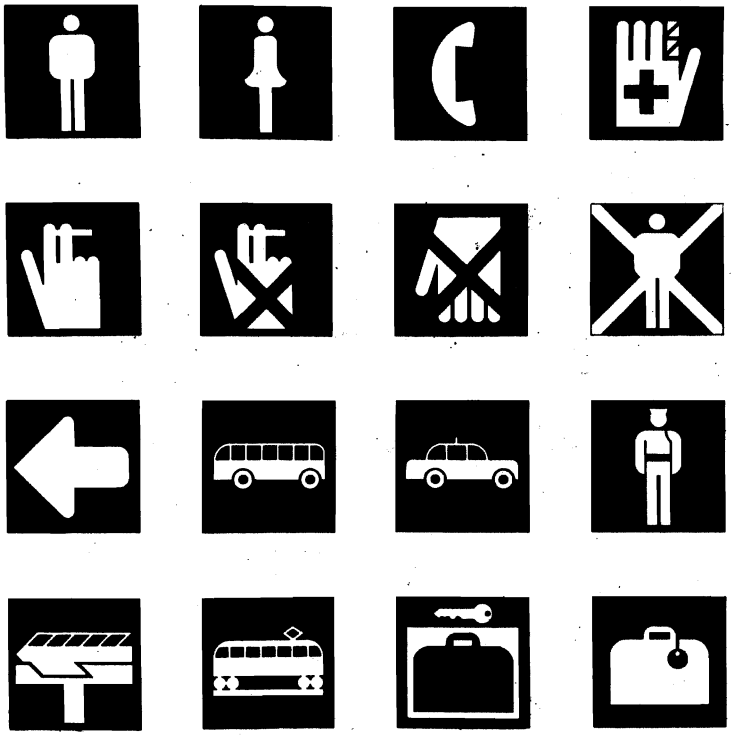
assunti dalla variabile; la trasformazione numerica risulterà in tal modo completamente trasparente al lettore.

Come riferimenti esterni per il lettore si può riportare in una nota il codice della tabella da cui è stata ricavata la rappresentazione, insieme con le altre necessarie indicazioni per la sua lettura corretta; conviene riportare tali informazioni in fondo al grafico, dove possono essere inseriti commenti o note di chiarimento.

Si vuole rimarcare che alcune regole importanti sono contenute negli standard proposti dal Joint Committee americano per la standardizzazione e da Cox nella memoria presentata alla Conferenza di Sheffield (1978), e che abbiamo in parte riportato nel paragrafo 2 del Capitolo 1.

Figura 2.1 — Alcuni esempi di ideogrammi

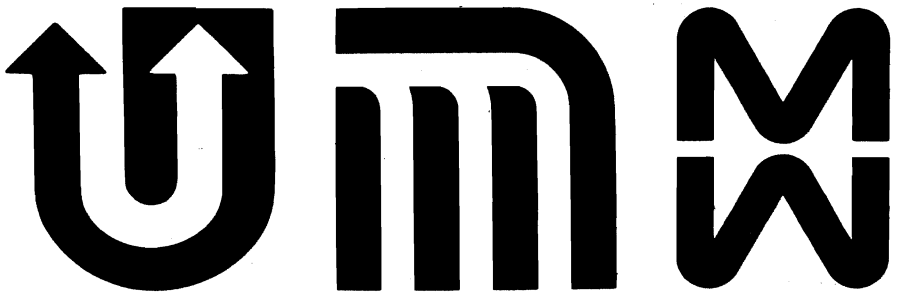
(a)



Segnali per il pubblico all'Expo internazionale di Osaka (Giappone), 1970.

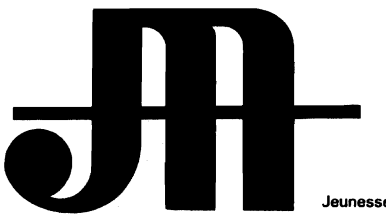
Metropolitana (Rft) Metropolitana di Milano

(b)



Metropolitana di Città del Messico

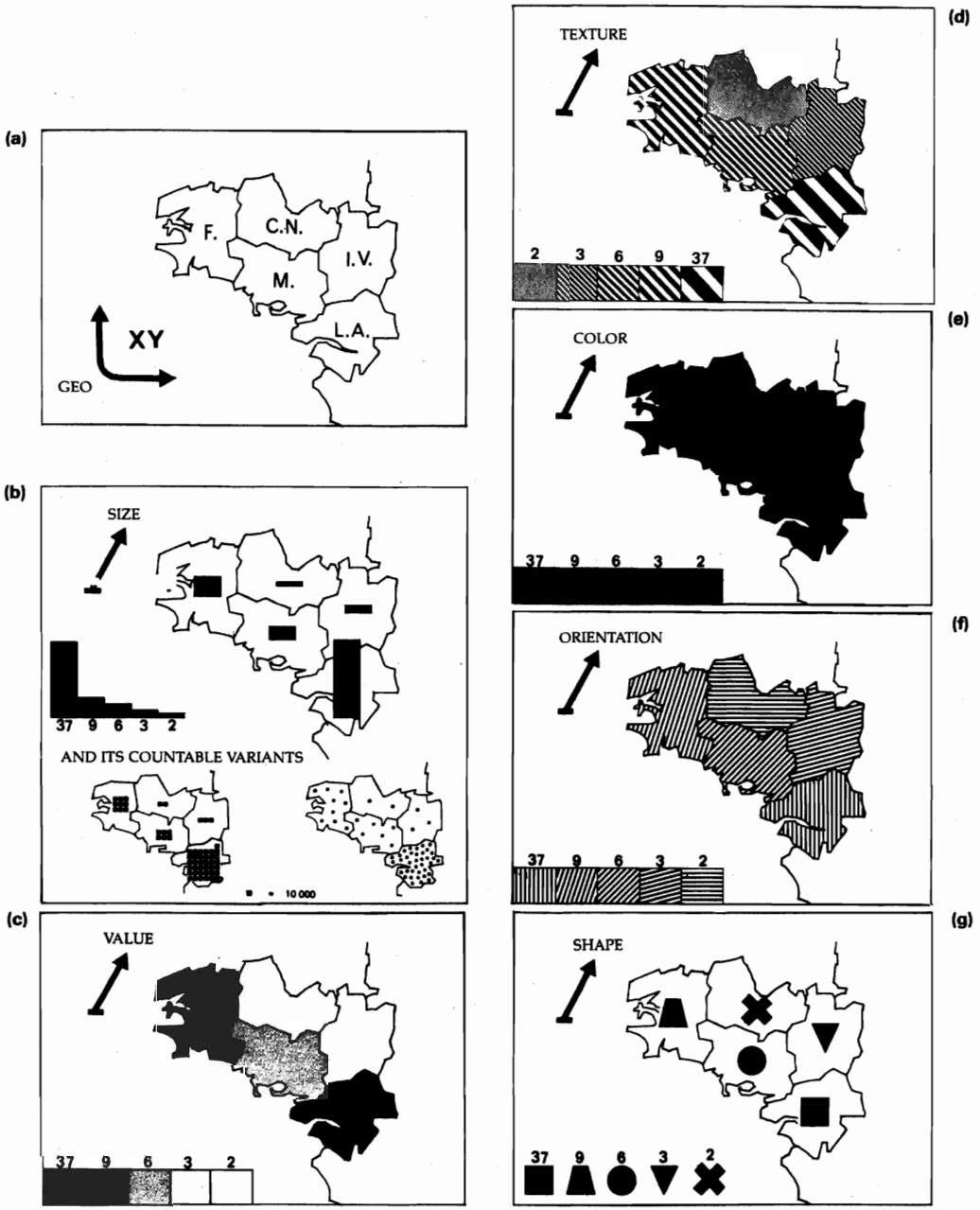
(c)



Fonte: Diethelm (1984)

Jeunesses Musicales de Lousanne (Svizzera)

Figura 2.2 — Le variabili visive per impianti di tipo areale secondo Bertin



Fonte: Bertin (1983)

Figura 2.3 — Piramidi delle età

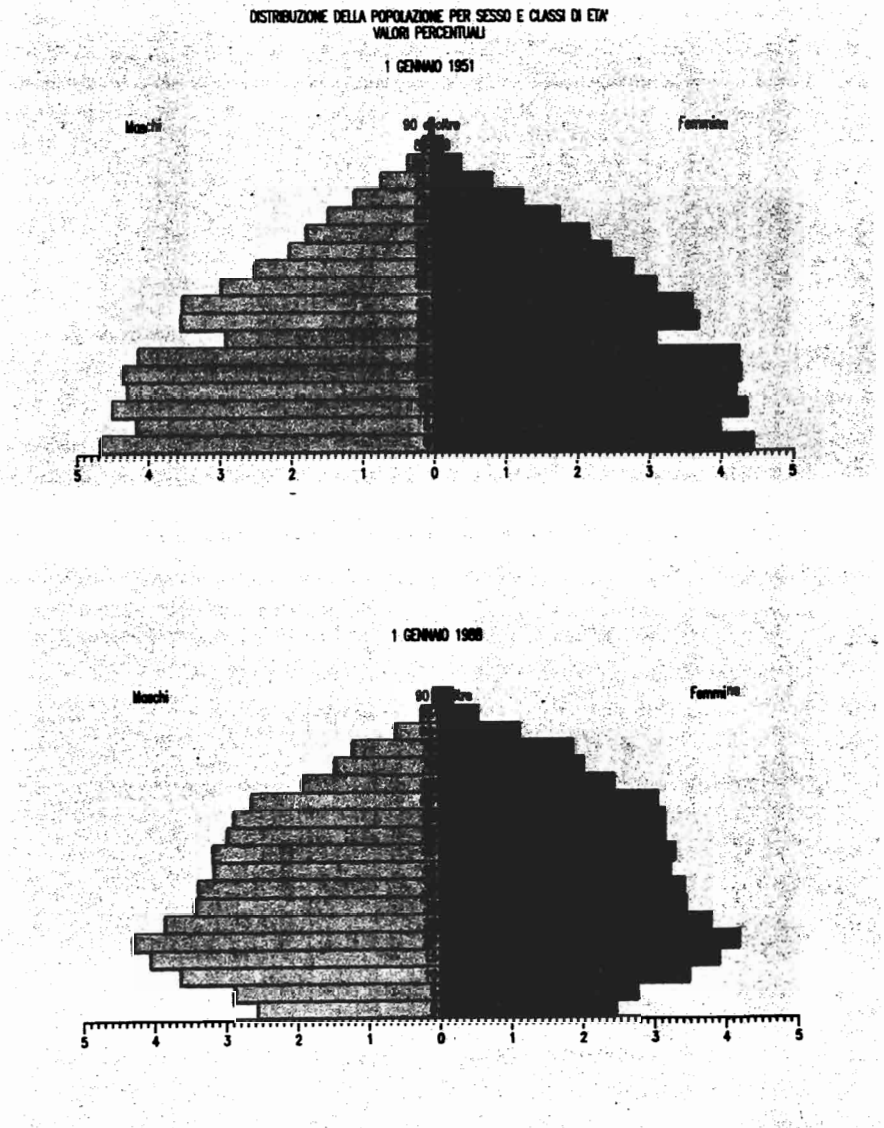
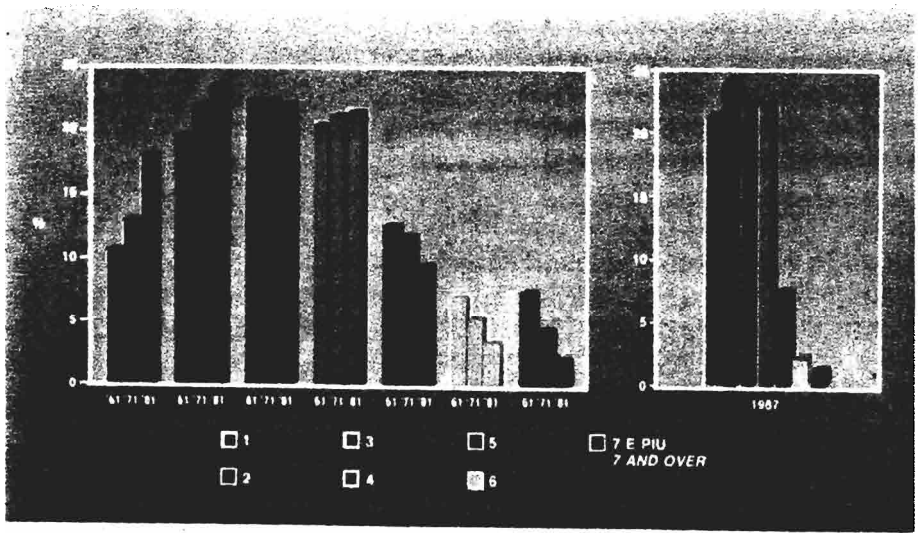
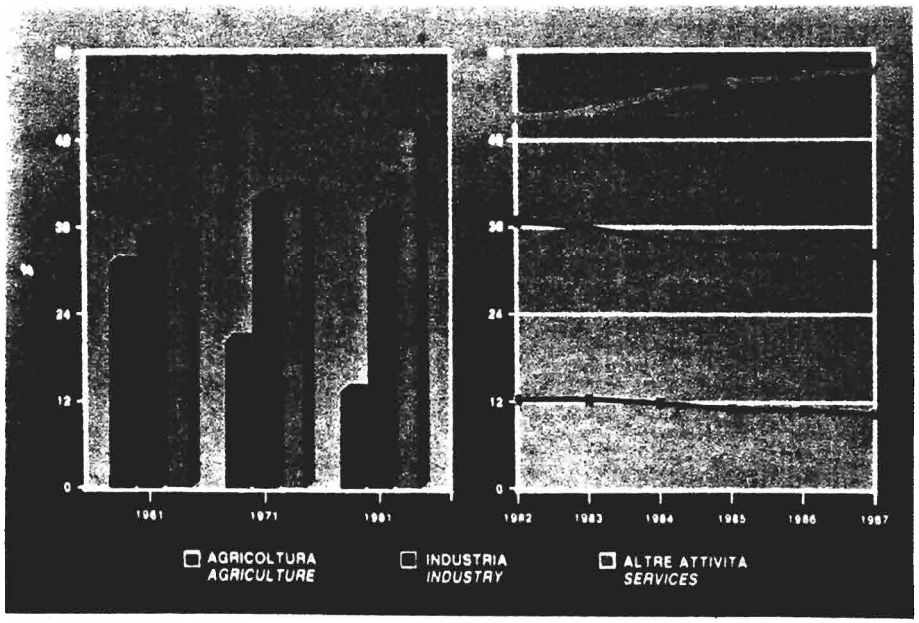


Tavola illustrata n. 2.8 - Piramidi della età della popolazione residente al 1° Gennaio 1951 e al 1°Gennaio 1988

Fonte: Annuario Statistico Italiano (1988)

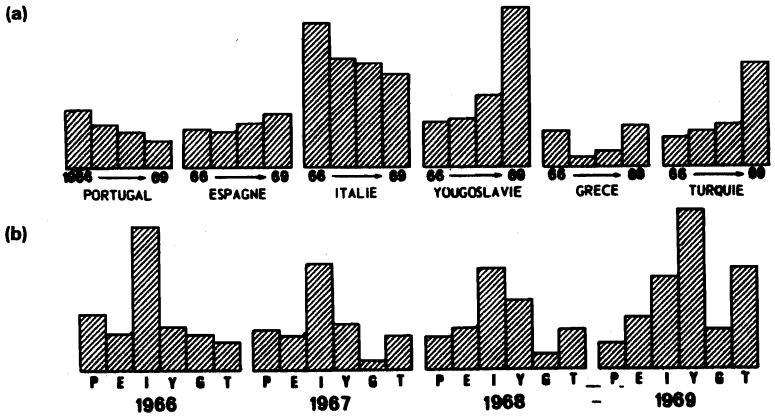
Figura 2.4 — Esempi di grafici in cui è presente come variabile il tempo





Fonte: Conoscere l'Italia - Istat 1988

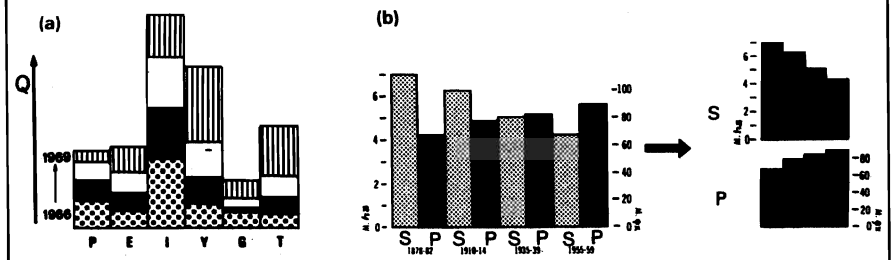
Figura 2.5 — Possibile sequenza dei diagrammi relativi al prospetto di pagina 31



Fonte: Bonin (1975)

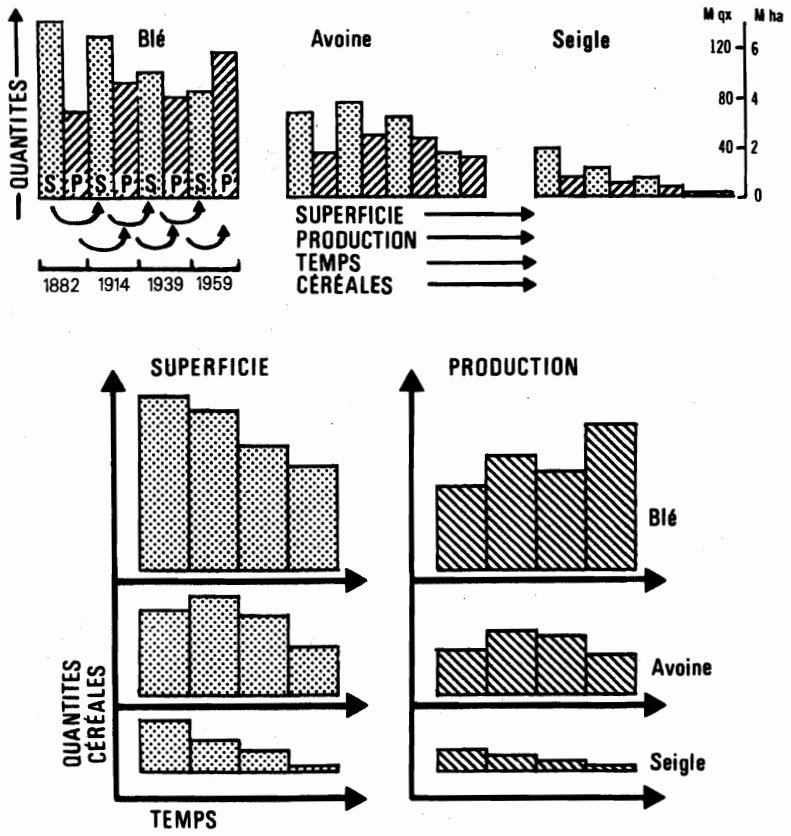
Nota: Nel grafico a) il carattere di prima specie è il tempo.
Nel grafico b) il carattere di prima specie è costituito dall'insieme dei paesi.

Figura 2.6 — Problemi derivanti dalla disomogeneità delle scale



Fonte: Bonin (1975)

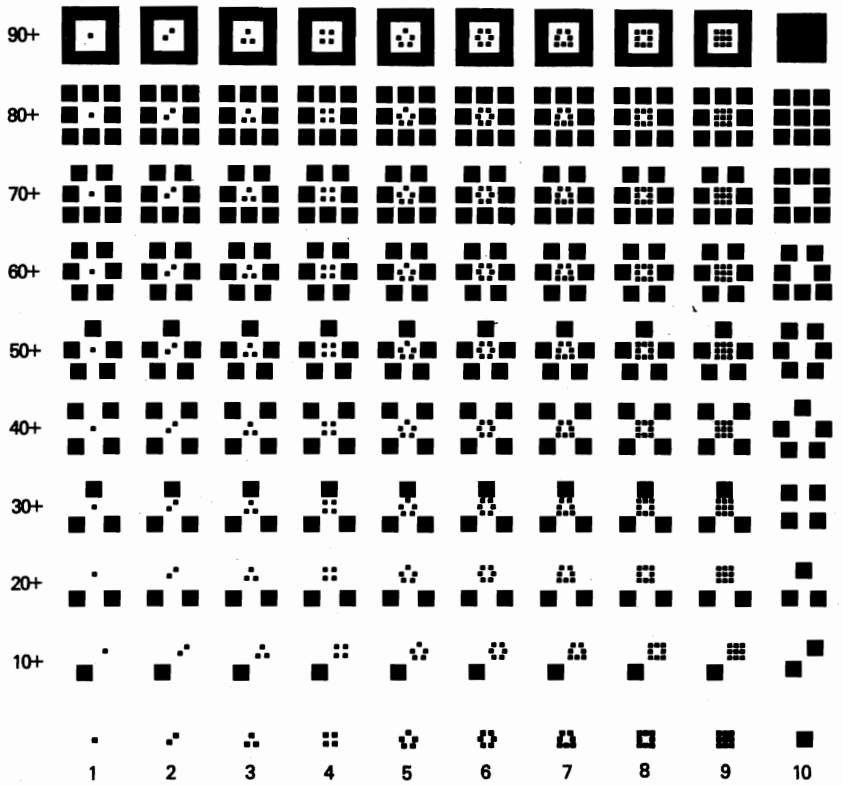
Figura 2.7 — Esempio di sequenza di diagrammi per una serie relativa a tre caratteri



Fonte: Bonin (1975)

Figura 2.8 — La scala dei Graphical Rational Patterns (GRP) da 1 a 100

FIGURE 2. SCALE OF SINGLE COMPUTERIZED GRP SHOWING INTEGERS FROM 1 TO 100

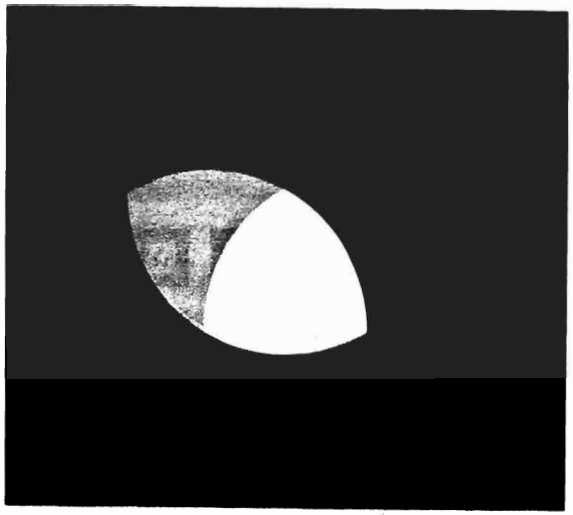


SOURCE: HEBREW UNIVERSITY, LABORATORY OF COMPUTER GRAPHICS, JERUSALEM, ISRAEL. METHOD BY ROBERTO BACHI.

Fonte: Bachi (1978)

Figura 2.9 — Sintesi additiva e sottrattiva dei colori

Tavola XVIII. Sintesi additiva delle luci, basata sui tre primari rosso, verde e blu.



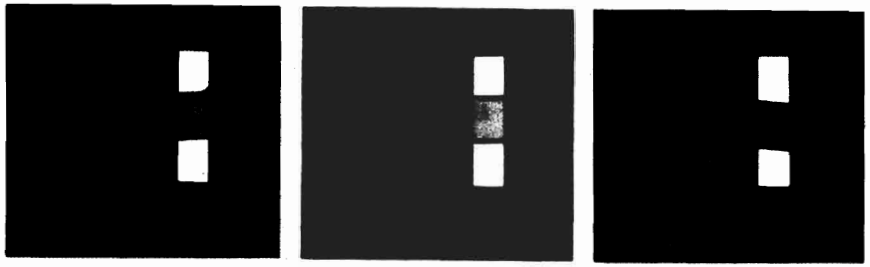
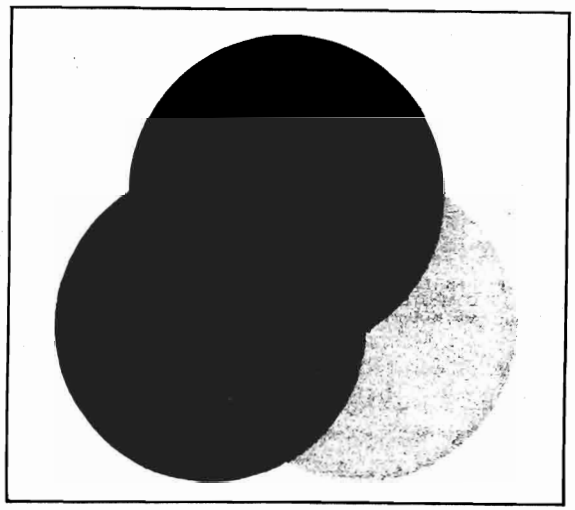
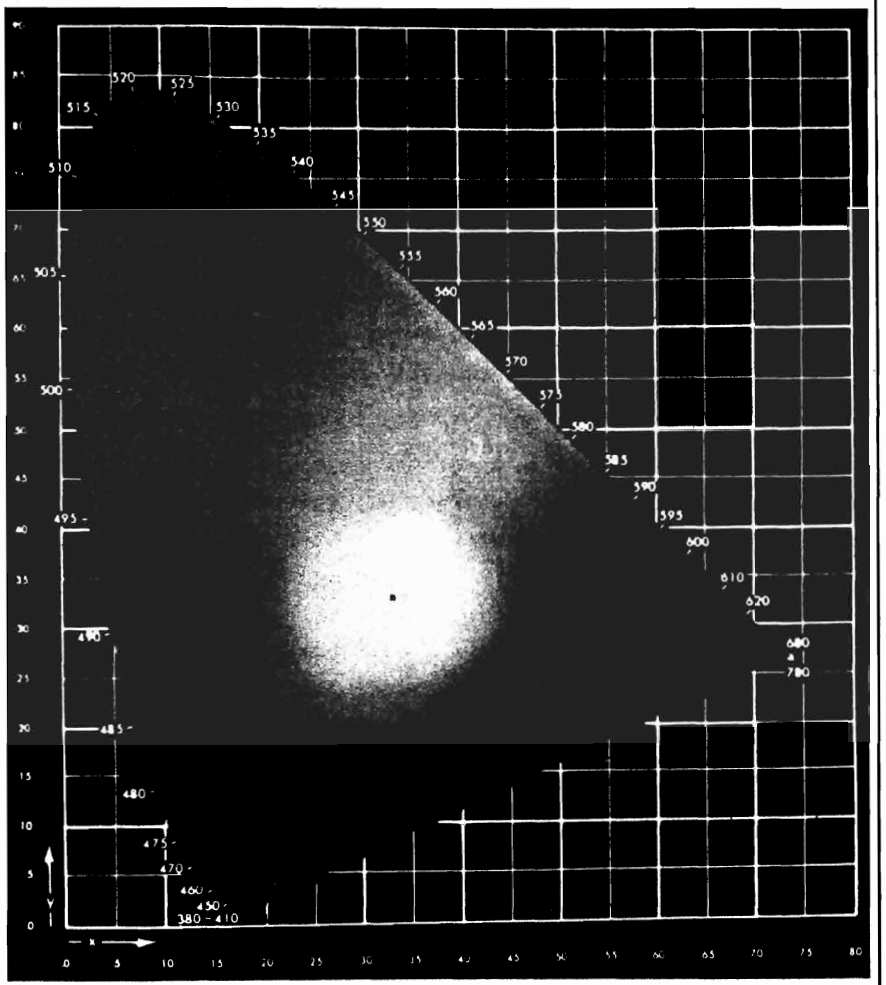


Tavola XIX. Sintesi sottrattiva dei filtri colorati, basata sui tre primari magenta, giallo e cian. Qui sopra, tre immagini ottenute con filtri trasparenti (foto Kodak).



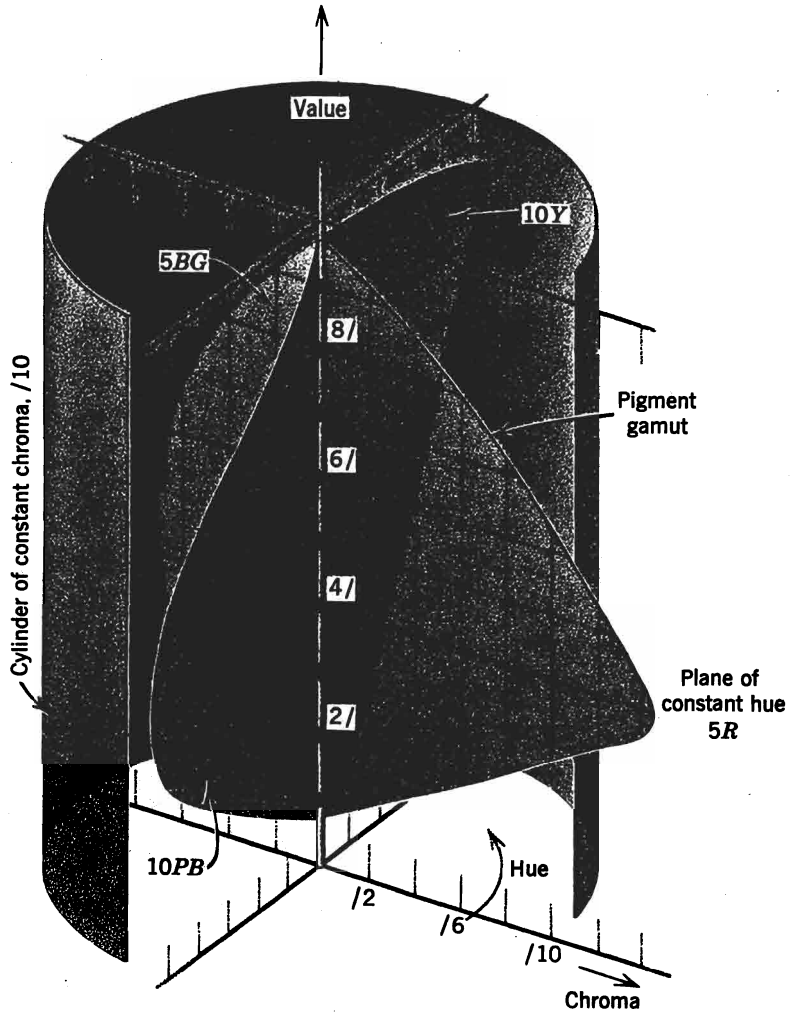
Fonte: Frova (1984)

Figura 2.10 — Diagramma di cromaticità CIE



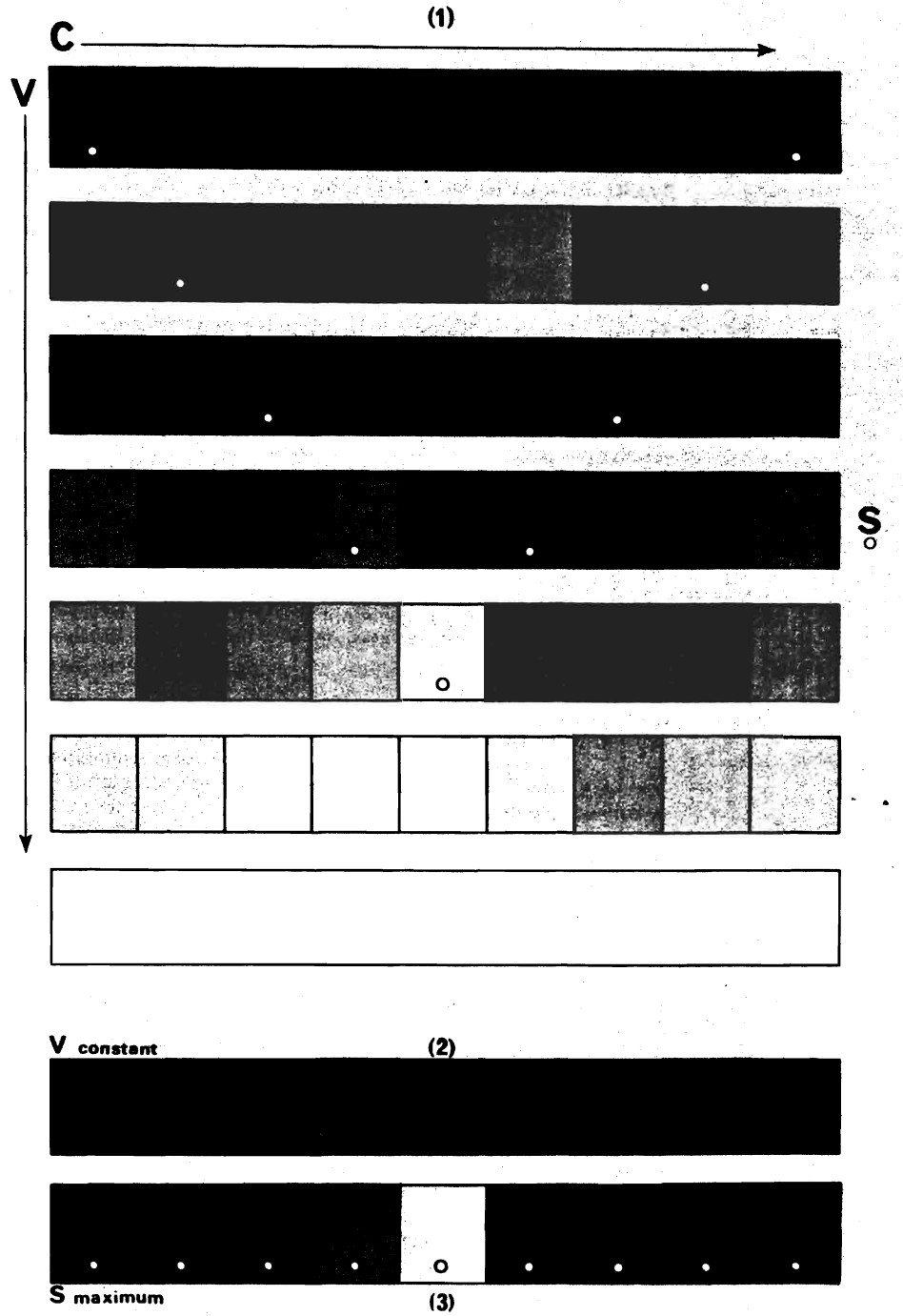
Fonte: Frova (1984)

Figura 2.11 — L'albero di Munsell



Fonte: Wyszecki, Stiles (1982)

Figura 2.12 — Scale cromatiche



Fonte: Bertin (1983)

CAPITOLO 3 - LE RAPPRESENTAZIONI GRAFICHE DELLE SERIE STATISTICHE NON TERRITORIALI

1. CONSIDERAZIONI GENERALI

Si vuole inizialmente rimarcare i vantaggi di attenersi a chiare e semplici regole nell'esecuzione dei grafici. Una razionale standardizzazione dei metodi grafici permette, infatti, di mantenere uniformità e consistenza nella esecuzione di grafici dello stesso tipo, anche se concepiti e destinati a pubblicazioni diverse e prodotti in tempi diversi, e di cumulare, nel corso del tempo, esperienze comuni in campo grafico. La standardizzazione, inoltre, facilita lo scambio di idee sistematico e la collaborazione fra funzionari appartenenti a settori differenti; evita le disfunzioni collegate allo sviluppo di processi non comunicanti fra loro e la dispersione di energie connesse con il ricercare soluzioni ad hoc per ogni progetto grafico.

In particolare questa ultima tendenza può portare a reinventare metodi già sperimentati ed a ignorare le ragioni per cui sono stati abbandonati.

Nei prossimi due capitoli si daranno delle indicazioni per una scelta razionale ed una efficace realizzazione di grafici statistici. Si tratterà in questo capitolo delle serie non territoriali, nel prossimo di quelle territoriali.

Allorquando il numero delle serie da rappresentare congiuntamente non è eccessivo, esistono metodi grafici che consentono una riproduzione dei fenomeni sufficientemente chiara; queste semplici tecniche non esauriscono certamente la tematica, che si presenta particolarmente viva e che ha visto, negli ultimi anni, varie proposte per la rappresentazione di dati multivariati.

I grafici che presentano la più alta frequenza di utilizzazione sembra siano i diagrammi cartesiani relativi alle serie storiche; nella ricerca di una efficace impostazione per queste rappresentazioni si è agevolati dagli standard proposti da parte di determinate istituzioni in alcuni Paesi, in particolare negli Stati Uniti d'America; oltre a ciò si può convenientemente tener conto delle riflessioni e delle proposte provenienti da studiosi quali Cox (1978) e Bachi (1975).

In riferimento alla classificazione presentata nel paragrafo 2 del capitolo 2, le serie storiche e le altre distribuzioni lineari di tipo quantitativo possono essere rappresentate su un piano cartesiano in un sistema di riferimento costituito da coordinate rettilinee, associate ad assi fra loro ortogonali.

Dal momento che l'evoluzione del tempo può pensarsi come un processo di tipo continuo, la presentazione delle serie storiche tramite linee continue, disegnate in un sistema di riferimento car-

tesiano, è una scelta, generalmente, naturale e razionale.

L'argomento richiede per le serie storiche relative a fenomeni di flusso alcune precisazioni che verranno date nel paragrafo seguente.

Quando il numero dei dati della serie è limitato questa può essere rappresentata tramite diagrammi a colonne in una successione finita di tempi distinti; ciò verrà indicato nel successivo paragrafo 4.

Nel paragrafo 3 si farà anche un breve cenno a quell'importante strumento grafico costituito dal diagramma semilogaritmico che è in grado, da una parte, di risolvere problemi di rappresentazione congiunta di serie, causati da escursioni molto elevate dei dati, e di fornire, dall'altra, ulteriori elementi di interpretazione dei dati.

Per le serie quantitative, di tipo continuo, le rappresentazioni sovente usate nel campo della statistica sono gli istogrammi; la loro costruzione richiede una preventiva suddivisione in classi del campo dei valori assunti dal fenomeno.

Per le serie quantitative di tipo discreto, come i voti agli esami universitari o il numero delle stanze in un appartamento, la rappresentazione più adatta è costituita da diagrammi ad aste (Leti 1983), o a colonne come suggerito nel paragrafo 4.

Per insiemi di dati costituiti da singole osservazioni quantitative riveste un certo interesse una rappresentazione che prescinda da una suddivisione in classi da fissare arbitrariamente; in tal caso i diagrammi stem-and-leaf e box-plot possono essere, per la loro flessibilità, dei validi strumenti grafici; di ciò discuteremo brevemente nel paragrafo 5.

Per quanto riguarda le serie di tipo qualitativo ordinabile e non un uso sobrio dei diagrammi a nastri e a colonne pare sia sufficiente a risolvere i principali problemi di rappresentazione di tali serie; di questo metodo grafico si discuterà nel paragrafo 4.

Nel paragrafo 7 si tratterà brevemente dei diagrammi areali, in particolare dei «diagrammi a torta» e dei diagrammi a barre suddivise; verranno sottolineati i problemi che essi pongono per una visione rapida e corretta dell'informazione e gli ambiti della loro utilizzazione.

2. LE RAPPRESENTAZIONI GRAFICHE TRAMITE LINEE CONTINUE IN UN SISTEMA DI COORDINATE CARTESIANE

Se entrambe le componenti dell'informazione sono quantitative, il modo naturale e più corretto di rappresentarle è tramite diagrammi riferiti ad un sistema di coordinate cartesiane del piano, associate ad assi fra loro ortogonali.

Tale sistema, come è noto, è individuato da due rette denominate x ed y, intersecantesi in un punto o, su ciascuna delle quali si fissa una direzione positiva, e da due segmenti u, v dei quali u è l'unità di misura dei segmenti sulla retta x, detta asse delle ascisse, e v è relativa alla retta y, detta asse delle ordinate. Le due rette sono dette assi cartesiani ed il punto o è chiamato origine degli assi. Se u = v il sistema è detto monometrico, altrimenti è chiamato dimetrico.

Per una razionale rappresentazione conviene fissare gli assi in modo che siano ortogonali fra loro.

Com'è noto, in questo sistema, è possibile associare in modo biunivoco a qualsiasi punto c del piano le sue coordinate cartesiane, date da una coppia di numeri reali (p,q), come è indicato nella figura 3.1a.

Conviene osservare che l'origine ha coordinate (0,0); i punti giacenti sull'asse y hanno ascisse uguali a zero; i punti giacenti sull'asse x hanno ordinata zero. Il sistema degli assi cartesiani, fra loro ortogonali, individua quattro quadranti retti, che vengono numerati in verso antiorario dal I al IV, a partire dal quadrante in cui entrambe le coordinate sono positive.

Avendo associato ai due assi le due componenti quantitative dell'informazione, questa si può perciò tradurre in un insieme di punti distribuiti sul piano, come è indicato nella figura 3.1b.

Nel caso delle serie storiche relative a fenomeni di stato, la componente temporale viene naturalmente associata all'asse delle ascisse, che viene perciò a rappresentare con il suo andamento rettilineo crescente l'evoluzione del tempo, mentre le quantità di stato vengono associate all'asse delle ordinate.

Dal momento che, generalmente, i valori posti in corrispondenza con il tempo sono quantità positive (frequenze od ammontari), i punti rappresentati sul piano ricadono, scegliendo opportunamente l'origine dei tempi, nel I quadrante.

Tali punti possono essere uniti per formare una linea continua, che ha la forma di una spezzata o poligonale, se i punti sono congiunti tramite segmenti, come appare nella figura 3.1c. Nel caso di serie storiche riferite a fenomeni di stato la linea che unisce i vari punti ha lo scopo di fornire al lettore una visione dell'andamento complessivo dei fenomeni, dando una stima, valida sotto semplici ipotesi, dei valori corrispondenti a istanti di tempo per cui non è stata effettuata la misura e che sono compresi fra istanti di tempo per cui il dato è conosciuto. La linea fornisce con la sua pendenza una misura semplicemente percepibile della rapidità di variazione del fenomeno fra due istanti successivi per cui il dato è noto.

Per la rappresentazione grafica delle serie storiche esistono degli standard proposti dal comitato Y15 della ANSI, del cui lavoro abbiamo già parlato nel paragrafo 1 del Capitolo 1; tali

standard sono stati ripresi da alcuni studiosi e, fra questi, dallo Schmidt nel suo testo del 1983. Riportiamo nella figura 3.2 una rappresentazione grafica «tipo» relativa alle serie storiche.

Il grafico è accompagnato da numerose didascalie, alcune evidenti e che non richiedono particolari commenti, altre che meritano di essere sottolineate.

Il titolo è stato posto in testa al grafico; una eventuale numerazione precederà a sinistra il titolo stesso; in basso vengono riportati possibili ulteriori delucidazioni e viene segnalata la fonte dei dati.

Nell'esempio di figura 3.2 manca la didascalia dell'asse delle ascisse in quanto il significato dei valori dell'asse è chiaramente autoesplicantesi; essa è invece riportata per l'asse delle y, con ulteriore chiarimento rispetto alle indicazioni fornite nel titolo.

Se è possibile, pare conveniente disporre le didascalie relative all'asse y orizzontalmente, come suggerito dai grafici 3.3a, nella stessa direzione del titolo, in quanto ciò favorisce la rapida lettura del grafico.

Su entrambi gli assi cartesiani sono riportati dei trattini, che segnalano particolari valori delle variabili rappresentate; essi sono a distanza uguale uno dall'altro e per le principali suddivisioni è riportato il valore numerico corrispondente. Conviene anche evitare di riportare sull'asse un numero eccessivo di cifre, che riuscirebbero di difficile lettura, come viene indicato nella figura 3.3b.

Per quanto riguarda le linee di riferimento interne al piano, esse partono da alcuni punti della suddivisione delle scale e vanno riportate in modo sobrio. Esse possono, in effetti, essere utili al lettore al fine di una precisa e rapida identificazione dei valori associati ai punti della curva.

È opportuno riportare un numero limitato di tali linee di riferimento e di renderle, ad un tempo, con un tratto leggero, rispetto a quello della curva che rappresenta il fenomeno, come è suggerito nella figura 3.3c.

Nella costruzione della scala relativa agli assi cartesiani, un aspetto importante è che, se la variabile ha come punto di partenza dei possibili valori uno zero dal significato intrinseco, ovvero se la variabile, come si usa dire in statistica, è di tipo rapporto, diventa necessario far partire i valori della scala dell'asse cartesiano, associato a tale variabile, dallo zero. Nelle applicazioni ci si trova sovente in casi del genere.

Ciò comporta l'inconveniente che, se la serie accentra i suoi valori in un intervento piccolo rispetto al livello di grandezza di questi, una scala rappresentata in modo completo risulta utilizzata solo in una sua piccola parte.

In questo caso appare opportuno interrompere la scala e limitarsi a rappresentare la zona d'interesse del fenomeno.È importante allora segnalare, come ad esempio viene fatto nella figura 3.2 per l'asse delle ordinate, l'avvenuta interruzione della scala. Ciò può essere indicato al lettore anche in altri modi, come quello suggerito nella figura 3.4, ripresa dall'Annuario Statistico Italiano 1988.

Per le variabili di tipo intervallo, lo zero ha un significato meramente convenzionale; in questo caso la scala può partire correttamente da qualsiasi valore opportunamente scelto.

Dal momento che gli assi cartesiani sono la base di partenza della numerazione della scala è opportuno che siano marcati in modo leggermente più pronunciato che non le altre linee di riferimento. Ancora più nettamente marcata rispetto ad entrambe deve essere la linea che rappresenta il fenomeno in studio. Nel caso che il numero dei punti della serie storica non sia eccessivo, sembra conveniente che questi siano rappresentati sul grafico, dal momento che costituiscono la base effettiva dell'informazione.

Quando si vuole rappresentare congiuntamente più serie storiche, esse vanno riprodotte graficamente usando per le rispettive linee simboli grafici diversi; ad esempio, come nella figura 3.2, si possono usare la linea continua e la linea a tratti; si possono immaginare altri metodi come la linea a punti, la linea a punti e tratti e così via. È certamente possibile utilizzare il colore per distinguere le diverse linee, anche al fine di rendere più valido esteticamente il grafico.

Se è di semplice realizzazione, conviene, come in figura 3.2, dare direttamente sul grafico le didascalie delle diverse curve. Nel caso in cui il grafico ne risultasse eccessivamente appesantito, si può pensare ad una legenda collocata in una zona libera del grafico o accanto ad esso.

La rappresentazione congiunta di più serie storiche su un medesimo sistema di riferimento è consigliabile fintanto che le curve rispettive non si intersechino e si sovrappongano in modo complicato tale da rendere difficoltosa la lettura del grafico.

Tale tipo di costruzione è possibile, va sottolineato, fintanto che le grandezze da rappresentare sono espresse nella stessa unità di misura, oppure in numeri adimensionali, come avviene per i numeri indici e nei diagrammi semilogaritmici. Se non è questo il caso è in generale sconsigliabile l'uso di due scale distinte sullo stesso diagramma e converrà ricorrere all'uso di diagrammi separati.

In questi casi, e comunque quando il numero delle serie è eccessivo, conviene rivolgersi ad una serie di diagrammi cartesiani con le ordinate spostate, posti parallelamente l'uno all'altro e con la scala dei tempi in corrispondenza; non è certo necessario riportare ogni volta l'asse dei tempi, anche se va chiaramente segnalato il valore di partenza della scala, o delle scale, se si tratta

di grandezze diverse.

Sarebbe importante che lo studioso, nel caso risultasse semplicemente realizzabile, ordinasse le diverse curve in tipi di andamento e che, quindi, le serie storiche venissero rappresentate secondo questi raggruppamenti, basati su una somiglianza di forma empiricamente valutata. Questa classificazione, realizzata su base empirica, acquista una rilevanza ed un interesse maggiore nel caso dei diagrammi semilogaritmici, nei quali è la pendenza delle curve e quindi la forma di queste a costituire l'informazione rilevante.

In proposito è importante sottolineare che lo scopo dei diagrammi cartesiani multipli, con e senza spostamenti delle ordinate, è di consentire confronti fra gli andamenti delle varie curve. Essi non consentono, invece, una valutazione pienamente soddisfacente delle differenze dei valori alle diverse ascisse; già C. Gini in una nota presentata alla XXI sessione dell'ISI nel 1933 espose sull'argomento precise osservazioni. Considerazioni analoghe sono state svolte in un recente articolo di Cleveland e McGill (1984), da cui riproduciamo alcuni grafici, connessi all'argomento in questione, nella figura 3.5.

In essa vengono rappresentate due serie di grafici, di cui la prima rappresenta coppie di curve e l'altra i grafici della loro differenza.

Nel caso che si ritenga importante dare al lettore l'informazione relativa alla differenza fra le due curve, conviene allora aggiungere il grafico che la rappresenta.

Riportiamo nella figura 3.6 un esempio di rappresentazione congiunta dell'indice generale della produzione industriale e di quelli per destinazione economica, entrambi destagionalizzati con il metodo X11-ARIMA [8].

Prima di passare alle considerazioni relative alle serie storiche relative a dati di flusso o di movimento, conviene puntualizzare un aspetto molto dibattuto e da lungo tempo presente nelle discussioni che vertono sulle rappresentazioni grafiche e che, a ragione di ciò, sono da tener presente nella scelta della struttura del grafico. Si tratta del problema, ben noto, legato alla scelta delle unità di misura degli assi cartesiani; variando tali unità di misura si può modificare fortemente l'impressione visiva suscitata dal grafico e quindi dal fenomeno rappresentato.

(8) A tal proposito si fa presente che in Istat è stato adottato per la destagionalizzazione delle serie storiche il metodo X11-ARIMA, realizzato dal Bureau of Census degli U.S.A. Esso è l'ultima versione del Census Method II, una delle più collaudate versioni dei metodi di destagionalizzazione basati sul *metodo del rapporto alla media mobile;* esso, per quanto privo di solide basi statistico-matematiche, ha una validità alternativa rispetto ai metodi lineari generalizzati, per la quantità di verifiche sperimentali a cui è stato sottoposto e per la qualità dei risultati raggiunti (Shiskin, Young, Musgrave 1965).

Ciò appare evidente dalla figura 3.7 dove si fanno variare, da grafico a grafico, congiuntamente le unità di misura di entrambi gli assi cartesiani. Si può dire che non esista, in effetti, una proporzione ottimale per le due unità di misura, anche se considerazioni fatte in proposito da Cox (1978) e Bachi (1975) sono da valutare attentamente.

Alcuni suggerimenti sono forniti in proposito anche nel testo redatto dallo Y15 Commitee sulle serie storiche ripresi dal testo dello Schmid (1983).

Sovente problemi di standardizzazione e di uniformità tendono a prevalere su considerazioni più specificamente analitiche. A tal fine, si potrebbe arrivare alla definizione, di alcune scale standard, relativamente all'asse dei tempi, concepite in funzione della lunghezza della serie, del tipo di periodicità del dato, del formato delle pubblicazioni a cui sono destinate e tenendo conto della capacità che l'occhio ha di distinguere due punti vicini.

Per quanto riguarda l'altro asse, che generalmente sarà l'asse y, è innanzitutto possibile trasformare i dati nei modi seguenti, in funzione delle loro caratteristiche e degli scopi della rappresentazione: a) si danno i valori effettivi della serie; b) si presentano i valori divisi per la media o per il dato di un certo tempo; c) si rappresentano i dati standardizzati. In ogni caso conviene riprodurre sul grafico per il lettore i valori effettivi, che risultano immediatamente comprensibili.

Rispetto alla definizione del valore da attribuire ad un intervallo standard dell'asse delle y, la scelta dipende dallo scopo della rappresentazione. Si può voler evidenziare le caratteristiche dell'andamento di una singola serie, oppure, al contrario, consentire la confrontabilità fra serie diverse.

Nel primo caso, nella scelta della scala, si terrà conto dei valori della singola serie. Nell'altro caso si considereranno congiuntamente i valori del gruppo di serie; si potrà quindi mantenere costante tale scala finché le caratteristiche assunte dai fenomeni non costringano a rivedere la struttura delle rappresentazioni.

Più semplicemente e schematicamente, al posto di una standardizzazione vincolante, si può tendere ad ampliare l'unità di misura relativa all'asse delle ordinate quando: a) il numero dei punti della serie è notevolmente ridotto; b) la serie presenta un forte trend o dei rapidi cambiamenti; c) si vogliono rappresentare insieme delle serie che con una scala della y ridotta verrebbero a confondersi; d) si vogliono evidenziare fluttuazioni che altrimenti rimarrebbero inosservate.

Riportiamo nella figura 3.8, ripresa dal testo di Schmid (1983), alcune scale temporali standard concepite all'interno di una istituzione degli U.S.A..

Un ulteriore aspetto, che una standardizzazione esauriente dovrebbe tenere in conto, è l'esistenza di periodizzazioni tempo-

rali differenti in funzione dei diversi fenomeni sociali e/o naturali.

Nel caso delle serie storiche relative a fenomeni di movimento o di flusso, il dato a disposizione dello studioso si riferisce ad un intero periodo di tempo e non più ad un preciso istante nel tempo; ad esempio il prodotto interno lordo relativo ad un anno, gli investimenti effettuati in un anno da una impresa, il numero dei nati in un mese sono quantità di flusso.

Supponendo che il dato, quella determinata grandezza riferita ad un periodo definito di tempo, si modifichi ad un tasso costante in quell'arco temporale, la rappresentanzione più logica cui pensare è l'istogramma.

In corrispondenza a ciascun intervallo temporale si costruisce un rettangolo avente per base il segmento i cui estremi sono le ascisse corrispondenti agli istanti iniziale e finale di tale intervallo e per altezza il rapporto fra il dato di flusso e l'ampiezza dell'intervallo temporale. Il valore dell'ordinata diventa, in questo modo, una misura della rapidità media con cui si genera in quel periodo la grandezza di flusso in questione; essa è anche interpretabile come la quantità della variabile di flusso che si è prodotta in media nell'unità di tempo.

Se gli intervalli temporali di riferimento sono uguali tra loro, allora l'altezza dei rettangoli è proporzionale al dato assoluto iniziale; in questo caso marcando la parte superiore della superficie limitata dai rettangoli si avrebbe come risultato grafico un diagramma a scalini simile a quello della figura 3.9, che potrebbe rappresentare una ipotetica serie storica mensile di dati di flusso.

Si usa, tuttavia, nella pratica corrente, sostituire a tale rappresentazione, che è la più vicina alle caratteristiche del dato, la spezzata o poligonale che unisce i punti intermedi a ciascun intervallo temporale. Ciò dà (Leti 1983) la possibilità di rappresentare congiuntamente in modo semplice più serie storiche; tale rappresentazione potrebbe essere raffinata sulla base delle considerazioni svolte nel successivo paragrafo 4.

In tal caso bisogna comunque associare con chiarezza il dato al periodo di tempo di riferimento ed evitare di rappresentare insieme sul medesimo diagramma serie di stato e di flusso, come ricorda Castellano (1988).

3. DIAGRAMMI SEMILOGARITMICI

Quando si è trattato nel precedente paragrafo 2 dei diagrammi cartesiani, in particolare dedicati alla rappresentazione delle serie storiche, il piano è stato semplicemente caratterizzato dai due

assi con le rispettive scale di tipo lineare; se un punto B, su uno dei due assi, ha una distanza doppia dall'origine rispetto ad un punto A, la grandezza corrispondente a B ha un valore doppio di quello corrispondente ad A; ciò pare naturale in quanto l'abitudine a trattare relazioni di tipo lineare è la più radicata ed ampiamente diffusa.

In numerosi fenomeni di natura demografica, economica ed ambientale, tuttavia, l'uso di una scala lineare può far perdere aspetti essenziali della loro evoluzione nel tempo, specialmente quando questi sono caratterizzati essenzialmente da variazioni relative da un tempo all'altro, piuttosto che da variazioni assolute.

La variazione, ad esempio, del prodotto nazionale lordo o del livello dei prezzi è misurato da indici appropriati in termini percentuali rispetto al livello raggiunto dal fenomeno; ciò accade spesso anche nello studio dell'evoluzione di popolazioni e dell'andamento di alcuni fenomeni ambientali.

In questi casi capita che, se si rappresenta la serie storica relativa ad un lungo periodo di tempo, il diagramma cartesiano con la scala delle ordinate di tipo lineare, mentre rende particolarmente evidenti le variazioni assolute relative ai periodi in cui il livello del fenomeno è elevato, appiattisce notevolmente le variazioni assolute dei periodi, generalmente quelli iniziali, in cui tale livello è basso.

Una osservazione simile si può fare nel caso che si voglia confrontare fra loro due fenomeni che evolvono da livelli notevolmente diversi ma con tassi di variazione relativa confrontabili. Se, ad esempio, si vuole confrontare la popolazione residente italiana e quella della città di Roma, ai censimenti dal 1871 al 1981, la rappresentazione congiunta delle due serie storiche in un diagramma con scale lineari vedrebbe i dati relativi a Roma appiattirsi nella parte bassa del grafico.

Si può dire che tutte le volte che interessa mettere in evidenza le variazioni relative da un tempo all'altro di un certo fenomeno o metterle a confronto per due serie simultanee, conviene rivolgerci a diagrammi semilogaritmici.

In questi, mentre l'asse delle ascisse mantiene la sua struttura lineare ad indicare i tempi, l'asse delle ordinate rappresenta la trasformata logaritmica dei dati; si fa ovviamento l'ipotesi che questi assumono valori positivi.

Generalmente, nella pratica usuale in campo statistico, si usano i logaritmi in base 10, indicati con il simbolo Log. In effetti la scelta della scala è dettata dal problema che si ha di fronte; nella teoria dell'informazione e nella informatica è di uso comune il logaritmo in base 2, come conseguenza dell'uso della codifica binaria dei dati, particolarmente adatta ai componenti elettronici; in studi di tipo scientifico e teorico viene naturale l'uso dei loga-

ritmi in base "e", la costante di Nepero, o logaritmi naturali, come conseguenza dell'apparire di tale costante in alcune espressioni teoriche.

Tuttavia, se lo scopo è la presentazione di dati al pubblico, può essere conveniente limitarsi all'uso dei logaritmi in base decimale, che sono quelli di più comune applicazione.

Riportiamo nella figura 3.10 una scala logaritmica, relativa a valori monetari; su di essa sono riportati i valori effettivi assunti dalla variabile non trasformata, al fine di consentire al lettore una valutazione delle quantità assolute. Dalla figura risulta che l'intervallo dei valori assoluti contenuti in un dato segmento dell'asse cresce con l'allontanarsi di detto segmento verso la parte alta della scala; raddoppiare una distanza, in tale scala, equivale in termini numerici ad una moltiplicazione per 10. Il segmento I indicato nella figura 3.10 ed immediatamente superiore al valore 1 contiene 9 unità nella scala lineare, mentre il segmento II immediatamente superiore, pur avendo la medesima lunghezza, contiene 90 unità della medesima scala lineare. È come se i valori assoluti subissero una contrazione progressiva sempre più marcata della loro scala in funzione della loro crescita.

Prima di riportare alcuni significativi esempi ed applicazioni, si vuole dare spunti ulteriori per semplificare ulteriormente la lettura di tale diagrammi.

Si può fare riferimento a due ipotetiche serie storiche, (A) e (B), relative ad istanti di tempo posti ad una distanza costante gli uni dagli altri, che si assume come unità di misura della scala dell'asse dei tempi. La prima, (A), assume i valori di una progressione aritmetica di ragione "a", che supponiamo sia un valore reale positivo, e che, perciò, vale all'istante k la quantità ak (k = 1, 2, 3...); la seconda serie (B) assume i valori di una progressione geometrica di ragione anch'essa pari ad a e che vale all'istante generico k la quantità a^k.

Nel caso di un diagramma cartesiano con scala lineare, i punti della serie (A) si dispongono lungo una retta di pendenza positiva a, mentre quelli della serie (B) si dispongono lungo la curva esponenziale a^k, il cui andamento dipende dal valore di a; se $a > 1$ i valori della serie crescono rapidamente verso grandezze molto elevate, se $a < 1$ tali valori tendono rapidamente a zero.

In un diagramma semilogaritmico la serie (B) risulterà semplicemente rappresentata da una retta, infatti $\log a^k = k \log a$.

Quindi i punti della serie (B) si dispongono, in un diagramma semilogaritmico, lungo una retta la cui pendenza è log a. Se $a > 1$, $\log a > 0$ e la retta avrà pendenza positiva; se $0 < a < 1$, $\log a < 0$ e la retta avrà pendenza negativa.

La figura 3.11 fornisce un esempio di applicazione di tale diagramma a una sequenza di 3 numeri in progressione geometrica.

È immediato verificare che una serie geometrica cresce, o decresce, ad un tasso costante dato dalla ragione che la caratterizza; nel nostro esempio teorico essa è pari ad a; il logaritmo di tale quantità equivale alla pendenza uniforme della retta che rappresenta tale serie in un diagramma semilogaritmico.

Nelle serie reali, laddove è ipotizzabile un meccanismo di crescita legato a variazioni determinate dal livello già raggiunto dal fenomeno, ben difficilmente il tasso di crescita risulta costante nel tempo. In questo caso l'andamento dei dati sul diagramma semilogaritmico può assumere forme diverse. Per facilitare l'utente nella lettura delle curve che possono risultarne si può fare riferimento alla figura 3.12, dove vengono riportati alcuni andamenti tipici di curve relative a serie storiche decrescenti e crescenti con variazione dei tassi di tipo, a sua volta, crescente, decrescente o costante.

Sempre nell'ottica di facilitare il lettore nella corretta interpretazione dei dati, si potrebbe pensare ad associare al diagramma semilogaritmico, una indicazione sul valore dei tassi di crescita associati alle varie pendenze; nella figura 3.13 è riportato un esempio ripreso da Bertin (1983) e relativo alle serie storiche della quantità di ettari destinati a vari usi del suolo in Svezia nel periodo dal 1937 al 1959.

Per la definizione dell'intervallo della scala logaritmica, questo può essere scelto in funzione dei valori assunti dal fenomeno; dal momento che l'aspetto informativo essenziale dei diagrammi semilogaritmici è la pendenza della curva e che il logaritmo dello zero non esiste, l'inizio della scala può essere scelto in funzione dei dati senza preoccupazione per la segnalazione di interruzioni.

Conviene riportare, in genere, come nella figura 3.10, sulla scala logaritmica i valori originali dei dati piuttosto che i valori logaritmici, per consentire al lettore una valutazione anche delle quantità assolute.

Riportiamo nelle figure 3.14 e 3.15 alcuni esempi di diagrammi semilogaritmici relativi a serie economiche e ripresi da Samuelson, Nordhaus (1987) e Lindbeck (1980); su alcuni di essi, visto lo scopo didattico per cui sono stati concepiti, sono riportate delle didascalie relative ad ampie periodizzazioni, utili nell'interpretazione del grafico. Nella figura 3.16 si riportano due esempi costruiti da Schmid (1983) e relativi a fenomeni demografici.

Nella figura 3.17 sono riportati due diagrammi semilogaritmici relativi alla popolazione in alcuni maggiori Comuni italiani dal 1871 al 1986; nella figura 3.18 si trovano alcuni diagrammi semilogaritmici relativi alle spese per diversi consumi; entrambe le figure sono state riprese dall'Annuario Statistico Italiano 1988.

4. ISTOGRAMMI - DIAGRAMMI A BARRE *(NASTRI E COLONNE)*

Se la distribuzione da rappresentare graficamente è relativa ad un carattere quantitativo di tipo continuo e se i valori si presentano suddivisi in classi, il diagramma più adeguato è l'istogramma; esso può essere pensato come un insieme di rettangoli allineati, ognuno dei quali corrisponde ad una determinata classe.

Il grafico dell'istogramma viene disegnato in un sistema di coordinate cartesiane in cui l'asse delle ascisse rappresenta, generalmente, le modalità del carattere e l'asse delle ordinate l'intensità del fenomeno. Tecnicamente la costruzione dell'istogramma è piuttosto semplice; una volta che le modalità del carattere sono state raggruppate in classi, a queste si associano dei segmenti sull'asse delle x aventi per estremi le ascisse corrispondenti ai valori iniziali e finali delle medesime; gli intervalli così costruiti possono essere aperti o chiusi, a destra o sinistra, secondo la corrispondente definizione della classe. La costruzione dell'istogramma si basa, quindi, sulla convenzione che l'area del rettangolo, avente per base il segmento associato ad una data classe, deve essere proporzionale alla frequenza o ammontare della classe medesima. Per realizzare ciò in modo consistente è sufficiente prendere come altezza del rettangolo il rapporto fra la frequenza od ammontare, relativo a quella classe, e l'ampiezza dell'intervallo della medesima.

In questo modo la quantità che viene ad essere rappresentata sull'asse delle ordinate diventa una densità, ovvero una frequenza od ammontare per unità di ampiezza della classe.

Si può certamente sostituire l'andamento a gradini che assume la funzione densità, in conseguenza della suddivisione in classi, con una funzione continua delle modalità del carattere.

A tal proposito è interessante riprendere un esempio proposto da Bachi in un lavoro del 1978.

In modo simile a quanto si usa fare per le serie storiche di flusso, si sostituisce generalmente all'istogramma la poligonale che unisce i punti intermedi delle estremità superiori di ciascun rettangolo.

Nella figura 3.19 è riportato il grafico di una distribuzione di frequenze fittizia, relativa a certe classi di reddito espresse in dollari. Le difficoltà concettuali di tale tipo di rappresentazione si possono rilevare in due osservazioni:

a) la poligonale attribuisce densità non nulla a redditi in corrispondenza dei quali il fenomeno è risultato assente;

b) l'area sottostante la spezzata corrispondente ad una certa classe di reddito non dà il reddito complessivo di quella classe.

Utilizzando dei criteri scientifici, come il metodo delle aree (Salvemini, Girone 1981), si può vincolare la curva continua che approssima l'istogramma a riprodurre, quando essa viene integrata sugli intervalli che definiscono le classi, i valori dei redditi complessivi corrispondenti.

Nella figura 3.20 è riportato l'esempio, realizzato dallo stesso Bachi, che supera le osservazioni critiche fatte al grafico 3.19.

Considerazioni simili possono essere estese alle rappresentazioni di serie storiche relative a dati di flusso; la costruzione di una curva continua potrebbe essere basata sulle due ipotesi principali: a) la rapidità con cui una data grandezza di flusso viene generata è una funzione continua del tempo; b) per ciascun intervallo temporale di riferimento iniziale la quantità nota complessiva della grandezza di flusso viene esattamente riprodotta dalla curva per integrazione.

Si vuol dedicare in questo paragrafo un po' di spazio alla presentazione delle piramidi delle età. Queste consistono in effetti di una coppia di istogrammi contrapposti, con le due scale delle classi di età disposte verticalmente e in corrispondenza fra loro. Come si è già detto in precedenza, in tale rappresentazione l'età costituisce la componente quantitativa di prima specie mentre il sesso costituisce la componente qualitativa di seconda specie.

Si vuole sottolineare che, in tale tipo di diagramma, l'aspetto essenziale dell'informazione è dato dalla forma della piramide. Presentiamo nella figura 3.21 una serie di piramidi dell'età riprese da Bonin (1975) e rese in modo stilizzato e senza tutte le necessarie informazioni per la loro accurata lettura. Tuttavia la sequenza dei grafici dà la possibilità di formarsi un'idea immediata della struttura per classi di età dei paesi scandinavi. Nella figura 3.21 le popolazioni sono state ricondotte ad un totale pari a 100 e la scala delle ascisse riporta le percentuali di popolazione ed è uguale per tutte le piramidi. Ciò consente di concentrare l'attenzione sulla forma delle distribuzioni, senza che questa sia distratta dalla diversa consistenza delle popolazioni, che potrebbe essere data a parte tramite un diagramma a nastri. Un grafico simile si potrebbe realizzare per le 20 regioni italiane, una volta che le rispettive popolazioni siano state rapportate ad un medesimo valore; i diversi ammontari potrebbero essere rappresentati con un diagramma a nastri.

Conviene aggiungere che, se una piramide dell'età viene presentata isolatamente, essa può avere una scala anche di tipo assoluto.

Nella figura 3.21 è rappresentata anche la piramide delle età relativa alla popolazione in complesso dei Paesi scandinavi. Nel grafico 3.22a vengono rappresentate le differenze, per ciascun Paese, fra le frequenze di ciascuna classe di età e quelle della classe di età corrispondente della popolazione in complesso;

esso consente di percepire con immediatezza come la struttura per classi di età dei singoli Paesi si discosta dalla struttura media per classi di età della popolazione scandinava.

Nel grafico 3.22b il confronto è invece realizzato fra le distribuzioni dei due sessi nell'ambito dello stesso Paese; in corrispondenza di ciascuna classe di età viene rappresentata la frequenza in eccedenza di un sesso sull'altro; il grafico consente una valutazione semplice dei valori della differenza fra le frequenze dei due sessi per ciascuna classe di età, valutazione che non è agevole sulla base della piramide delle età.

La rappresentazione di ulteriori modalità su una medesima piramide delle età, come quelle relative allo stato civile, rende difficili i confronti fra le frequenze delle varie modalità e di conseguenza piuttosto problematica la percezione delle relative piramidi delle età. Come viene suggerito nella figura 3.23 (Bonin 1975) conviene, in questi casi, pensare alla rappresentazione di più piramidi delle età, ognuna relativa ad una modalità del carattere; in questo caso può essere opportuno rispettare nella rappresentazione le proporzioni reali delle varie sub-popolazioni.

La difficoltà a percepire l'esatta forma della distribuzione cresce ovviamente col crescere delle modalità da rappresentare ed è direttamente legata ai limiti della vista dell'uomo a fare confronti fra segmenti o barre che partono da livelli differenti.

Nel caso di distribuzioni di frequenze od ammontari relativi ad un carattere quantitativo di tipo discreto, la rappresentazione più adeguata è quella che rende graficamente l'intensità delle quantità con dei segmenti perpendicolari all'asse delle ascisse ed applicate ai punti corrispondenti alle modalità discrete del carattere; tale grafico viene anche chiamato diagramma ad aste (Leti 1983). Dal momento che i valori compresi fra due interi successivi non hanno significato, non ha senso unire l'estremità superiori dei segmenti con una linea continua.

Prima di trattare le rappresentazioni di serie relative a caratteri qualitativi tramite i diagrammi a barre, si vuole sottolineare come questo tipo di grafico nella sua variante a colonne si presti a rappresentare in modo adeguato una serie temporale, nel caso che questa sia costituita da pochi valori.

Nel diagramma a colonne corrispondente, l'associare all'asse delle ascisse una struttura metrica o semplicemente ordinale è una scelta da farsi in funzione degli scopi della rappresentazione. Come possibile esempio tipo per la realizzazione di un diagramma a colonne si può considerare il grafico riportato nella figura 3.27a e ripreso da Schmid (1983).

Quando si passa a rappresentare le distribuzioni relative a caratteri qualitativi, avendo come base la struttura del piano propria dei diagrammi cartesiani, come è stata descritta nel precedente paragrafo 2, si può constatare che la struttura metrica

dell'asse associato al carattere viene a perdersi; l'altro asse, associato alla componente costituita delle frequenze, ammontari o quantità derivate, conserva, ovviamente, una scala di tipo quantitativo, con una origine ed unità di misura. I possibili tipi di rappresentazione in questo caso sono i cosiddetti diagrammi a nastri o a colonne, dove ciascun nastro o colonna è in corrispondenza con una modalità del carattere e dove l'intensità del fenomeno relativa a ciascuna modalità è rappresentata dalle diverse lunghezze delle barre, disposte orizzontalmente o verticalmente, in riferimento ad una ben definita scala; come esempio riportiamo i grafici 3.24a e 3.25a.

Nel caso che il carattere sia qualitativo ordinabile, si possono rappresentare le modalità disponendole nella corretta successione lungo l'asse delle x e distanziando due modalità successive in modo uguale; ciò è una semplice convenzione di comodo e va sottolineato che non riflette alcuna struttura metrica sottostante alle modalità ordinate del carattere.

Come standard di rappresentazione anche in questo caso si può far riferimento alla figura 3.27a. Sul grafico sono riportate alcune didascalie il cui significato è abbastanza evidente; sono da sottolineare solo alcuni punti, importanti al fine di realizzare un diagramma chiaro e semplice da leggere, che valgono sia per i diagrammi a colonne che per i diagrammi a nastri, per i quali è dato, nella figura 3.27b, un esempio ripreso dal Pedroni (1968).

Per quanto riguarda il disegno della scala conviene evitare i troncamenti di questa nella parte inferiore e segnalare con evidenza lo zero come base di partenza per la valutazione dei valori di frequenze o di ammontari; vanno evitati, in questo caso, anche le interruzioni della scala che diventerebbero evidenti interruzioni delle barre.

Nel caso che nella serie vi siano valori particolarmente elevati rispetto al resto dei dati, conviene calcolare la scala escludendo tali valori eccezionali; nel caso che questi non risultino rappresentabili da barre compiutamente disegnate, queste possono essere interrotte nella loro parte superiore, segnalando accanto il valore numerico corrispondente; il numero di tali troncamenti dovrebbe essere, comunque, limitato.

Possiamo riprendere per i diagrammi a barre alcune considerazioni già fatte a proposito dei diagrammi a linee continue, come, ad esempio, che i valori della scala da riportare sul grafico dovrebbero essere in linea di massima multipli interi dell'unità o del cinque; che è opportuno riportare con tratto leggero alcune linee di riferimento, al fine di facilitare la valutazione delle quantità rappresentate e che, comunque, queste non devono interferire con il disegno dei nastri o delle colonne.

Per campire le barre conviene generalmente usare un colore scuro distribuito in modo uniforme; l'uso del colore può essere adatto per rappresentare quantità di segno diverso o per indicare modalità diverse anche di uno stesso carattere. Esempi di campiture sono riportati nella figura 3.26.

Se si vuole usare il tratteggio, conviene che questo sia semplice e che ricopra il più uniformemente possibile l'area delle barre.

Si vuole aggiungere che la separazione delle barre l'una dall'altra non è strettamente necessaria; colonne contigue come nella figura 3.24b possono consentire una valutazione dell'andamento complessivo più efficace, se questo ha un senso, anche se a scapito del confronto fra le singole barre vicine.

Per quanto riguarda i diagrammi a nastri si può aggiungere che la disposizione orizzontale delle barre può consentire una più semplice ed efficace descrizione delle modalità del carattere qualitativo. Nel caso che su una pagina appaiono numerosi grafici relativi allo stesso carattere, può essere conveniente associare a ciascuna modalità un simbolo alfanumerico e riportare la simbologia in una legenda.

Il diagramma a nastri si presta facilmente alla rappresentazione congiunta. Le singole rappresentazioni delle due distribuzioni possono essere contrapposte come nella figura 3.25b.

Nel caso che le misure da confrontare siano numerose o che il carattere di seconda specie possegga numerose modalità, i differenti diagrammi a barre possono essere disposti parallelamente l'uno all'altro. Un esempio è stato fornito dalle selezioni dei grafici prodotti in Istat nella figura 1.1h. Lo scopo di questi grafici è consentire al lettore una visione complessiva di come variano le forme della distribuzione relativa al carattere di prima specie, al variare delle modalità del carattere di seconda specie. Si pone allo studioso il problema se convenga o meno riportare il valore complessivo di ciascuna popolazione ad un medesimo ammontare.

Per consentire al lettore valutazioni delle differenze relative alle quantità corrispondenti alla stessa modalità del carattere di prima specie per due diverse modalità del carattere di seconda specie, la cosa più opportuna è disegnare direttamente tali differenze tramite grafici simili a quelli presentati nelle figure 3.24c e 3.25c. Si deve ovviamente tener conto nella costruzione della scala della possibile presenza di valori sia positivi che negativi; si può anche pensare all'uso di colori differenti per i valori di diverso segno.

Si vuole terminare questo paragrafo con due ulteriori considerazioni. Talvolta può essere opportuno ordinare le modalità di un carattere qualitativo non ordinabile in modo monotono rispetto ai valori dell'intensità associate, dal momento che ciò facilita i confronti.

Ciò chiaramente non è possibile per caratteri qualitativi ordinabili, dal momento che in questo caso l'ordinamento delle modalità costituisce parte integrante dell'informazione.

In entrambi i casi non ha alcun senso unire con una linea continua le estremità delle barre.

5. ALCUNI STRUMENTI GRAFICI PER LE OSSERVAZIONI INDIVIDUALI: STEM-AND-LEAF E BOX-PLOT

In questo paragrafo si vuole presentare alcuni strumenti grafici che hanno rilevanza, di preferenza, nell'ambito della analisi dei dati e che soprattutto risultano essere utili allo studioso nel suo lavoro di ricerca empirica.

Tuttavia, nella misura in cui il grafico riesce, attraverso un uso appropriato, ad evidenziare aspetti essenziali dell'informazione, può essere proficuo utilizzare tali metodi grafici anche a fine di presentazione. Le lievi difficoltà presenti nella corretta lettura di questi diagrammi possono essere risolte con una semplice introduzione tecnica.

Si darà una breve descrizione dello stem-and-leaf ed il box-plot, recenti ideazioni di Tukey (1977), che li ha proposti come strumenti grafici nell'ambito di un'analisi di tipo iniziale in cui non si fa direttamente intervenire ipotesi di tipo probabilistico e si cerca di caratterizzare in modo semplice l'insieme dei dati.

A) Stem-and-leaf

Per poter rappresentare una serie di dati quantitativi tramite l'istogramma è necessario suddividere l'insieme dei valori della variabile in classi; questo avviene spesso sulla base di scelte aprioristiche o convenzionali con una perdita d'informazione.

Se il nostro gruppo di dati non è eccessivamente numeroso, approssimativamente non superiore a 300-400 numeri, può essere vantaggioso, per costruire una visione d'insieme, usare la rappresentazione grafica «stem-and-leaf» come proposta dal Tukey (1977) e presentata da Lombardo (1984). Tale espressione può essere resa in italiano come diagramma «ramo-foglia».

Un pregio notevole è che esso prescinde da una rigida suddivisione dell'intervallo dei valori in classi. Tale diagramma, inoltre, facilita la lettura dei diversi aspetti della distribuzione con particolare riferimento a:

1. l'ampiezza del campo di variabilità dei dati;
2. l'esistenza dei valori intorno a cui i dati si concentrano;
3. la simmetria o asimmetria della distribuzione;
4. salti nell'intervallo dei valori ove non si osservano misurazioni;
5. valori che appaiono isolati rispetto al nucleo dei dati.

Prendiamo come base del nostro esempio l'insieme di prezzi di una merce ipotetica, secondo i diversi modelli in cui viene prodotta;

supponiamo che i prezzi espressi in migliaia di lire siano i seguenti: 290, 370, 470, 350, 485, 375, 540, 445, 280, 1970, 425, 380, 430, 320, 345, 350.

L'idea base del diagramma stem-and-leaf (ramo-foglia) è di utilizzare nella rappresentazione le stesse cifre che compongono i dati.

Eliminando eventuali punti decimali, avremo n numeri di k cifre: di queste scegliamo le prime k' più significative, in numero tale che diano dimensioni accettabili al grafico. Queste, scritte le une sotto le altre in modo crescente a percorrere l'intero campo dei valori, costituiscono il ramo del nostro diagramma.

Nel nostro esempio, il valore minimo è 280, il massimo 1970; questo appare però isolato dal resto dei valori e conviene perciò rappresentarlo su una riga a sé preceduto da un simbolo speciale, ad esempio HI (per high). Il valore massimo che rimane è 540. Siamo praticamente obbligati a scegliere come «ramo» del nostro diagramma i valori delle centinaia di migliaia di lire, la prima delle cifre che compongono i nostri numeri. Una scelta differente, diluendo eccessivamente il grafico, non faciliterebbe in modo apprezzabile la lettura dei dati.

Il ramo apparirà in questo caso nel modo seguente:

```
2
3
4
5
HI 197
```

Accanto a ciascun elemento del ramo e separato da esso con uno spazio, si porrà la prima delle cifre restanti, a costituire una foglia del ramo; ciò viene fatto per ogni numero della collezione, ponendo l'uno accanto all'altro sulla stessa linea le cifre che hanno le prime k' uguali.

L'esempio completato apparirà nel modo seguente:

```
unità: 10,0
2 89
3 1455778
4 23478
5 4
HI 197
```

In testa si è posta l'unità di misura che consente di leggere effettivamente i numeri; si fa osservare che a meno delle migliaia i singoli valori sono esattamente ricostruibili.

In generale, l'unità di misura potrà essere opportunamente modificata, per consentire rappresentazioni più concentrate o diluite secondo le necessità e tenendo conto che incrementarla di un fattore 10 comporta la perdita di una cifra significativa nella rappresentazione grafica.

Come variazione della rappresentazione grafica, a consentire una diluizione parziale delle cifre, senza modificare la quantità di informazione, si possono porre su ciascuna riga del ramo non tutte le cifre da 0 a 9, ma rappresentarne 5 o 2 con conseguente moltiplicarsi degli elementi del ramo in 2 o 5 righe differenti che potremo ad esempio rappresentare al modo seguente:

```
5*    5*
5.    5T
      5F
      5S
      5.
```

dove, nel primo caso, accanto a * faremo seguire le cifre da 0 a 4 e a . da 5 a 9; nel secondo a * i valori 0 e 1, a T 2 e 3, a F 4 e 5, a S 6 e 7 e a . 8 e 9. [9]

Utilizzando per il nostro esempio la prima delle due possibilità avremo:

```
unità = 10,0
2. 89
3* 24
3. 55778
4* 234
4. 78
5* 4
HI 197
```

La maggiore diluizione delle cifre consente di ricavare delle informazioni significative in modo probabilmente più immediato. Il grafico ci dice che i prezzi sono distribuiti in modo abbastanza simmetrico intorno ad un valore centrale di circa 37; il valore 197 è chiaramente un valore 'anomalo' che, se lo vogliamo, richiede un'analisi a sé.

Presentiamo nelle figure 3.28-29 esempi che utilizzano i dati relativi alle temperature medie annuali, espresse in °C, e alle precipitazioni annuali in mm per 39 stazioni termopluviometriche ed osservatori del bacino del Tevere nel 1980, ricavati dalla pubblicazione Istat (1981) *Annuario di statistiche meteorologiche*.

Dalla figura 3.28, osservato il caso «anomalo» del monte Terminillo con media annuale di 3,8 °C, possiamo dire che la temperatura sembra mostrare una distribuzione abbastanza uniforme fra gli 11 e 15 gradi centigradi. [10]

(9) T per two e three, F per four e five, S per six e seven.

(10) I grafici presentano sulla sinistra la colonna della profondità, che dà per ciascuna riga del ramo il numero dei valori presenti sulla riga e su quelle più vicine al limite più prossimo del gruppo dei dati. Per la riga che contiene il valore centrale, la mediana, viene riportato fra parentesi il numero di elementi della riga stessa.

Per le precipitazioni si può osservare una forte concentrazione intorno ai valori fra gli 800 e 1000 mm, con zone dai valori più elevati, intorno a 1150, 1500 e 1700 mm.

Riportiamo la definizione dell'indice di aridità del de Martonne (Dajoz 1972) (uno dei vari indici climatici utilizzato soprattutto per cercare di spiegare le ripartizioni della vegetazione), data dalla seguente formula:

$$i = \frac{P}{T + 10}$$

dove P è la piovosità annuale espressa in mm e T la temperatura media annuale in °C. Se rappresentiamo i valori di tale indice per le 39 zone, sembra di poter discriminare meglio alcune zone che già si potevano individuare nel grafico precedente. Si veda in proposito la figura 3.30.

B) Box-plot

Il box-plot, o «diagramma a scatola» (Tukey 1977, Lombardo 1984) è una rappresentazione che dà una visione sintetica della distribuzione e, come vedremo, per la sua struttura unidimensionale, consente facili confronti fra serie diverse. Per la sua costruzione sono necessarie alcune semplici definizioni preliminari.

Siano $X_1, X_2, X_3, \ldots, X_i \ldots, X_n$ i numeri che costituiscono il nostro gruppo di dati. Da essi formiamo l'insieme ordinato $X_{(1)} \ldots, X_{(i)} \ldots, X_{(n)}$ dove $X_{(i)}$ è la i-esima osservazione più piccola.

Per ciascun elemento definiamo suo rango superiore e inferiore la misura della distanza, intesa come conteggio di elementi, rispettivamente dal minore e dal maggiore degli elementi dell'insieme ordinato, percorrendo i dati una volta nel verso crescente e l'altra nel verso decrescente. Chiaramente, se l'elemento X_j diventa $X_{(i)}$ nell'insieme ordinato, il suo rango superiore sarà i e quello inferiore $(n + 1 - i)$.

Per ragioni di simmetria, ovvero per dare uguale importanza ai due estremi dell'insieme ordinato, sembra conveniente definire per ciascun elemento un'altra quantità : la sua profondità; essa è il più piccolo fra i due ranghi associati al dato.

La profondità di un elemento ci dà la distanza di questo dal più vicino dei due estremi del gruppo dei dati. Utilizzando questa nuova quantità possiamo ricavare una serie di quantità chiave nel gruppo dei dati ordinati. A profondità $(n + 1)/2$ abbiamo la mediana, quel valore centrale della distribuzione che ha fra i dati tanti elementi ad essa superiori quanti sono quelli inferiori.[11] Indicheremo la mediana con la lettera M.

(11) Nel caso di n pari calcoleremo la mediana come media aritmetica dei due valori vicini di rango superiore $(n + 1)/2 - 1/2$, $(n + 1)/2 + 1/2$.

A profondità 1 troviamo chiaramente i due valori estremi della distribuzione. In questo caso, come per ogni altro, escludendo la mediana, avremo a una certa profondità due valori posti simmetricamente rispetto ad M.

Un modo che riesce utile per caratterizzare il gruppo dei dati è considerare successivamente i punti centrali delle code della distribuzione definite dai valori precedentemente calcolati. Partendo dalla mediana si possono calcolare i quartili della distribuzione, che indicheremo con la lettera F, come i valori mediani delle code della distribuzione a destra e a sinistra di M.

Per definire il box-plot è sufficiente far riferimento esclusivamente ai due quartili, superiore ed inferiore, ed alla loro differenza, detta differenza interquartile. Questa può essere designata, indicando con F_u il quartile superiore e F_l il quartile inferiore, al modo seguente:

$$d_F = F_u - F_l$$

Tramite d_F possiamo ricavare dei criteri empirici per la individuazione di valori anomali. Un criterio possibile (Tukey, 1977) consiste nel definire anomali quei dati che cadono al di fuori dell'intervallo di troncamento 1,5 d_F: $(F_l - 1{,}5\ d_F, F_u + 1{,}5\ d_F)$.

Nel nostro esempio abbiamo come valori anomali 1700 e 1760 delle stazioni rispettivamente di Atina e Casamari. Sulla base delle precedenti definizioni e posizioni è possibile ricavare una semplice ed efficace rappresentazione grafica che consente di evidenziare, senza concentrarsi nella analisi dei singoli valori, le caratteristiche di localizzazione, di dispersione, di simmetria, di lunghezza delle code e la eventuale presenza di dati anomali. Questo grafico viene costruito nel modo seguente.

Su una linea orizzontale, scelta una scala adatta per rappresentare l'intero gruppo dei dati, si indicherà con un segno + la posizione della mediana.

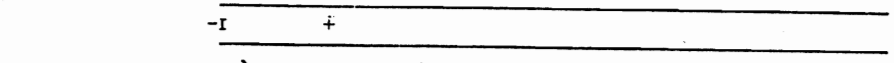
Ai lati vengono segnate con due segmenti verticali le posizioni dei due quartili superiore ed inferiore. Per evidenziare che in questa zona del grafico ricade la metà dei valori della distribuzione, uniamo gli estremi corrispondenti dei due segmenti a formare una scatola rettangolare.

Partendo dai due segmenti tracciamo verso l'esterno due linee fino ai punti che, sulla scala, corrispondono ai valori della distribuzione immediatamente superiore ed inferiore rispettivamente a $F_l - 1{,}5\ d_F$ e $F_u + 1{,}5\ d_F$. Questi punti danno una idea della lunghezza delle code della distribuzione. Indicheremo con un * i valori esterni all'intervallo di troncamento 1,5 d_F; se fra questi ce ne sono alcuni lontani dal rispettivo quartile per una quantità superiore a $3d_F$ li evidenzieremo con il simbolo O; questi ultimi li chiameremo «valori molto anomali». Riportiamo come esempio, nella figura 3.31, i box-plot dei dati sulle precipitazioni dell'anno 1980

del bacino del Tevere e del bacino medio sinistro del Po, sulla medesima scala, uno sotto l'altro, per consentire il loro confronto.

Si vuole osservare che un uso meccanico di questo metodo può portare a valutazioni fuorvianti.

Ad esempio il seguente box-plot:



è stato ottenuto dai dati ricavabili dal diagramma stem-and-leaf qui sotto riportato.

unità = 0,0010

```
 6  229.   888889
(2) 230*   11
 7  230T
 7  230F
 7  230S
 7  230.   9
 6  230T*  000000
```

Da questo si ricava chiaramente che esistono due gruppi di dati posti all'estremità della distribuzione e che questa non presenta alcun dato nella zona intermedia.

Un'utilizzazione abbastanza naturale del box-plot è il confronto fra differenti gruppi di dati. Ciò è consentito in modo semplice dal fatto che i box-plot relativi a distribuzioni differenti possono essere accostati l'uno all'altro senza difficoltà. Tale accostamento consente di percepire immediatamente la posizione relativa delle mediane e dei quartili riportati sul grafico.

La figura 3.32 ripresa dal testo di Hoaglin, Mosteller e Tukey (1983) riporta i box-plot relativi alle popolazioni nell'anno 1962 delle 10 maggiori città in alcuni paesi.

6. IL PROBLEMA DELLA RAPPRESENTAZIONE GRAFICA DELLE QUALITÀ

Le qualità possono essere rese graficamente secondo metodi diversi; quelli più comuni fanno uso di:

a) espressioni verbali;
b) variazione di valore;
c) variazione di tessitura o grana e di orientamento;
d) variazione di colore;
e) ideogrammi.

Il sistema verbale è praticamente inevitabile dal momento che l'uso di un qualsiasi simbolo richiede la sua associazione semantica tramite una apposita didascalia. Dal momento che la lettura è un processo che richiede tempo ed un minimo di concentrazione, conviene che la didascalia sia breve e sintetica; eventuali spiegazioni più dettagliate possono essere fornite nelle note al grafico.

In taluni casi la spiegazione verbale può essere data direttamente sul grafico, ad esempio accanto alla linea che indica l'andamento di un fenomeno, invece che essere data in modo mediato attraverso l'uso aggiuntivo di un altro simbolo, come il colore o il tratteggio. Ciò ha il vantaggio di evitare che l'occhio del lettore passi continuamente dalla legenda al grafico.

La disposizione orizzontale delle scritte può facilitare la lettura; i diagrammi a barre disposte orizzontalmente si prestano ovviamente a tale tipo di esemplificazione riuscendo immediata l'associazione fra didascalia e simbolo.

Un altro metodo talvolta utilizzato per differenziare le modalità di un carattere qualititativo è la variazione di valore, resa ad esempio molto semplicemente tramite il tratteggio. Tale metodo appare utilizzabile solo quando il numero delle modalità da rappresentare è limitato, in quanto altrimenti le barre potrebbero diventare scarsamente visibili per i valori iniziali.

Per ottenere un sistema di campiture che dia un effetto visivo equivalente come intensità percettiva, è possibile utilizzare la variabile visiva «orientamento», di cui un esempio è dato nella figura 2.2, ripresa dal Bertin (1983); tuttavia la possibilità di discriminare fra i diversi orientamenti è piuttosto limitata.

La variabile visiva «grana» o «tessitura» ha proprietà percettive simili, da questo punto di vista; infatti la quantità di colore che viene distribuita su una data superficie, risulta costante come per la variabile «orientamento».

Tale tipo di variabile si presta anche alla rappresentazione delle modalità di caratteri ordinabili, di tipo non quantitativo.

L'uso del colore, fra i possibili metodi proposti, ha certamente il vantaggio della vivezza della presentazione e della rapida discriminabilità fra le diverse tinte. Vanno riprese alcune considerazioni già svolte nel paragrafo 6 del capitolo 2 sulla scelta adeguata di una scala cromatica; se le modalità sono sconnesse i toni dei diversi colori della scala dovrebbero essere di pari luminosità.

Per quanto riguarda gli ideogrammi, si riprendono alcune considerazioni già riportate nel paragrafo "Classificazione dei tipi di grafici". L'uso generalizzato di tali simboli comporta delle limitazioni alla impostazione di un sistema di rappresentazioni grafiche articolato e flessibile. Essi si prestano soprattutto ad associare, nelle rappresentazioni di tipo areale, una informazione di tipo qualitativo ad una di tipo spaziale; in questi casi vanno distinti gli ideogrammi di tipo generico, come quelli della figura 2.2, che sono concepiti in funzione della loro massima discernibilità, e di tipo specifico, quando essi con la loro forma cercano di richiamare l'idea del fenomeno.

7. I DIAGRAMMI DI TIPO AREALE E A BARRE SUDDIVISE

Si discute inizialmente dei diagrammi areali, cioè di quei diagrammi che rendono la componente quantitativa dell'informazione tramite la variazione del valore delle aree che li compongono. Nella loro forma più semplice essi si presentano come diagrammi rettangolari o a settori circolari, più noti questi ultimi come diagrammi a torta. Dal momento che la valutazione da parte del sistema visivo del valore di aree è meno immediata e precisa di quella della lunghezza di segmenti, per la rappresentazione di frequenze, ammontari e quantità derivate, i diagrammi a barre, che nelle loro varie combinazioni si prestano a risolvere la maggior parte dei problemi che si presentano nella pratica, mostrano una maggiore efficacia.

Uno dei vantaggi che i diagrammi areali di tipo rettangolare offrono è che essi possegono due dimensioni indipendenti da utilizzare per la presentazione di dati quantitativi; ad esempio nel caso di quozienti il loro uso consente la rappresentazione del quoziente stesso e della grandezza della popolazione presente a denominatore.

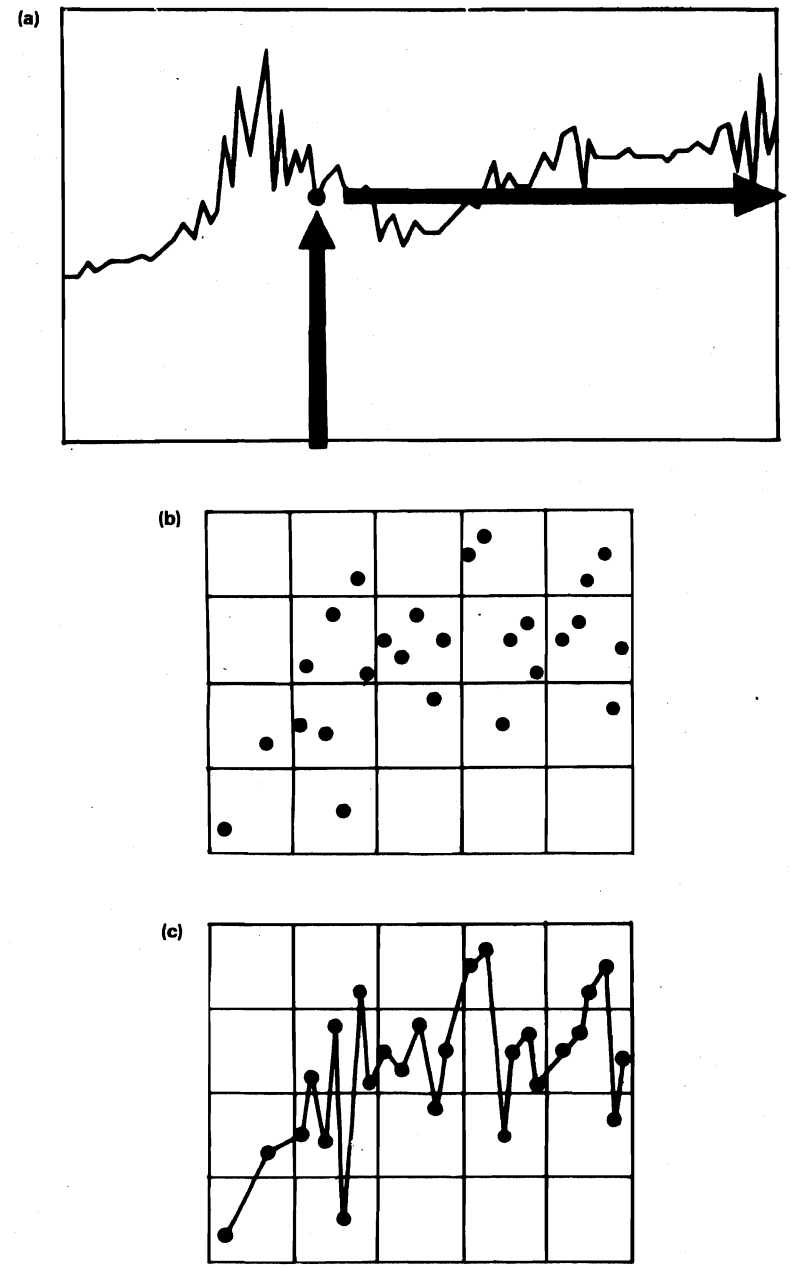
Per la loro vasta utilizzazione si vuol dedicare alcune osservazioni aggiuntive alle proprietà del diagramma a torta. Tale tipo di rappresentazione è decisamente inefficiente, come risulta da numerosi contributi, fra cui ricordiamo quello recente di Cleveland e McGill (1984), specialmente quando il carattere è composto da molte modalità e quando queste assumono valori simili, come appare dalla figura 3.33.

Esso soprattutto non si presta al confronto fra serie diverse; infatti, dal momento che la posizione dei settori all'interno del cerchio varia generalmente da diagramma a diagramma, la sovrapposizione dei due settori corrispondenti richiede generalmente una rotazione ideale di uno dei due settori; questa operazione complica e rende inefficace il confronto fra le due grandezze.

Se lo scopo del grafico è quello di consentire la valutazione della parte sul tutto, come nel caso della composizione percentuale di una grandezza suddivisa fra poche modalità, il diagramma a torta presenta in generale una semplicità di lettura che lo rende altrettanto valido che il diagramma a barre semplici (Simkin e Hastie 1987).

L'uso dei diagrammi a barre suddivise presenta difficoltà simili; come è chiaramente mostrato dalla figura 2.6a del Capitolo 2 il confronto fra le quantità relative a distribuzioni diverse è certamente difficile. La sua utilizzazione può essere giustificata da considerazioni di tipo teorico, come, per esempio, quando per la rappresentazione della composizione del conto economico delle risorse e degli impieghi si vuole evidenziare l'eguaglianza del valore complessivo delle due parti del conto; a tal proposito si fornisce nella figura 3.34 un esempio ripreso dall'Annuario Statistico Italiano 1988.

Figura 3.1 — Esempi di rappresentazioni grafiche a scopo didattico



Fonte: Bertin (1983)

Figura 3.2 — Rappresentazione grafica «tipo» di due serie storiche fittizie

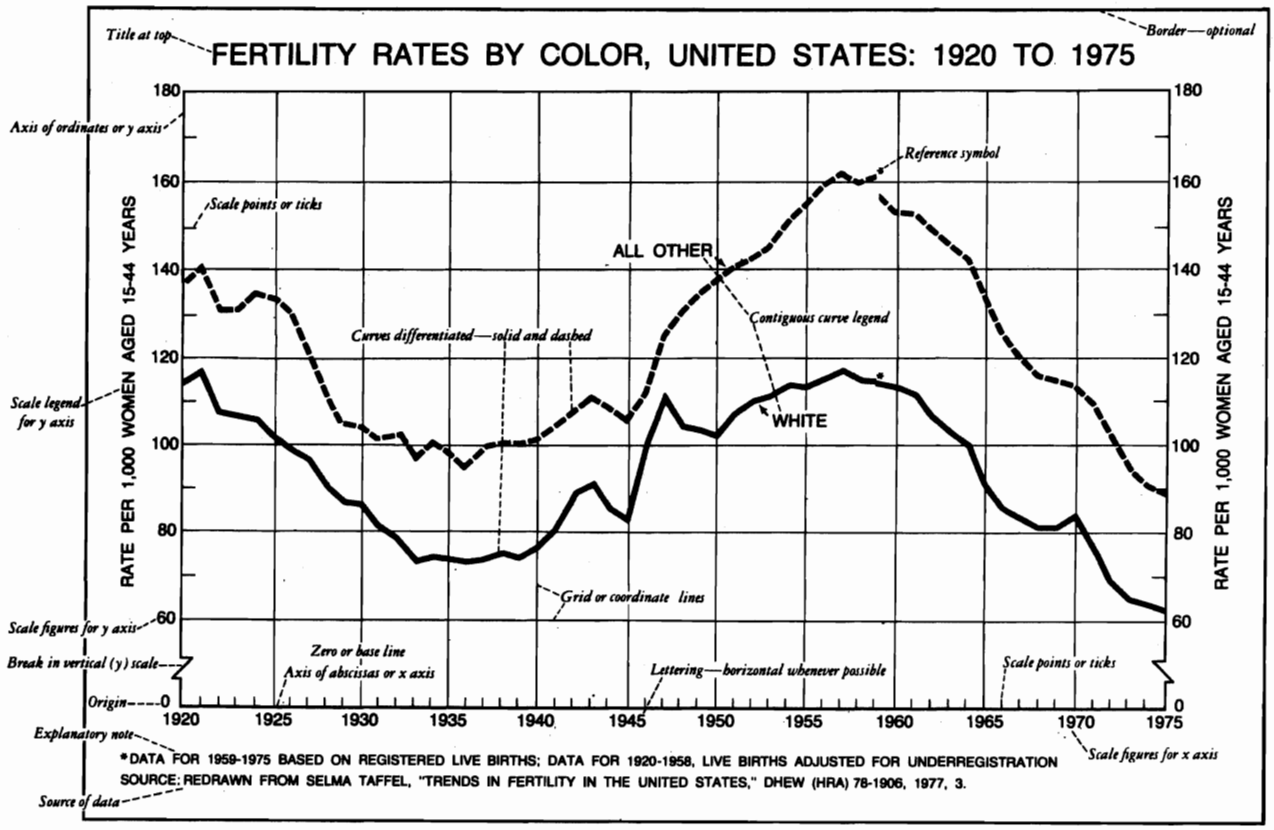


Figure 2-1. Chart designed as a model to illustrate the essential elements or components of a rectilinear coordinate line chart.

Fonte: Schmid (1983)

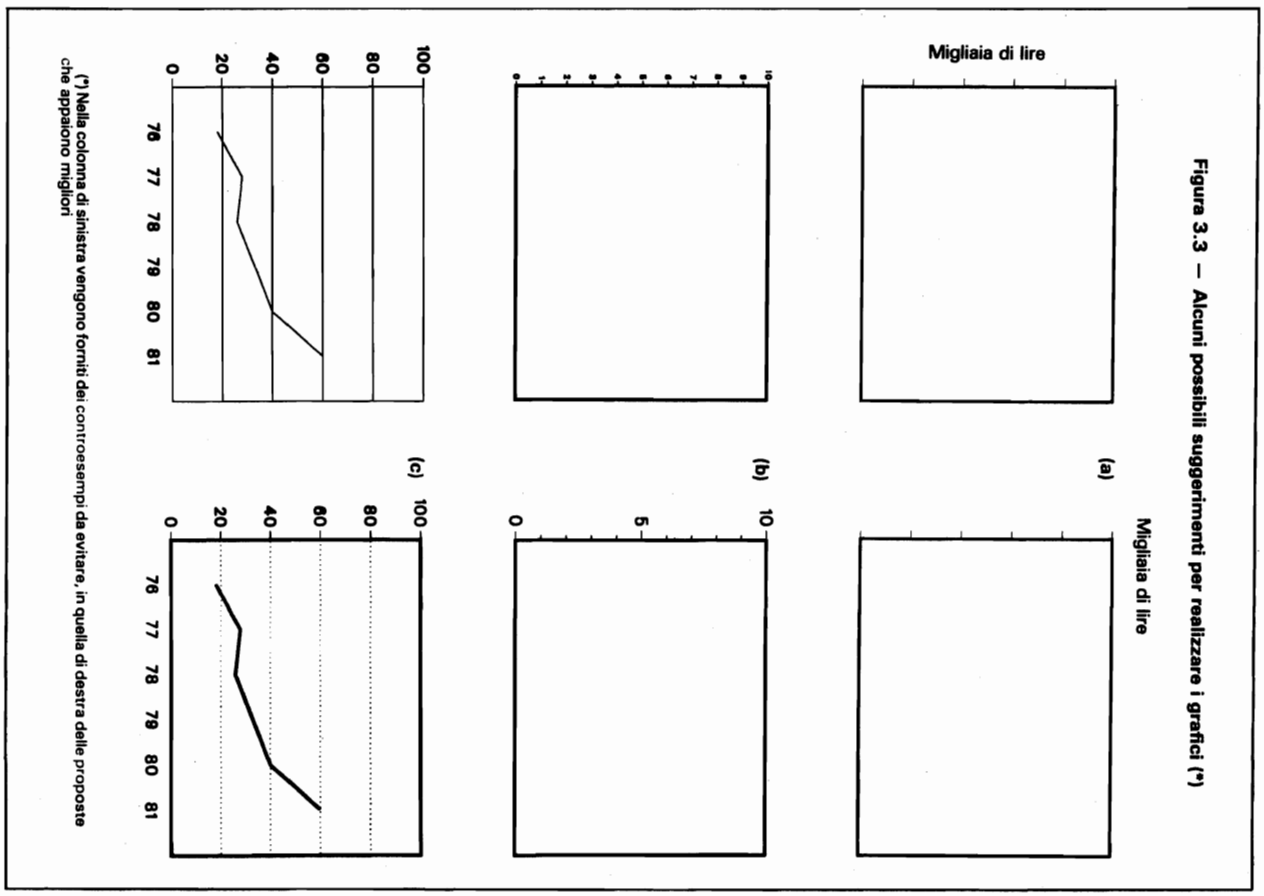
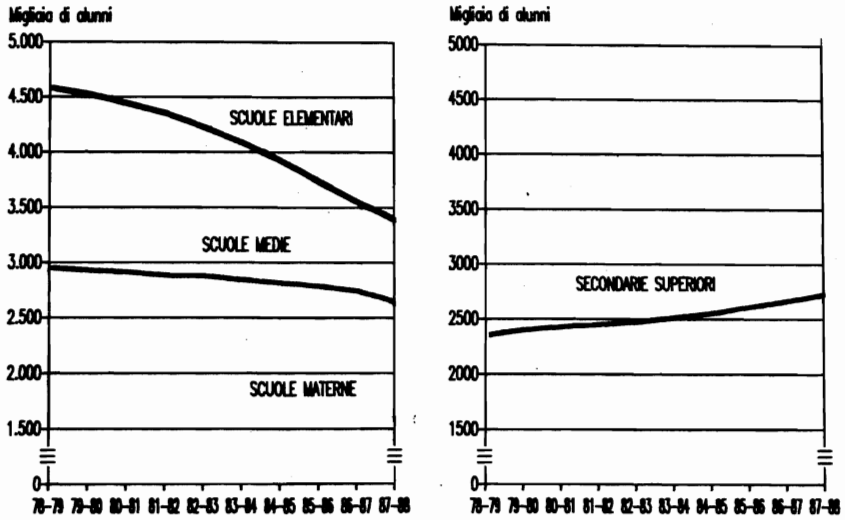


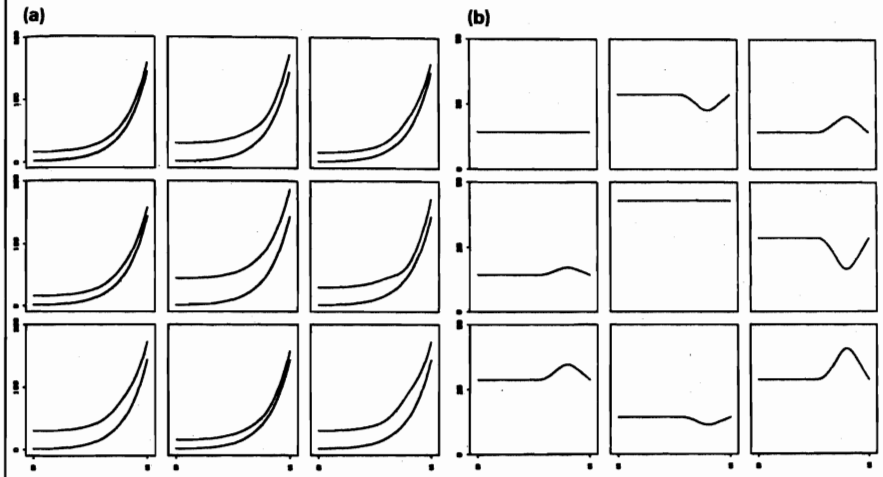
Figura 3.3 — Alcuni possibili suggerimenti per realizzare i grafici (*)

Figura 3.4 — Segnalazione dell'interruzione della scala



Fonte: Annuario Statistico Italiano (1988)

Figura 3.5 — Confronto fra due curve



Nota: a ciascuna coppia di curve nei grafici a) corrisponde nella parte b) il grafico della loro differenza.

Fonte: Cleveland e McGill (1984).

Figura 3.6 — Esempi di rappresentazione congiunta di serie storiche

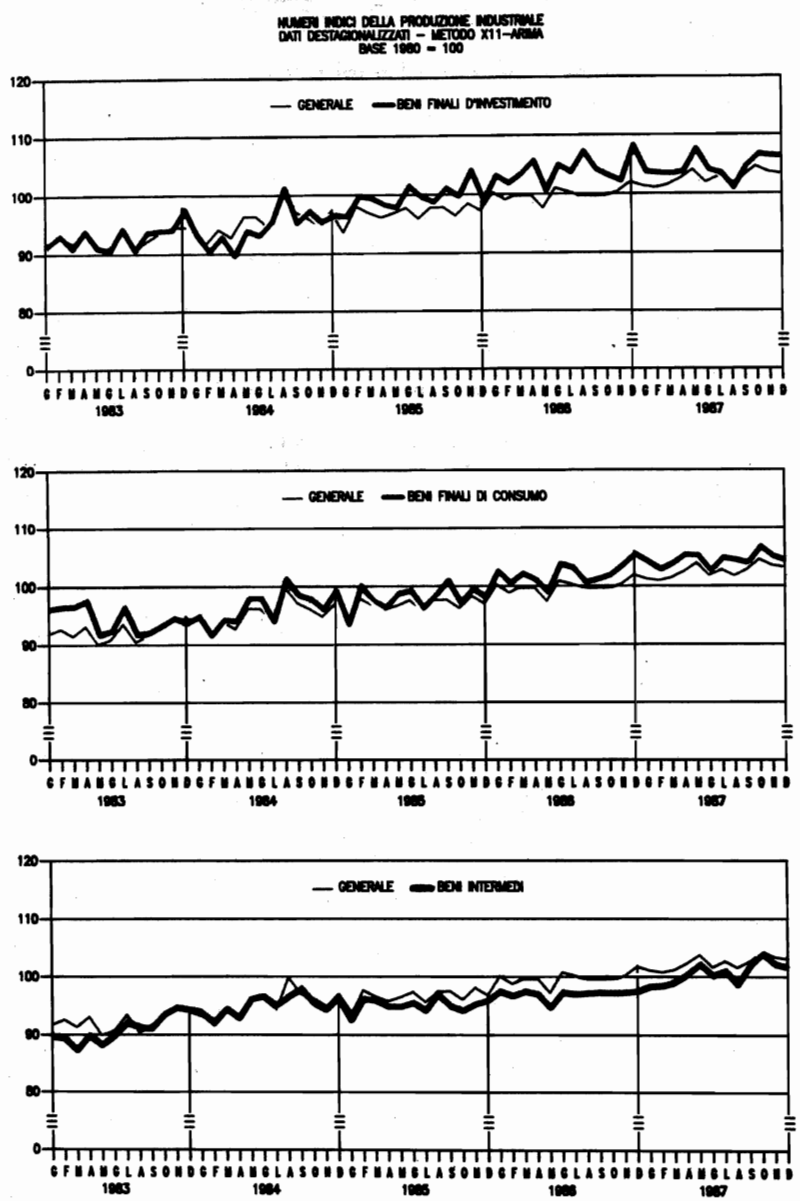
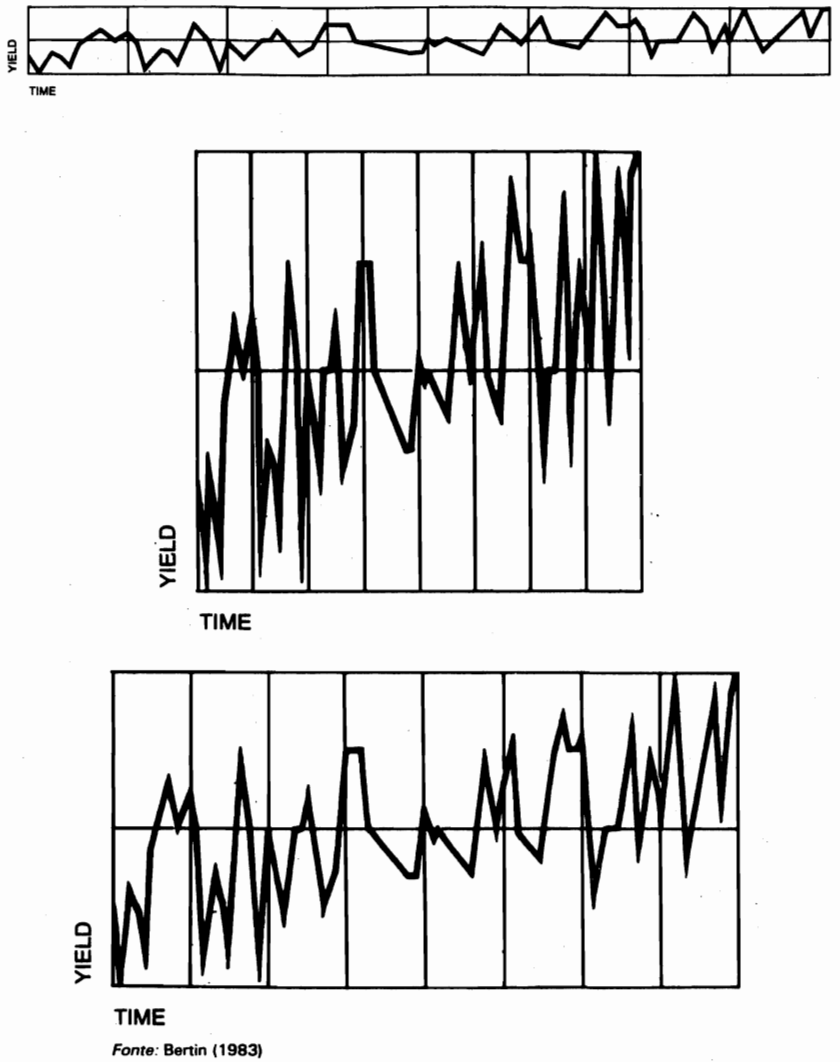


Tavola illustrata n. 11.2 - Numeri indici della produzione industriale per destinazione economica - Base: 1980 = 100 - Dati destagionalizzati - Anni 1983 - 1987

Fonte: Annuario Statistico Italiano (1988)

Figura 3.7 — Differenti impressioni visive al variare della scala degli assi



Fonte: Bertin (1983)

Figura 3.8 — Scale standard per l'asse dei tempi

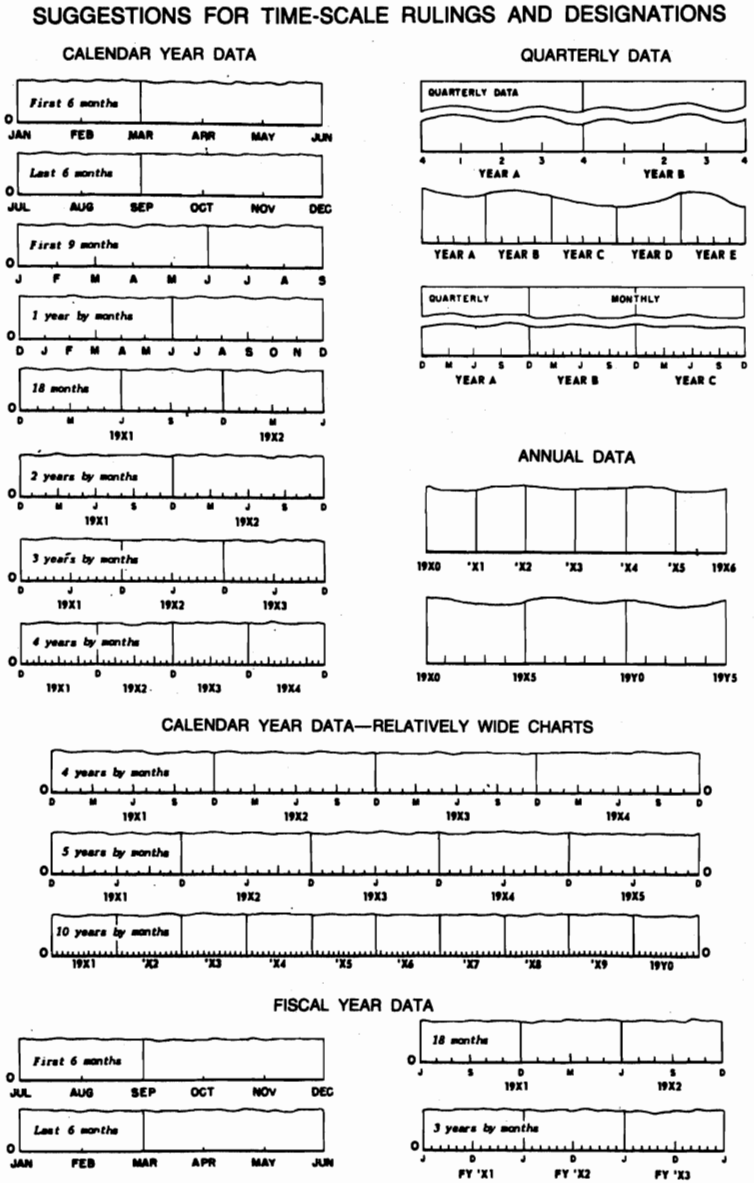


***Figure 2-14.** It will be observed from these sketches that in general, the size of the chart, length of the period covered, and nature of the time interval determine the number of rulings and scale designations for the time axis of rectilinear coordinate charts. For example, every month should be labeled for periods up to 1 year; every third month for periods from 15 to 24 months. Relatively wide charts should have more scale labels.* **(From Department of the Army, Standards of Statistical Presentation, *Department of the Army Pamphlet 325–10, 1966, pp. 86–87.*)**

Fonte: Schmid (1983)

Figura 3.9 — Diagramma a scalini

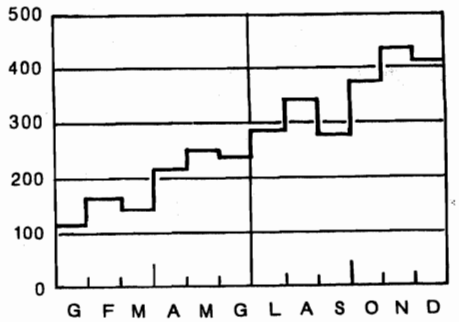


Figura 3.10 — Esempio di scala logaritmica

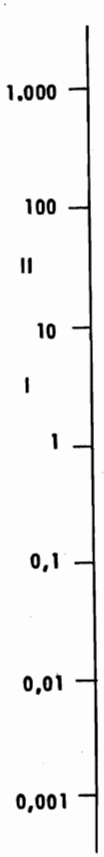


Figura 3.11 — Rappresentazione di una serie temporale in scala naturale ed in scala semilogaritmica

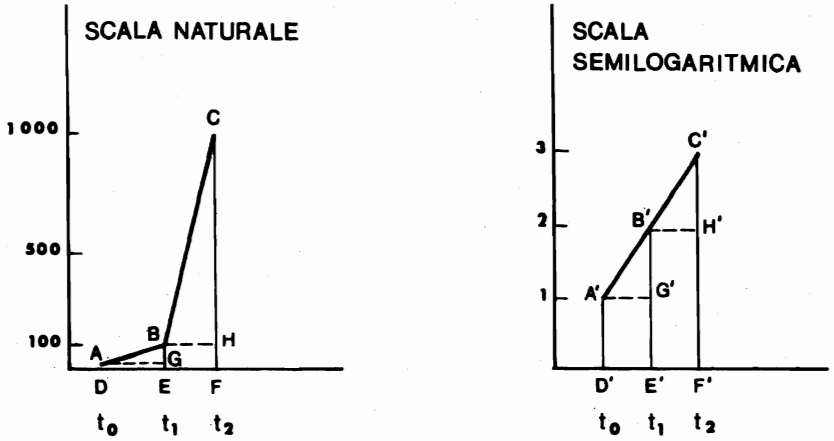


Figura 3.12 — Significato dei vari tipi di curve sul diagramma semilogaritmico.

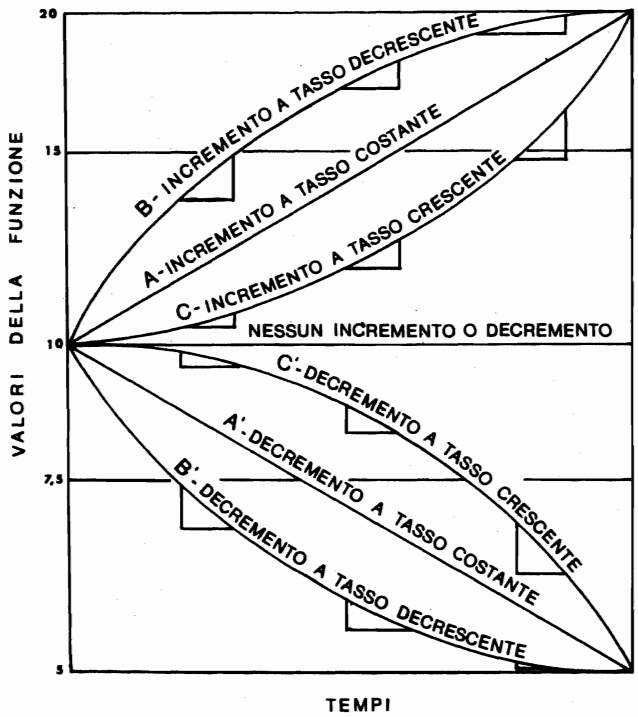
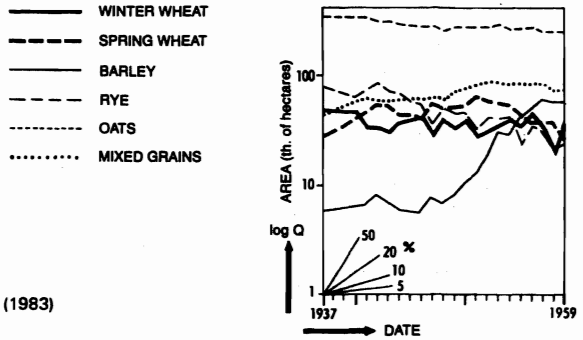
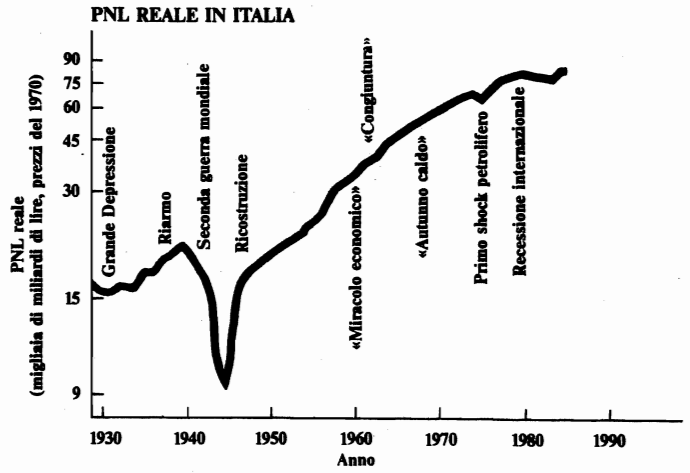


Figura 3.13 — Diagramma semilogaritmico con indicazione della corrispondenza fra pendenza e tasso di crescita



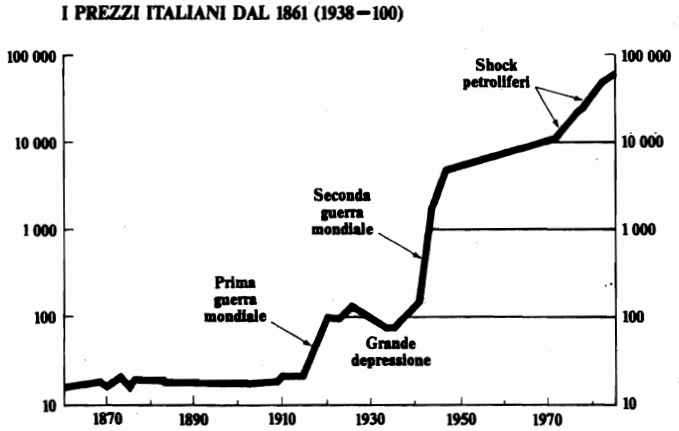
Fonte: Bertin (1983)

Figura 3.14 — Applicazione di diagrammi semilogaritmici a serie economiche relative al PNL

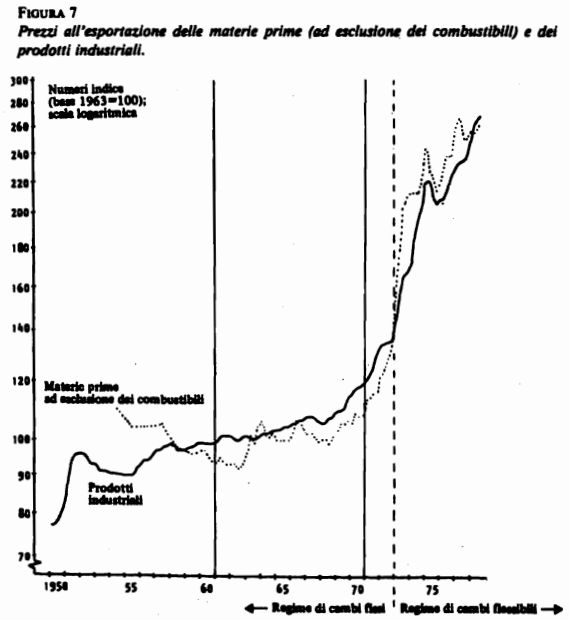


Fonte: Samuelson, Nordhaus (1987)

Figura 3.15 — Applicazione dei diagrammi semilogaritmici a serie economiche relative a numeri indice dei prezzi



Fonte: Samuelson, Nordhaus (1987)



Fonte: Lindbeck (1980)

Figura 3.16 — Applicazione dei diagrammi semilogaritmici a serie demografiche - valori assoluti e tassi

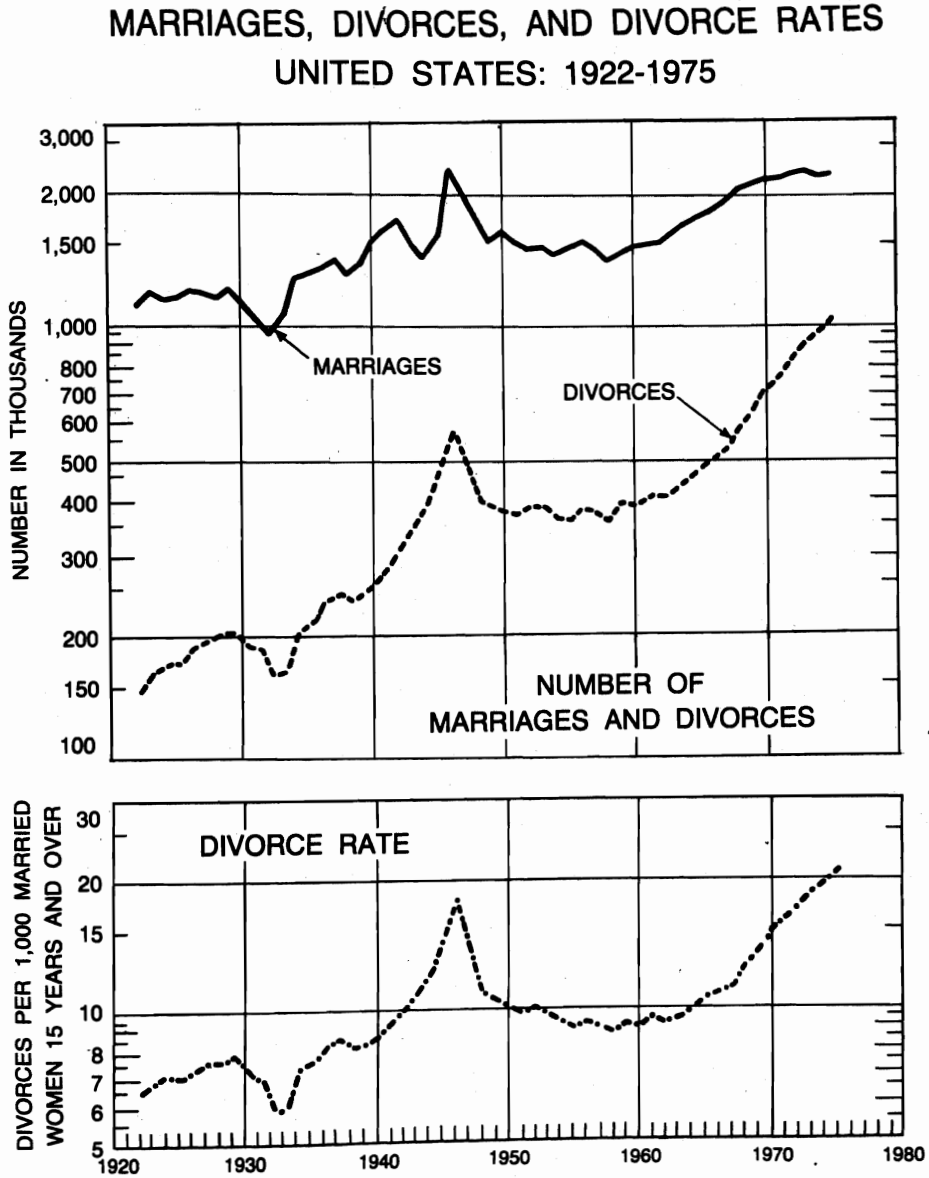


***Figure 5-12.** This chart is a reconstruction of Figure 5-11. See text for a discussion of the revisions that have been made.*

Fonte: Schmid (1983)

Figura 3.17 — Esempi di diagrammi semilogaritmici ripresi dall'Annuario Statistico Italiano (1988)

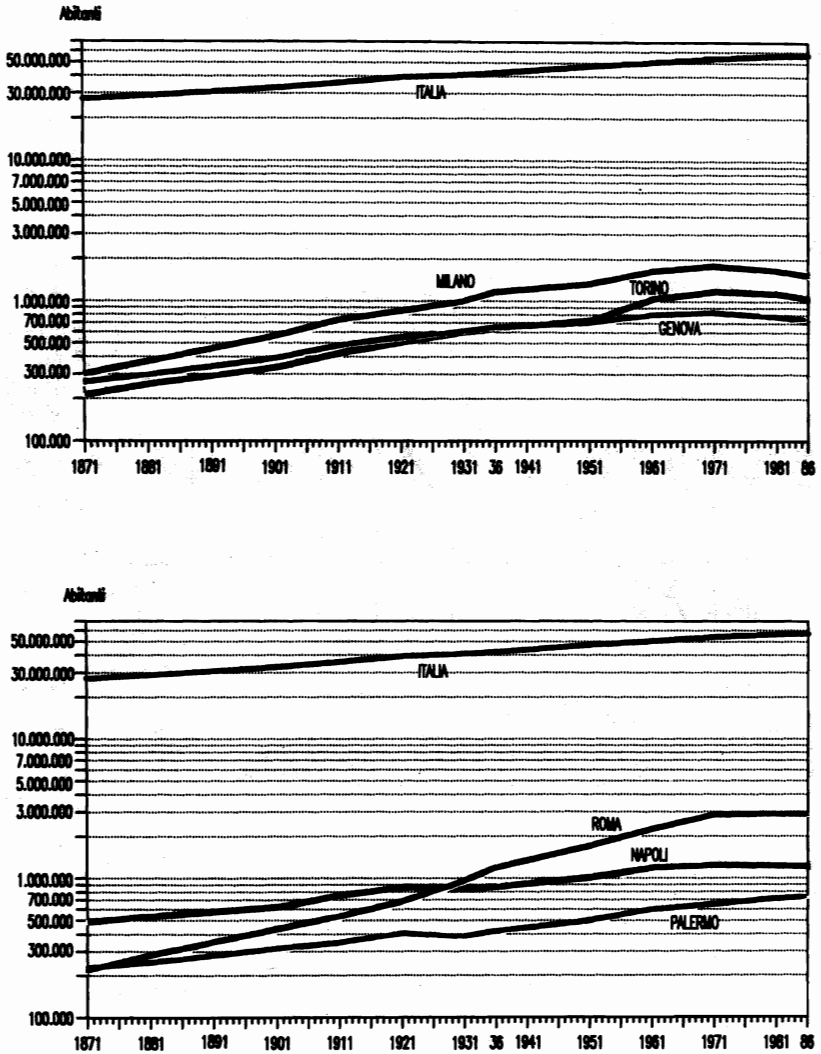


Tavola illustrata n. 2.9 - Popolazione residente in Italia e in alcuni principali Comuni dal 1871 al 1986. Diagrammi semilogaritmici

Figura 3.18 — Esempi di diagrammi semilogaritmici ripresi dall'Annuario Statistico Italiano (1988)

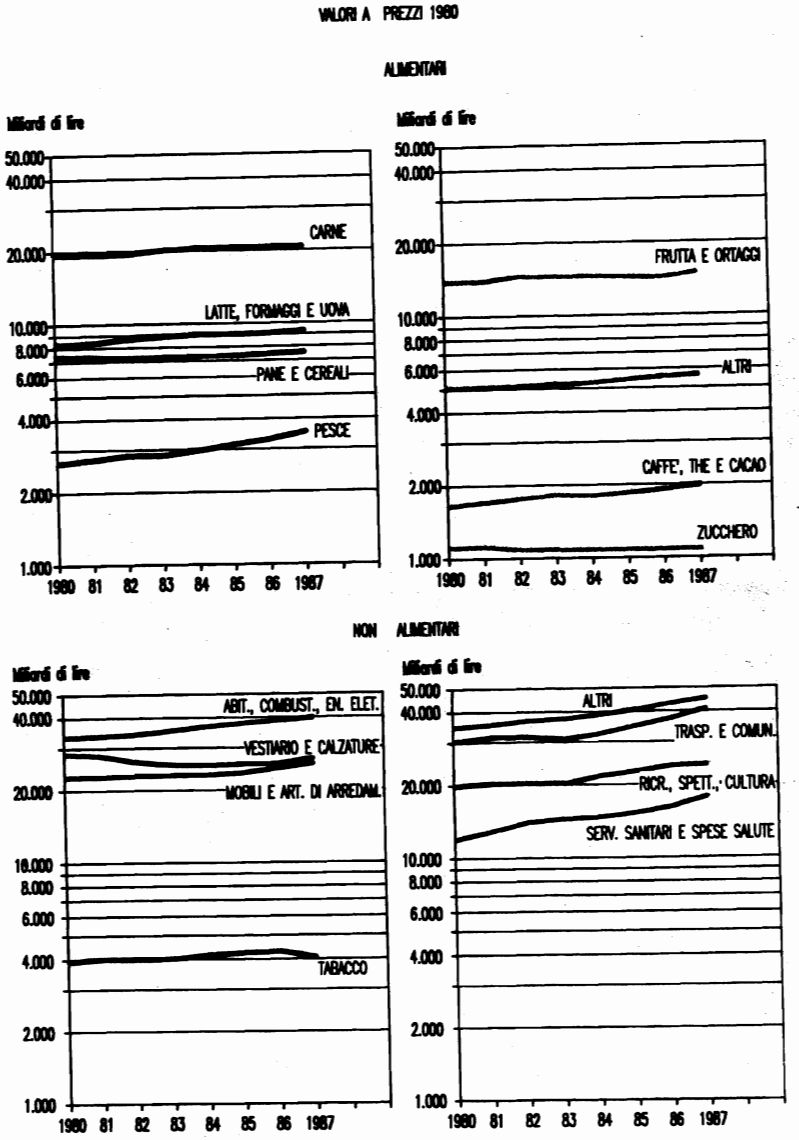
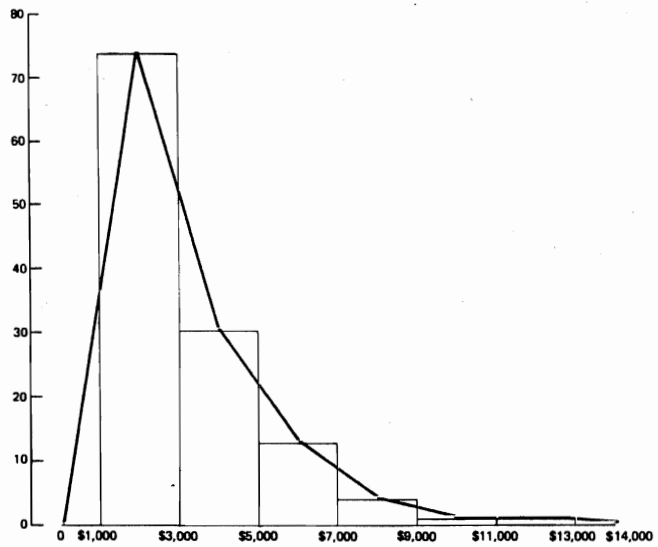


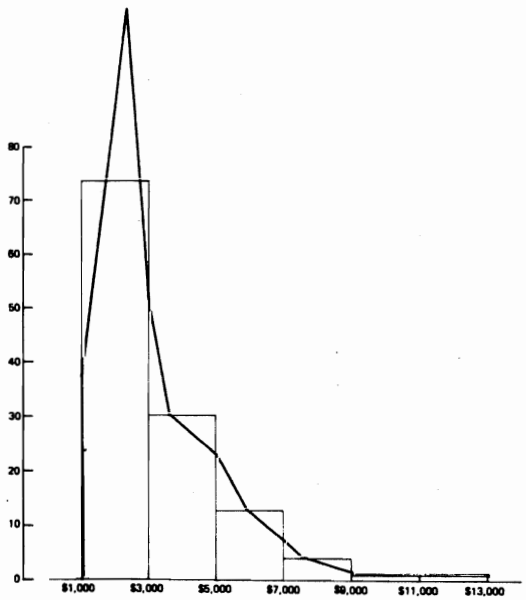
Tavola illustrata n. 8.2 - Consumi finali delle famiglie - Valori a prezzi 1980
Diagrammi semilogaritmici - Anni 1980 - 1987

Figura 3.19 — Esempio di poligonale costruita da un istogramma di frequenze



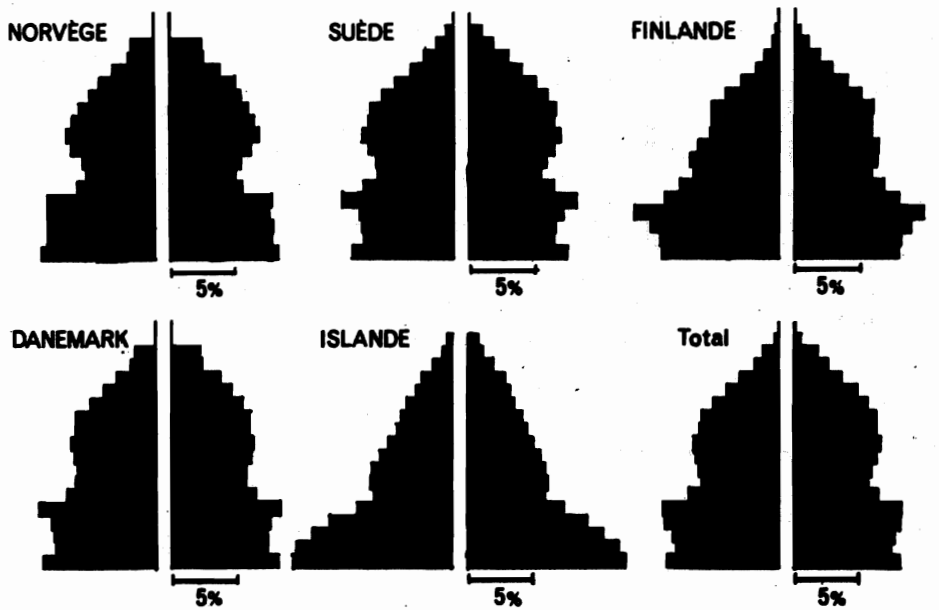
Fonte: Bachi (1978)

Figura 3.20 — Poligonale basata sul metodo delle aree



Fonte: Bachi (1978)

Figura 3.21 — Confronto fra piramidi delle età

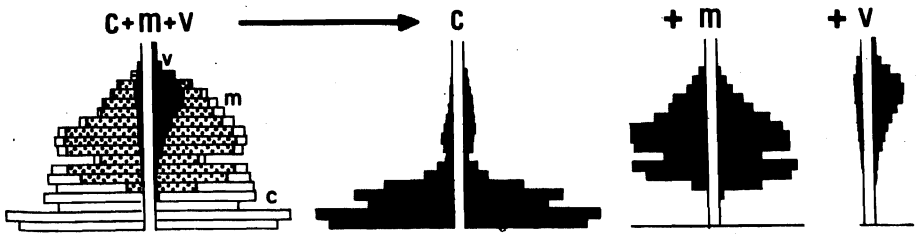


Fonte: Bonin (1975)

Figura 3.22 — Ulteriori confronti analitici fra piramidi delle età

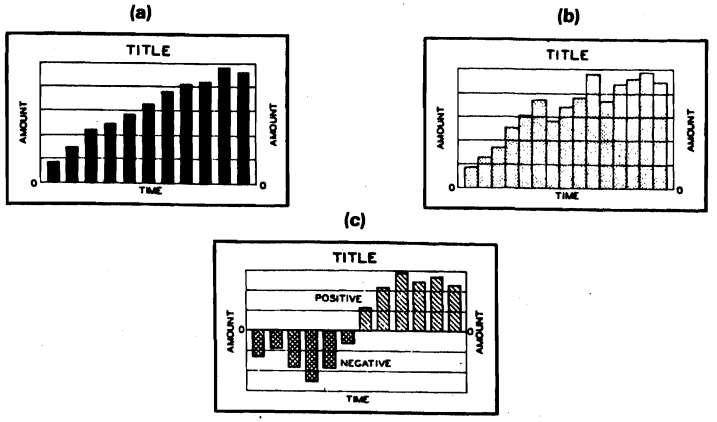
Fonte: Bonin (1975)

Figura 3.23 — Piramide delle età di una popolazione distinta per stato civile



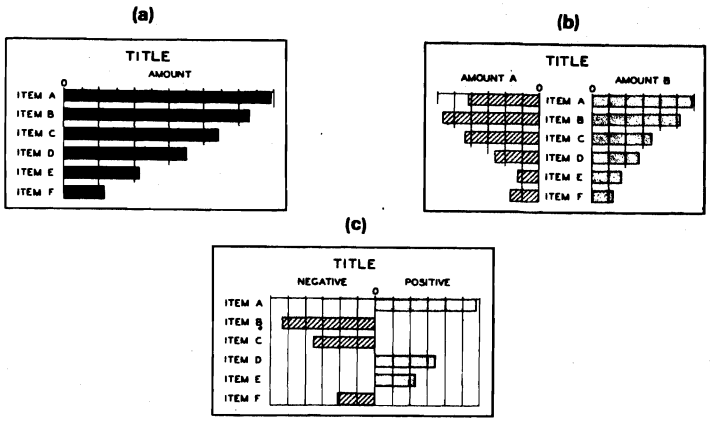
Fonte: Bonin (1975)

Figura 3.24 — Esempi di diagrammi a colonne



Fonte: Schmid (1983)

Figura 3.25 — Esempi di diagrammi a nastri



Fonte: Schmid (1983)

Figura 3.26 — Esempi di campitura delle barre

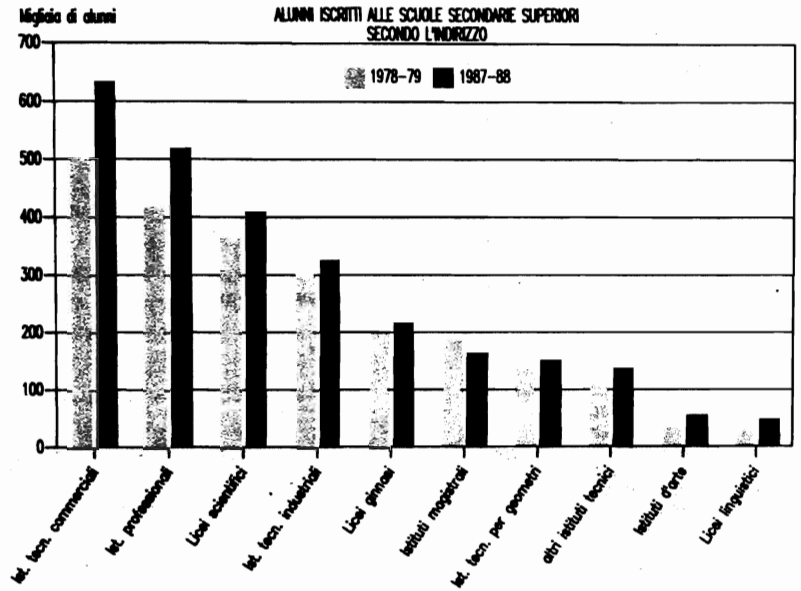
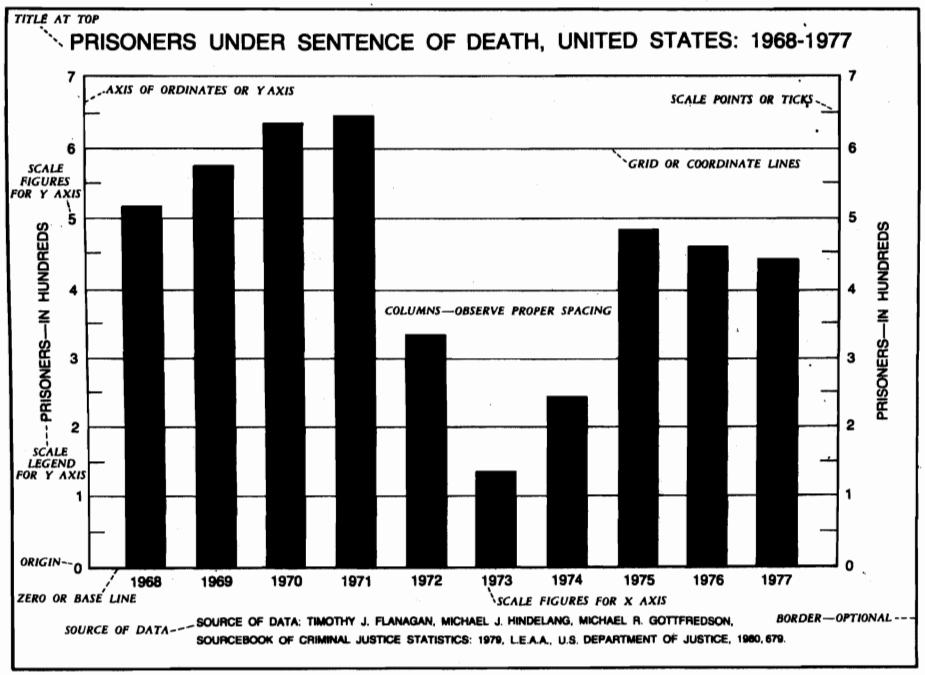


Tavola illustrata n. 4.1 - Alunni iscritti secondo il tipo di scuola

Fonte: Annuario Statistico Italiano (1988)

Figura 3.27 — Esempi standard di diagrammi a barre

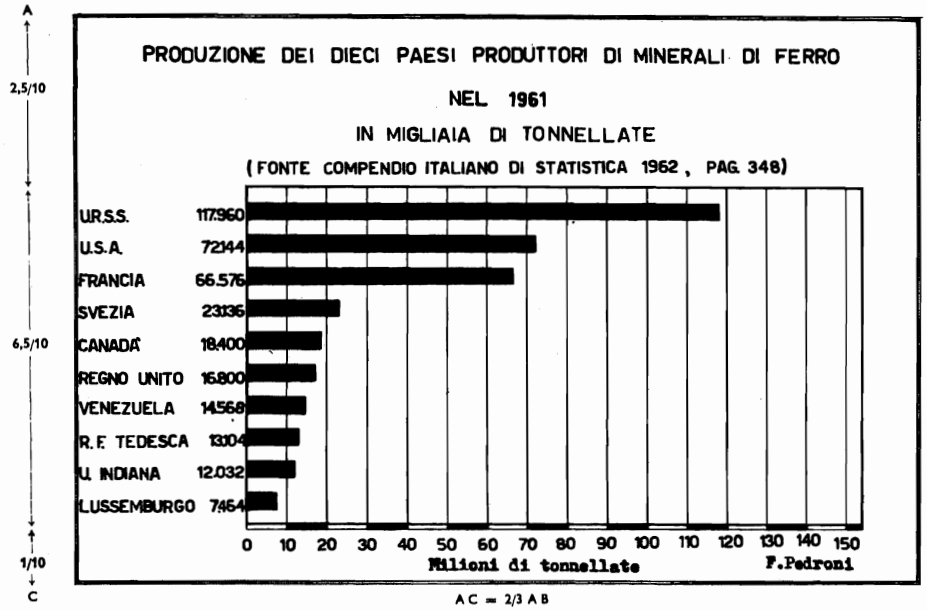
(a) colonne



Fonte: Schmid (1983)

Figura 3.27 — Esempi standard di diagrammi a barre

(b) nastri



Fonte: Pedroni (1968)

Figura 3.28 — Temperature medie annuali - Bacino del Tevere - Anno 1980

```
unità = 0,10
LO 38

 2   9. 8
 3  10* 2
 5  10. 57
 9  11* 3344
10  11. 6
13  12* 224
17  12. 6899
(3) 13* 013
19  13. 66799
14  14* 0334
10  14. 5578
 6  15* 004
 3  15. 68
 1  16* 0
```

Figura 3.29 — Precipitazioni annuali - Bacino del Tevere - Anno 1980

```
unità = 10,0

 1   5 4
 2   6 2
 3   7 4
12   8 011467799
19   9 1234568
(6) 10 333359
14  11 688
11  12 0223
 7  13
 7  14 3
 6  15 244
 3  16 1

HI 170, 176
```

Figura 3.30 — Indice di aridità - Bacino del Tevere - Anno 1980

unità = 10,0

```
 3 2.  579
 5 3*  23
16 3.  557778399
(7) 4* 0111224
16 4.  7
15 5* 0111
11 5.  667
 8 6* 2
 7 6.  57
 5 7* 1222
 1 7* 1222
```

Figura 3.31 — Box-plot delle precipitazioni del bacino del Tevere e del medio Po sinistro - Anno 1980

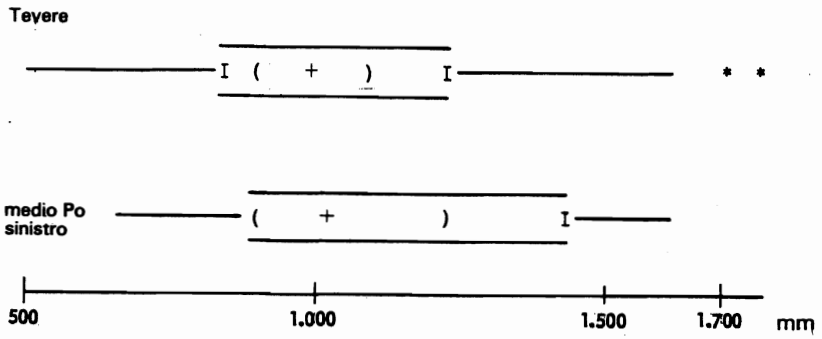
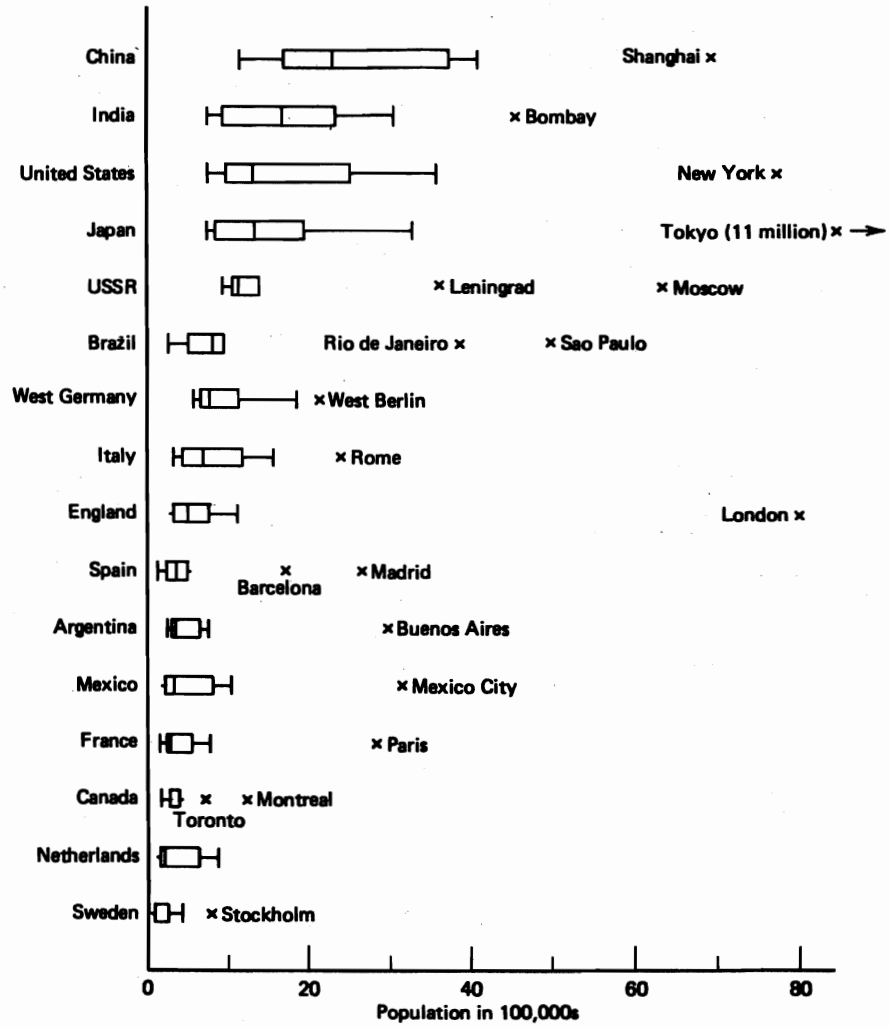


Figura 3.32 — Box-plot della popolazione delle 10 maggiori città di 16 paesi nel 1962



Fonte: Hoaglin, Mosteller, Tukey (1983)

Figura 3.33 — Confronto fra il diagramma a punti e a «torta»

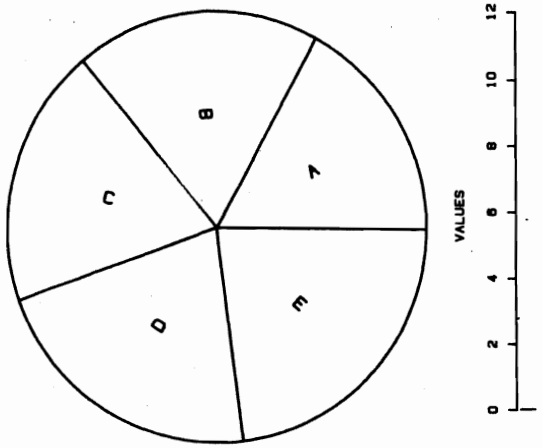


Figure 22. Pie chart.

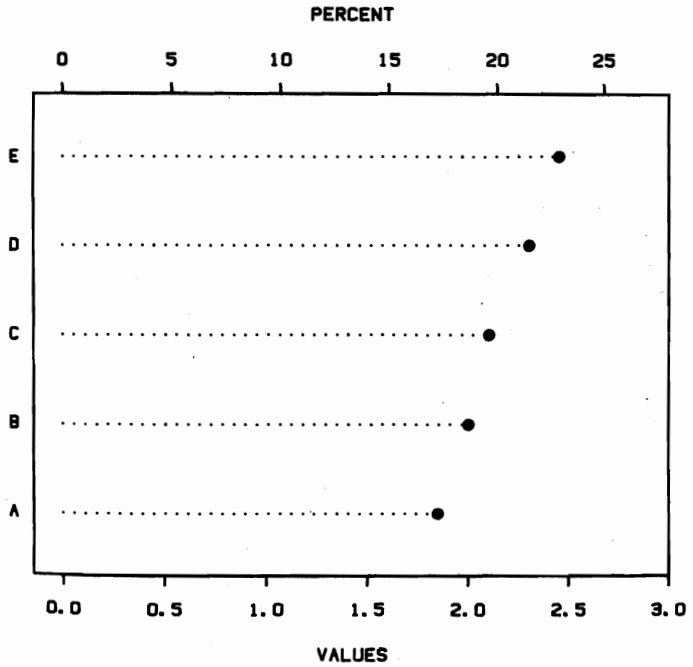
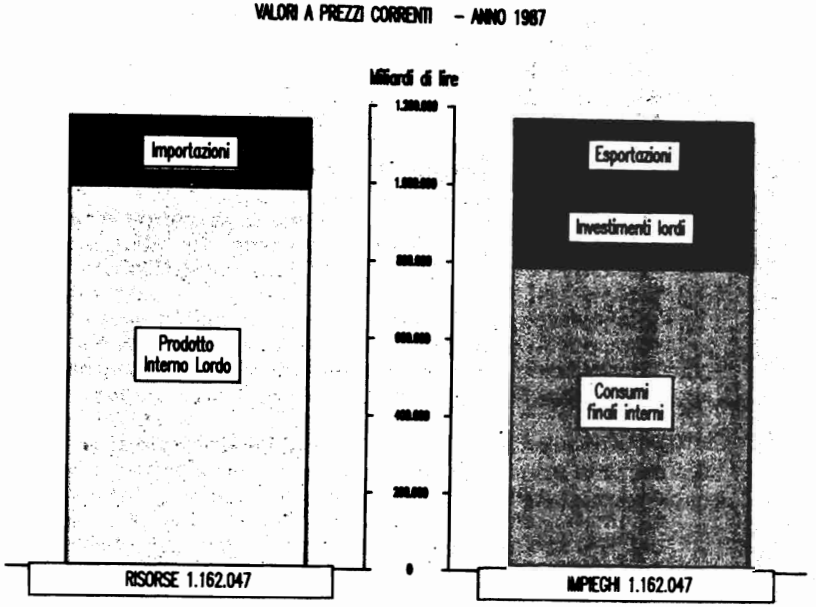


Figure 23. Dot chart.

Fonte: Cleveland e McGill (1984)

Figura 3.34 — Esempio di diagramma a barre suddivise



Fonte: Annuario Statistico Italiano (1988)

CAPITOLO 4 - LA RAPPRESENTAZIONE GRAFICA DELLE SERIE TERRITORIALI

1. CONSIDERAZIONI INTRODUTTIVE

La rappresentazione grafica delle serie territoriali riveste una particolare importanza, dal momento che tale tipo di dati statistici riesce generalmente difficile da interpretare senza l'ausilio di un adeguato strumento grafico.

Nella tabella statistica relativa a serie territoriali è implicitamente contenuta l'informazione sull'organizzazione spaziale delle unità territoriali di riferimento, che il grafico ha il compito di rendere esplicite, attraverso un modello analogico dello spazio. Ciò è tanto più importante quanto più complesso è il sistema di zone utilizzato.

In questo capitolo l'interesse sarà rivolto pressoché esclusivamente alle serie territoriali di tipo areale, riferite cioè a zone di area non nulla e si tratterà di dati relativi a suddivisioni del territorio in zone ampie, sufficienti comunque alla rappresentazione delle principali caratteristiche delle varie distribuzioni, a livello nazionale e regionale.

Queste serie territoriali, come vedremo nel successivo paragrafo 2, si presentano per lo più come distribuzioni di frequenze, ammontari o quantità derivate, associate, da una parte, alle modalità di un carattere e dall'altra, ovviamente, alle varie zone in cui il territorio è suddiviso.

Le tecniche di rappresentazione di tali serie devono soddisfare requisiti di rilevanza statistica ed accuratezza nella resa grafica e facilitare l'accesso all'informazione al più vasto pubblico possibile; l'importanza dei temi attinenti al territorio preme per una attenta selezione dei metodi grafici, al fine di semplificare e di preservare ad un tempo la ricchezza informativa dei dati a disposizione.

Su un piano metodologico generale, si può riprendere la critica rivolta da Tukey (1979) a quel tipo di cartogrammi, da lui definite «patch map», o carte a mosaico, che si limitano a colorare in modo brillante ampie zone del territorio, fornendo un dato che risulta essere uniforme all'interno di ampie regioni e che presenta invece salti bruschi da una regione all'altra.

Per chiarire gli aspetti metodologici connessi a tale tipo di rappresentazioni si può aggiungere che esse danno la visualizzazione della classificazione delle zone in cui è stato suddiviso il territorio secondo classi di valori; se la variabile di partenza è di tipo quantitativo ciò significa una perdita di informazione; la semplificazione dell'informazione che viene così raggiunta può essere, in alcuni contesti, espressamente richiesta.

Un altro aspetto importante da sottolineare per le rappresen-

tazioni di quantità associate a zone è la scarsa significatività, in generale, di rappresentare direttamente i valori assoluti; dal momento che questi sono fortemente dipendenti dalla ampiezza delle varie zone, sul grafico può essere difficile distinguere fra le caratteristiche del fenomeno e la variazione della superficie da zona a zona; per evitare ciò si possono rapportare i valori assoluti ad altre quantità, come ad esempio la superficie o la popolazione delle varie zone, mettendo così in evidenza le peculiarità del fenomeno.

Per superare il problema dell'appiattimento territoriale dell'informazione, si può utilizzare, per i formati usuali delle pubblicazioni Istat, dati a livello comunale; questi sembrano consentire per i cartogrammi relativi all'intera Italia e, presumibilmente, per quelli a livello regionale, una risoluzione spaziale accettabile per i fenomeni d'interesse.

Se si dovesse pensare a cartogrammi relativi a zone subregionali, si potrebbe ricorrere ad un dettaglio spaziale superiore a quello comunale, ad esempio le sezioni di censimento.

Di notevole importanza in campo cartografico sono le applicazioni del metodo delle curve isometriche, che richiedono spesso il passaggio da dati associati a zone o punti distribuiti in modo irregolare sul territorio a valori associati ai punti di un sistema regolare; l'utilizzazione di tale metodo è stato oggetto di notevoli critiche nelle applicazioni alle distribuzioni di densità di popolazione, a causa del carattere discontinuo di tali fenomeni (Unwin 1986).

Un aspetto importante per la rappresentazione grafica delle serie territoriali è l'informatizzazione del processo di produzione dei cartogrammi insieme con una efficace gestione dei dati a livello comunale.

Nella produzione automatizzata dei cartogrammi il passaggio ad un sistema regolare di zone rende più semplice ed efficace l'elaborazione dei dati e la loro rappresentazione.

Il riferimento ad un sistema regolare di zone consente di confrontare fra loro serie territoriali relative ad insiemi di zone differenti e di poter utilizzare graficamente informazioni anche incomplete relative a dettagli spaziali più spinti. Di alcune semplici proprietà di un sistema regolare di zone si discuterà nel seguente paragrafo 3.

Per preservare l'informazione quantitativa contenuta nei dati di partenza occorre rivolgersi all'uso di un adeguato sistema di simboli grafici. Per rendere dati di tipo quantitativo va utilizzata la variabile visiva dimensione, come si è argomentato nel paragrafo 3 del Capitolo 2. Ciò prescinde dalla scala della rappresentazione, come è esemplificato nelle figure 4.1 e 4.2; qui sono riprodotte due applicazioni di Bertin (1983), l'una relativa ad una sequenza di cartogrammi della popolazione della penisola iberica in alcuni anni dal 1530 al 1960 e l'altra relativa alla popolazione di un quartiere di

Parigi.

Al fine di rendere graficamente i dati quantitativi, in Istat è stato adottato il sistema dei Graphical Rational Patterns descritto nel paragrafo 5 del Capitolo 2; di esso è consigliabile l'uso in una sequenza ridotta per semplificare la lettura e rendere più semplice la rappresentazione, specialmente quando il formato del grafico è piccolo.

Un aspetto peculiare della rappresentazione grafica delle serie territoriali consiste nel fatto che una delle due componenti della informazione è di per sé costituita dal territorio di riferimento, la cui rappresentazione occupa entrambe le dimensioni del piano.

Per preservare la possibilità di una visione d'insieme della distribuzione conviene limitare la rappresentazione grafica ad una sola modalità del carattere.

La rappresentazione della serie secondo le varie modalità richiede perciò l'uso di più grafici da disporre in una sequenza "parallela" di cartogrammi; la sequenza può far riferimento al tempo, cioé ai diversi anni cui si riferisce il fenomeno, come nell'esempio della figura 4.1, ripresa dal Bertin, oppure alle diverse modalità di un carattere. Come esempio ulteriore si può considerare la popolazione residente in Italia alla data dei censimenti, come viene presentata nel Capitolo 1.1 dell'Atlante Statistico Italiano (1988).

L'avere preservato il carattere quantitativo dell'informazione consente di andare al di là della semplice rappresentazione delle serie originarie; ciò infatti dà la possibilità di confronti fra serie diverse, operando sui valori associati a ciascun punto del reticolo e di procedere ad ulteriori rappresentazioni grafiche, in un sistema semplice e coerente.

Ciò consente anche di associare, in modo consistente, alla rappresentazione della distribuzione sul territorio la riproduzione grafica di parametri statistici come il baricentro e gli assi di dispersione, che di per sé consentono di valutare alcune principali caratteristiche della distribuzione.

L'obiettivo è quello di proporre un sistema di rappresentazione di tipo grafico-numerico coerente e flessibile, in cui gli aspetti metodologici si intrecciano organicamente con gli aspetti più propriamente grafici.

Nel paragrafo 4 che segue, si tratterà della presentazione di distribuzioni di popolazione; nel paragrafo 5 si descriverà e si fornirà un'interpretazione dei cartogrammi delle differenze fra due distribuzioni di popolazione; nel paragrafo 6 si discuterà della rappresentazione dei rapporti fra frequenze ed ammontari relativi a due distribuzioni di popolazione; infine, nel paragrafo 7 si darà un cenno sulla possibilità di rappresentare classificazioni di zone del territorio, secondo caratteri qualitativi di tipo ordinabile e non, utilizzando come variabili visive la variazione di valore e di colore; si

darà anche un cenno alla possibilità di rappresentare dati quantitativi con l'uso appropriato dei Graphical Rational Patterns in riferimento ad un sistema arbitrario di zone.

Affinché tale sistema possa essere messo in opera è necessario disporre delle coordinate relative alle entità spaziali da rappresentare; per lo sviluppo di tale sistema l'Istat ha adottato le coordinate predisposte dalla società ITALECO, ottenute previa elaborazione delle coordinate Gauss-Boaga relative alla cartografia ufficiale dell'Istituto Geografico Militare alla scala di 1:100.000.

Per le indicazioni fondamentali sulla cartografia ufficiale italiana si può far riferimento al testo di Franchi (1950).

2. LA CLASSIFICAZIONE DELLE SERIE TERRITORIALI

Ciò che caratterizza la raccolta ed il trattamento dei dati territoriali è il riferimento di questi allo spazio; tale raccordo può essere realizzato secondo modi diversi.

Un appartamento ad uso residenziale od una unità locale possono essere referenziati tramite il loro indirizzo, come anche dal numero mappale attribuito dal Catasto alla particella che li contiene.

Una famiglia può essere associata, a sua volta, all'appartamento in cui vive o semplicemente alla sezione di censimento o al Comune di residenza.

Quale che sia il livello di risoluzione del fenomeno in studio, nel dato territoriale è comunque riscontrabile una coppia di elementi messi in relazione fra loro: un elemento spaziale ed un elemento descrittivo (Bonfatti, 1985).

L'elemento spaziale è generalmente caratterizzato da un insieme di coordinate relative a punti che ne definiscono la posizione e la forma nello spazio; tradizionalmente i dati territoriali vengono classificati rispetto all'elemento spaziale secondo le tre forme seguenti: punti, linee ed aree.

Di particolare interesse per la nostra trattazione è il caso di dati riferiti a zone che costituiscono una partizione di un territorio, determinato a priori in base a criteri convenzionali; ad esempio il territorio italiano. Per partizione intendiamo una suddivisione di tale territorio in un insieme di zone che risultano disgiunte fra loro e che ricoprono l'intero territorio di riferimento. Ad esempio si può pensare alla suddivisione dell'Italia in Regioni, Province, Comuni, oppure anche ad un sistema regolare di zone di forma ed area uguale.

Il processo di aggregazione per i dati territoriali può avvenire rispetto ad entrambi gli elementi che li definiscono; rispetto all'elemento spaziale, si può passare, ad esempio, da dati assegnati per punti, a dati assegnati per zone, da informazioni associate ad un numero di zone maggiore, ad un numero di zone minore, per aggregazione delle zone iniziali. È in particolare quest'ultimo tipo di aggregazione che ha suscitato un notevole interesse da parte di statistici e geografi.

Come esempi di dati territoriali riferiti a punti si può far riferimento alle serie associate all'insieme dei porti, od aeroporti del territorio nazionale, ai valichi di frontiera, alle dighe di un bacino idrografico. Come esempi di dati geografici associati a linee si può far riferimento alle misure di flusso lungo certe direzioni, quali strade od autostrade e agli spostamenti fra zone prefissate, quali i fenomeni di pendolarismo e di migrazione.

Per completezza si ricorda come ulteriore tipo di dato geografico, quello determinato dai fenomeni che si distribuiscono in modo continuo sul territorio; ad esempio l'altitudine, le precipitazioni atmosferiche, le temperature minime o massime medie giornaliere in un mese costituiscono serie territoriali di questo ultimo tipo.

Per quanto riguarda la componente descrittiva si vuole premettere la seguente considerazione a carattere generale.

La componente descrittiva si può limitare ad un semplice codice identificativo, ad esempio ad un codice binario (urbano, non urbano) stabilito attraverso analisi statistiche preliminari o attraverso delle convenzioni. In questi casi si può dire che nella rappresentazione delle serie territoriali prevale la componente geometrica dell'informazione; molte carte tematiche attualmente prodotte possono ricondursi a dati territoriali di questo tipo, che si limitano ad identificare sul territorio componenti spaziali determinate, supposte omogenee al loro interno.

Ai fini della classificazione delle serie territoriali la componente descrittiva può essere opportunamente distinta secondo lo schema ben noto:

$$\text{qualitativa} \begin{cases} \text{non ordinabile} \\ \text{ordinabile} \end{cases}$$

$$\text{quantitativa.}$$

Tale informazione si può presentare come un attributo direttamente associato all'entità spaziale.

Nel caso di suddivisione spinta del territorio, che è un campo di rappresentazione proprio della cartografia tradizionale, ciascuna zona può essere caratterizzata da un codice che le attribuisce una modalità di un carattere qualitativo; ad esempio per il carattere

«uso del suolo» possiamo avere la modalità «ad uso residenziale», oppure «destinato a coltivazione» con l'ulteriore specificazione del tipo di coltivazione. Alle zone può anche essere attribuita una modalità relativa ad un carattere quantitativo come, ad esempio per un terreno agricolo, il reddito prodotto in un anno.

Nel caso di serie territoriali qualitative il carattere può essere di tipo ordinabile o sconnesso; la classificazione dell'uso del suolo è certamente qualitativa non ordinabile; in certe classificazioni del catasto rurale, che per ogni tipo di coltura individuano i territori migliori cui viene attribuita la prima classe ed i peggiori, cui si attribuisce l'ultima classe, intercalando le classi intermedie, è riscontrabile un carattere qualitativo ordinabile.

A livello microgeografico, il carattere omogeneo delle zone consente, nel caso di caratteri qualitativi, la rappresentazione congiunta delle diverse modalità del carattere.

Le persone, le famiglie, le unità locali o altre entità, distribuite sul territorio, ma concettualmente autonome rispetto a questo, una volta che siano state poste in relazione con entità spaziali, vengono a costituire l'elemento descrittivo del dato territoriale; ad esse si possono associare caratteri qualitativi o quantitativi rispetto ai quali le entità di descrizione vengono generalmente aggregate. In effetti, in questo caso, i dati territoriali vengono per lo più forniti come frequenze, ammontari o quantità derivate associate alle diverse entità spaziali.

A livello macrogeografico, d'altronde, si perde quasi necessariamente l'omogeneità delle zone; in questo caso anche i dati territoriali, che a livello microgeografico associano alle singole entità spaziali una ben precisa modalità di un carattere, una volta aggregate vengono a formare delle serie territoriali costituite da frequenze o ammontari associati alle varie modalità.

Ad esempio per la classificazione del territorio secondo l'uso del suolo, la serie territoriale può associare a ciascuna zona, ottenuta per aggregazione dalle zone elementari, la percentuale di superficie associata a ciascun tipo di uso.

Con il termine «serie territoriali» intenderemo essenzialmente dati come frequenze, ammontari o quantità derivate (valori medi, quozienti, percentuali) riferite alle modalità di un carattere ed associate ad entità spaziali ben definite, di tipo macrogeografico. Nella maggior parte dei casi il sistema di raccolta dei dati e di elaborazione in uso all'Istat consente il riferimento dei dati ad entità territoriali come Comuni, Province o Regioni.

Si vuole sottolineare come l'aggregazione dell'elemento spaziale comporti inevitabilmente una perdita di informazione associata alla risoluzione spaziale dei fenomeni; nel passaggio dal dettaglio comunale a quello provinciale il dato viene uniformemente distribuito sull'intera provincia con notevole perdita d'informazione.

In tale processo va tenuto conto di un altro importante aspetto, approfondito in modo particolare da Openshaw (1987), relativo alla dipendenza dei risultati delle analisi non solo dal livello di aggregazione ma anche dal sistema di zone prescelto, sovente arbitrariamente, a quel dato livello di risoluzione spaziale.

Per ovviare a questi problemi conviene preservare, per quanto possibile, il dettaglio spaziale, cioé usare nella rappresentazione il dettaglio comunale fino a quando il formato del grafico lo consente.

Nello stesso tempo conviene riferire il dato ad un sistema regolare di zone per favorire la corretta rappresentazione della distribuzione con l'uso di simboli appropriati, e per facilitare i confronti fra fenomeni diversi.

Allorquando, nel processo di aggregazione, si preserva l'omogeneità delle zone anche a livello macrogeografico è corretto pensare alla rappresentazione delle modalità del carattere associandole direttamente alle zone iniziali.

3. L'USO DI UN SISTEMA REGOLARE DI ZONE

L'estrema variabilità della superficie dei Comuni italiani (da 1.507,60 chilometri quadrati del Comune di Roma ai 0,1 chilometri quadrati di Atrani in provincia di Salerno), la base storica e convenzionale con cui essi sono stati definiti, che sovente non corrisponde ad una reale omogeneità del territorio, costituisce un ostacolo ad una razionale rappresentazione delle serie territoriali basate su dati comunali.

Inoltre per i cartogrammi da produrre in formato ridotto, la rappresentazione della fitta rete dei confini comunali verrebbe a costituire una interferenza con la componente essenziale della informazione costituita dall'elemento descrittivo.

L'esigenza di una gestione semplice ed efficace dell'informazione, l'opportunità di rendere più regolari le forti variazioni sovente associate ai fenomeni territoriali, (ad esempio la densità di popolazione nel 1981 in termini di abitanti per chilometro quadrato varia dai 17.750 di Portici in provincia di Napoli al solo abitante per chilometro quadrato di Rhêmes-Notre Dame in Val D'Aosta), la necessità di utilizzare un sistema standard di simboli grafici, al fine di preservare il carattere quantitativo dell'informazione, l'importanza, infine, di rendere facilmente realizzabile il confronto fra distribuzioni differenti, sono argomentazioni a favore dell'utilizzazione di un sistema regolare di zone come base di riferimento per la rappresentazione grafica delle serie territoriali.

Inoltre l'uso di una griglia regolare di punti (centri delle zone), consente di leggere i simboli che rappresentano gli ammontari

delle popolazioni distribuite sul territorio contemporaneamente come valori assoluti (e relativi) e come valori di densità. Il problema del rapporto con l'area delle zone è infatti implicitamente risolto dalla uguaglianza delle aree associate a ciascun punto ed è sufficiente fornire una legenda aggiuntiva con i valori assoluti divisi per l'area costante di ciascuna zona per avere la lettura diretta dei valori di densità.

L'applicazione di un sistema regolare di punti in campo statistico per la rappresentazione grafica di popolazioni finite può fare riferimento ai lavori di Bertin (1983) e Bachi (1984).

La scelta del numero dei punti è collegata in modo opposto alle due seguenti considerazioni. Da una parte il dettaglio di risoluzione spaziale del dato è direttamente proporzionale al numero dei punti: quanto questo è maggiore, tanto più raffinata, se l'informazione di base lo consente, è la precisione della collocazione del fenomeno nello spazio. D'altra parte una crescita del numero dei punti della griglia comporta, a parità di formato, il progressivo accentuarsi della densità dei punti e della difficoltà di distinguerli singolarmente e, di conseguenza, di operare confronti di tipo quantitativo fra i valori assunti dalla serie nelle varie zone.

Esemplificazioni sono riportate in proposito da Bertin (1983). Nella figura 4.3 viene riprodotta una serie di cartogrammi del territorio della Francia (circa 551 mila kmq), sui quali viene distribuito un numero variabile di punti su una griglia quadrata secondo le seguenti tre densità: 1 punto ogni 1.000 chilometri quadrati, 1 punto ogni 250 chilometri quadrati e infine 1 punto ogni 100 chilometri quadrati.

Su ciascun punto della griglia viene sovrapposto un cerchio di ampiezza variabile, proporzionale alla densità della popolazione. Dalla sequenza dei grafici è percepibile la progressiva difficoltà ad operare confronti quantitativi ed il prevalere, in particolare per l'ultimo cartogramma, della percezione di varie gradazioni di grigio associate ai diversi dipartimenti della Francia, essendo il medesimo valore di densità assegnato ai punti appartenenti allo stesso dipartimento.

Se si vuole consentire di operare, in modo semplice, confronti, che preservino la natura quantitativa dei dati, fra i valori assunti nelle diverse zone del territorio, conviene selezionare un numero di punti appropriato in modo da evitare che questi si addensino in modo eccessivo gli uni agli altri, rendendo difficile all'occhio distinguere con chiarezza i simboli associati a ciascuna zona.

Alla considerazione che la scelta del numero dei punti deve tener conto del grado di addensamento di questi, si deve aggiungere la osservazione che tale scelta va anche regolata sul grado di risoluzione spaziale dei dati da elaborare, in quanto un numero di punti eccessivo porterebbe a distribuire in modo uniforme l'informazione su ampie zone del cartogramma.

Per quanto riguarda le applicazioni in Istat si è ritenuto opportuno adottare come soluzione un numero di punti pari a 1.000, con densità di circa 1 punto ogni 300 chilometri quadrati. La perdita di informazione nel passaggio da circa 8.000 valori relativi ai dati comunali ai 1.000 valori relativi ad una griglia regolare non pare eccessivo, tenendo soprattutto conto delle scale consentite dai formati standard delle pubblicazioni Istat.

L'uso di un reticolo con un numero superiore di punti, rimanendo il dettaglio comunale, relativamente all'intero territorio nazionale, ed in riferimento a cartogrammi da presentarsi sui formati delle attuali pubblicazioni Istat non appare giustificabile.

Uno dei vantaggi più importanti nell'uso di un sistema regolare di punti è che l'elemento geometrico del dato è stato semplificato ed al lettore è consentito di concentrarsi sulle caratteristiche della distribuzione territoriale del fenomeno.

Potendo disporre di formati maggiori o produrre delle carte ad una scala maggiore, il numero dei punti potrebbe crescere in modo proporzionale; in tal caso potrebbe divenire comunque opportuno utilizzare informazioni di dettaglio più spinto, come ad esempio i dati a livello di sezioni di censimento, per evitare di produrre rappresentazioni che appiattiscano l'informazione sull'area comunale. Il sistema regolare di punti consentirebbe di risolvere semplicemente l'assegnazione dei contributi di ciascuna sezione ai punti interni all'area comunale, anche in presenza di una acquisizione solo parziale dell'elemento spaziale relativo alle sezioni.

Nelle applicazioni si trovano generalmente proposte due soluzioni per la disposizione dei punti secondo un reticolo regolare: una basata su un reticolo quadrato e l'altra su un reticolo di triangoli equilateri.

Tali possibili soluzioni sono esemplificate nella figura 4.4, dove ciascun punto della griglia è identificabile da una coppia di numeri interi. Ogni reticolo regolare può essere caratterizzato da un parallelogramma, che per traslazione lo genera, e dalla distanza minima fra due nodi del reticolo (Hilbert, Cohn-Vossen 1960).

Intorno a ciascun punto del reticolo quadrato è possibile costruire un quadrato di lato pari alla distanza minima fra centri contigui. Sui centri del reticolo triangolare si possono costruire esagoni regolari di apotema "a", pari a metà della distanza "d" fra due centri contigui; il lato dell'esagono (l) è pari a $2\,a / \sqrt{3} = d / \sqrt{3}$. In questo senso consideriamo associato al sistema regolare di punti il corrispondente sistema regolare di zone.

L'assegnazione dell'informazione iniziale ai punti del reticolo si può effettuare secondo procedure diverse; se l'informazione di partenza è data per punti (centroidi di zone) si distribuisce il dato fra i punti più vicini della griglia, tenendo presente il problema della equivalenza delle aree delle zone iniziali e delle celle del sistema regolare.

Se l'informazione geografica è assegnata tramite il contorno chiuso che delimita la zona di riferimento, si può ripartire l'ammontare in modo uguale fra i punti interni all'area.

Nella figura 4.5 è riportato un possibile esempio del passaggio da un reticolo, quadrato od esagonale, ad un altro reticolo di forma simile ma con zone di area quattro volte quella delle celle di partenza; il processo di aggregazione consiste nell'utilizzare un quarto dei punti del sistema iniziale.

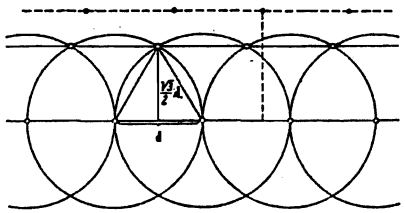
La base standard per la rappresentazione grafica dei cartogrammi da utilizzare nelle applicazioni in Istat è costituita da una griglia regolare di 1.000 celle esagonali; essa viene riprodotta nella figura 4.6. Alcuni aspetti vengono sviluppati nell'introduzione e nelle appendici dell'Atlante Statistico Italiano (1988). Qui viene presentata una tabella che associa ciascun Comune del 1981 ad uno o più esagoni; in una seconda tabella si presenta come gli esagoni sono stati assegnati alle 95 Province italiane.

Si vuole, infine, ricordare una importante proprietà del sistema regolare di punti distribuiti secondo un reticolo di triangoli equilateri. Esso, fra tutti i reticoli regolari di pari distanza minima, è tale che il numero dei punti di esso distribuiti all'interno di una data superficie di ampie dimensioni è il massimo possibile fra quelli relativi ad ogni altro reticolo regolare.

Se con centro in tutti i nodi del reticolo descriviamo delle circonferenze di raggio uguale pari alla metà della distanza minima fra due di essi, otteniamo un sistema di cerchi che si toccano ma non si sovrappongono mai. Un sistema di cerchi costruito in questo modo viene definito strato di cerchi. Si dice che uno strato di cerchi, di raggio assegnato, è tanto più denso quanto è maggiore il numero di cerchi che trovano posto in una regione assegnata. Di conseguenza il sistema proposto del reticolo di triangoli equilateri ci dà lo strato di cerchi di densità massima.

Riprendiamo l'argomentazione svolta in Hilbert, Cohn-Vossen (1960) per dimostrare la proprietà su accennata che il reticolo di triangoli equilateri, fra tutti i reticoli regolari di pari distanza minima, è tale che il numero dei punti di esso distribuiti all'interno di una data superficie di ampie dimensioni è il massimo possibile fra quelli relativi ad ogni altro reticolo regolare. Un qualsiasi reticolo regolare può essere generato da un parallelogramma appropriato, traslandolo lungo la direzione di una sua coppia di lati parallelamente a se stesso, in entrambi i versi, e, quindi, traslando la striscia così ottenuta lungo la direzione definita dall'altra coppia di lati. Una proprietà generale dei reticoli regolari, come dimostrato in Hilbert, Cohn-Vossen (1960), è che ogni reticolo può essere generato da opportuni parallelogrammi aventi area uguale.

Ciò premesso si deve costruire il reticolo di distanza minima pari a d *e che sia generato da un parallelogramma di area minima; consideriamo i punti distanti l'uno dal successivo di tale quantità e disposti lungo una retta; se immaginiamo, come nella figura seguente, di costruire cerchi di raggio* d *e con centro nei suddetti nodi del reticolo, si ottiene una striscia di cerchi. Sulla base del fatto che* d *è la distanza minima, nessun punto del reticolo può cadere all'interno di tale striscia di cerchi.*



*Il reticolo avente distanza minima assegnata "*d*" e generato da un parallelogramma di area minima è quello che possiede come ulteriore sequenza di punti quella determinata dalle intersezioni delle circonferenze; il reticolo di punti così generato è un reticolo di triangoli equilateri.*

La proprietà su specificata deriva quindi dall'ipotesi che il numero dei punti del reticolo distribuiti su un dato territorio può ritenersi proporzionale al rapporto fra l'area del territorio e l'area del parallelogramma generatore.

4. LA RAPPRESENTAZIONE GRAFICA DI DISTRIBUZIONI DI POPOLAZIONE SUL TERRITORIO — PARAMETRI STATISTICI CONNESSI

Indichiamo con j la generica zona iniziale di riferimento dei dati territoriali; j varierà da 1 a n, dove n = 95 per le Province, e n = 8086 per i Comuni nel 1981. Si indica con i, per i = 1, ..., N, la generica cella esagonale del sistema regolare di zone in cui è stato suddiviso il territorio italiano; tale indice nella procedura standard in uso all'Istat, come dinanzi accennato, varia da 1 a 1000. La suddivisione, effettivamente impiegata, è riportata nella fig. 4.6 del paragrafo precedente.

A ciascuna cella esagonale i rimane associato il suo centro, di coordinate (X_i, Y_i), tramite le quali sarà possibile calcolare importanti parametri statistici connessi alla distribuzione della popolazione in studio.

Indichiamo con $\tilde{A}$ una generica distribuzione di popolazione sul territorio dell'Italia; essa può consistere di frequenze od ammontari, che supponiamo già associati alle singole celle esagonali i, con i = 1 ... 1000; le frequenze od ammontari associati alle zone i vengono indicati con A_i; l'ammontare complessivo è

$$A = \sum_{i=1}^{1000} A_i$$

I valori delle frequenze relative o degli ammontari relativi in i sono indicati da $a_i = A_i/A$.

Per la generica distribuzione di popolazione $\tilde{A}$ una possibile soluzione per la sua rappresentazione sarebbe di calcolare il valore massimo e minimo degli A_i, escludendo eventualmente dei valori eccezionali; si potrebbe quindi ricondurre l'intervallo contenuto fra questi estremi al campo di variazione da 1 a 100 ed approssimare, infine, i valori così trasformati all'intero più vicino.

Tuttavia ad una tale procedura conviene preferire una soluzione standard a carattere generale: quale che sia la distribuzione degli A_i, si riporta il totale delle frequenze o degli ammontari ad un valore costante, pari ad M. Ciò consente di operare confronti significativi fra distribuzioni territoriali differenti, fissando l'attenzione sulla forma delle varie distribuzioni, evitando che i diversi ammontari possano influenzare la percezione delle caratteristiche geografiche di queste.

Codesta è una pratica abbastanza usuale in statistica; ne abbiamo vista una esemplificazione a proposito delle piramidi dell'età e del loro confronto. La distribuzione territoriale che si ottiene riducendo l'ammontare complessivo delle A_i ad un valore M dato ed approssimando tali quantità ai valori interi più vicini viene definita la distribuzione statistica equivalente ed indicata con (A).

Le frequenze od ammontari ad essa pertinenti sono pari a:

$M_i = [a_i \times M]$, dove con [·] si indica la funzione «intero più vicino».

Per il valore di M sono possibili scelte diverse; esse dipendono, fra l'altro, dal tipo di scala di GRP prescelta al fine di rappresentare le serie. Sulla base di esperienze effettuate, viene proposto di utilizzare, insieme con la scala di GRP che va da 1 a 100, il valore di M = 10.000 e di associare direttamente a ciascun valore M_i il simbolo s_i secondo una ben definita funzione.

Può essere utile riportare un esempio, per spiegare come avviene la determinazione del simbolo da associare alla generica cella i. Si calcola la frequenza od ammontare relativo alla distribuzione equivalente (A) nella cella i; esso è pari all'intero più vicino a $a_i \times M$, che si indica M_i; tale valore ci fornisce direttamente il simbolo grafico razionale da rappresentare nella cella i. Nel caso della popolazione residente al 1.1.1987, se nella cella i A_i = 26.993, riesce (26.993/57.290.000) x 10.000 = 4,72, dove A = 57.290.000 è la popolazione residente in Italia, e quindi $M_i = 5$; si pone quindi il simbolo grafico da rappresentare $s_i = 5$.

Dal momento che il totale della popolazione equivalente è pari a 10.000, l'unità verrà a rappresentare 5.729 individui; sulla base di tale valore è possibile costruire l'intera legenda dei simboli grafici razionali (GRP).

Si vuole osservare che i valori assoluti sono immediatamente trasformabili in valori relativi (attribuendo ai simboli i loro valori interi, divisi per 10.000), come anche in valori esprimenti la densità di abitanti per kmq. A tale fine basta dividere i valori assoluti per 301 kmq, corrispondenti all'area costante delle celle esagonali.

Anche in questo metodo di rappresentazione, pur molto preciso, esiste un'approssimazione che dipende dal valore di M e dalla scala dei simboli prescelta. Nell'applicazione descritta, per i valori rappresentati dai simboli compresi fra 1 e 100, tale approssimazione è al massimo pari a 0,5 X A/10.000; nel caso della popolazione del 1987 l'errore è pari 2.862 individui. Ciò significa, in primo luogo, che le celle con $A_i < 2.862$, ovvero con densità inferiore a 9,5 abitanti per kmq, risultano prive dei simboli grafici.

Tale errore assoluto è costante rispetto al livello del dato rappresentato, e viene, perciò, ad esserci notevole disparità fra gli errori relativi dei valori bassi e quelli corrispondenti ai valori più elevati della serie. Sulla base di questa considerazione e avendo anche l'obiettivo di semplificare la lettura della legenda e la realizzazione del cartogramma, è stata concepita una scala ridotta consistente nella sequenza riprodotta nella figura 4.7; essa riporta i simboli da 1 a 15 con salti unitari e i simboli dal 20 al 100 (che rende i valori 100 e più) con le decine. Un'analisi accurata di tale scala ridotta viene condotta nella introduzione dell'*Atlante Statistico Italiano 1988*, ove essa è estesamente utilizzata. Nella figura 4.8 viene riprodotto il cartogramma della popolazione residente italiana al 1.1.1987, che utilizza la su descritta scala. In essa appaiono anche i due assi che si intersecano perpendicolarmente e di cui parleremo tra poco.

La flessibilità del sistema consente di scegliere la rappresentazione, oltre che in funzione della scala prescelta, anche in funzione delle caratteristiche della distribuzione delle A_i; ad esempio, per una distribuzione la cui densità di occupazione per cella si concentra intorno al valore medio (A/1000), si può far coincidere tal valore medio con il simbolo 50, posto al centro della scala standard completa. Ciò consente di differenziare più chiaramente i valori.

Un altro aspetto della flessibilità del sistema è che, quando si è interessati ad evidenziare l'esistenza di ammontari complessivi diversi e, quindi, a confronti di tipo assoluto fra popolazioni diverse, si può far variare il valore di M da una distribuzione all'altra. Tale tema verrà ripreso nel successivo paragrafo 5; qui ci limitiamo a far presente due possibili esempi di uso di M diversi. In una serie territoriale-cronologica, come la popolazione residente in Italia alla data dei censimenti, si può voler mantenere nelle rappresentazioni il peso complessivo effettivo di ciascuna popolazione in

confronto alle altre, per dare al lettore l'idea del processo di crescita in valore assoluto della popolazione, oltre che del suo progressivo concentrarsi in certe zone del territorio; questo esempio è riportato nei primi 12 cartogrammi dell'*Atlante Statistico Italiano 1988*.

Nella rappresentazione delle varie componenti di un certo fenomeno, come ad esempio la distribuzione del territorio italiano secondo i principali tipi di utilizzazione: 1) seminativi, 2) coltivazioni legnose-agrarie, 3) colture foraggere permanenti, 4) boschi, 5) altre colture, 6) superficie non produttiva, se si vuole evidenziare la differenza relativa di utilizzazione del territorio, si possono riproporzionare le varie distribuzioni ad ammontari complessivi proporzionali agli ammontari effettivi delle distribuzioni.

L'avere preservato il carattere quantitativo dell'informazione consente di arricchire il cartogramma della distribuzione territoriale con la rappresentazione grafica di importanti e significativi parametri statistici. In particolare conviene fissare l'attenzione su alcune misure sintetiche di localizzazione e di dispersione della serie, resi graficamente sul cartogramma della fig. 4.8.

Fra i possibili parametri statistici atti a fornire una misura di localizzazione della distribuzione, per la sua semplicità e per alcune sue proprietà è da menzionare il baricentro.

Ricordiamo che con (X_i, Y_i) indichiamo le coordinate dei centri delle celle esagonali di indice i e che a_i misura le frequenze od ammontari relativi della distribuzione della popolazione $\bar{A}$ nelle celle i. Il baricentro della distribuzione ha le seguenti coordinate:

$$\overline{X} = \sum_{i=1}^{1000} a_i X_i \qquad \overline{Y} = \sum_{i=1}^{1000} a_i Y_i$$

Per misurare la dispersione (Bachi 1962, 1984), si può utilizzare la distanza-varianza, definita dalla seguente formula:

$$d^2 = \sum_{i=1}^{1000} a_i \left[(X_i - \overline{X})^2 + (Y_i - \overline{Y})^2 \right];$$

Se consideriamo tale espressione come funzione delle coordinate del punto da cui si calcolano le distanze dei centri degli esagoni, essa risulta minima quando per tale punto viene scelto il baricentro.

Si vuole ricordare alcuni semplici proprietà di tali parametri.

Per il baricentro ricordiamo che: a) le somme ponderate degli scarti delle coordinate (X_i, Y_i) dalle corrispondenti coordinate del baricentro $(\overline{X}, \overline{Y})$ è nulla, separatamente per le X e per le Y; b) le coordinate del baricentro sono invarianti rispetto ad arbitrarie suddivisioni del territorio, nel senso che il baricentro di una distribuzione su un dato territorio può essere direttamente calcolato dai centri delle 1000 celle esagonali, come anche dai punti che sono a loro volta i baricentri della medesima distribuzione relativamente a zone che costituiscono una partizione esaustiva del territorio, attribuendo a tali punti pesi determinati dagli ammontari della distribuzione nelle rispettive zone.

Quest'ultima proprietà è particolarmente vantaggiosa dal momento che consente di passare da analisi parziali ad analisi globali con relativa semplicità.

Occorre tuttavia una certa attenzione nell'interpretazione del baricentro, dal momento che esso può avere comportamenti non gradevoli, come nel caso di popolazioni distribuite su territori concavi per le quali può accadere che il baricentro cada al di fuori dei confini del territorio di riferimento; allo stesso modo nel caso di popolazioni composte da grappoli, ovvero che si concentrano in certe zone del territorio distanti fra loro, è possibile che il baricentro si trovi in una zona dove il fenomeno ha scarsa consistenza.

Una proprietà del baricentro, che, a seconda delle circostanze, può talvolta essere utile, altre volte spiacevole, è la sua sensibilità allo spostamento anche di pochi elementi della popolazione sul territorio, all'aggiunta di nuovi elementi e alla presenza di dati anomali, che possono allontanare in modo marcato il baricentro dal cuore della distribuzione, verso la periferia del territorio.

Per ovviare a tali inconvenienti sono state definite altre misure di localizzazione, sulla base di opportune ipotesi di minimo; tuttavia non è questa la sede per affrontare tale analisi e qui basta avervi fatto cenno.

La distanza-varianza consiste, come risulta dalla espressione su riportata, nella somma ponderata dei quadrati delle distanze dei centri delle singole celle i dal baricentro $(\overline{X}, \overline{Y})$; una delle sue più interessanti proprietà è la possibilità di scomporla in modi differenti.

Nel caso in cui si siano individuati dei clusters (o grappoli), ovvero delle zone del territorio in cui la popolazione è prevalentemente concentrata, si può utilizzare, al fine di valutare il livello dell'accentramento, un indice ricavabile dalla scomposizione della distanza-varianza in una parte dovuta alla dispersione della popolazione all'interno dei vari grappoli ed in una parte dovuta alle distanze, opportunamente ponderate, dei baricentri dei singoli grappoli dal baricentro della distribuzione nel suo complesso.

È possibile una scomposizione della distanza-varianza che interessa direttamente la rappresentazione grafica delle distribuzioni equivalenti,come quella della figura 4.8.

A tal fine è necessario dare alcune definizioni preliminari.

Definiamo le seguenti due quantità:

$$\sigma^2_x = \sum_{i=1}^{1000} a_i (X_i - \overline{X})^2$$

$$\sigma^2_y = \sum_{i=1}^{1000} a_i (Y_i - \overline{Y})^2;$$

esse misurano la distanza-varianza della distribuzione (A) lungo l'asse x e y rispettivamente.

Sulla base di tali posizioni la distanza-varianza è immediatamente scomponibile nel modo seguente:

$$d^2 = \sigma^2_x + \sigma^2_y$$

Si può dimostrare che d^2 è invariante rispetto alle trasformazioni di coordinate ottenute per arbitrarie rotazioni intorno al baricentro degli assi del sistema di riferimento iniziale; la stessa proprietà non vale singolarmente per le distanze-varianze σ^2_x e σ^2_y; esse, infatti, variano in modo tale che, ruotando gli assi di riferimento, mentre l'una aumenta l'altra diminuisce, o viceversa (Bachi 1984).

Può essere utile fornire le direzioni degli assi in corrispondenza delle quali, per una di esse, la distanza-varianza di (A) raggiunge il suo valore massimo e, quindi, per l'altra, di converso, assume il suo valore minimo; tali due direzioni, fra loro perpendicolari, consentono di identificare le direzioni di massima e minima dispersione della distribuzione (A).

Tali direzioni vengono definite gli assi principali della distribuzione e indicate con x' e y'; le rispettive distanze-varianze sono indicate da $\sigma^2_{x'}$ e $\sigma^2_{y'}$. L'angolo α (Bachi 1984), di cui bisogna ruotare gli assi di partenza per raggiungere la posizione degli assi principali, è dato dalla seguente formula:

$$2\alpha = \text{arctg} \frac{2 \sum_{i=1}^{1000} a_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sigma^2_x - \sigma^2_y}$$

dove si sceglie $0^0 \leqslant |\alpha| < 45^\circ$ e si suppone $\sigma^2_x \neq \sigma^2_y$. La rotazione è da effettuare in verso antiorario od orario a seconda del valore rispettivamente positivo o negativo di α.

Nella rappresentazione grafica della fig. 4.8 sono stati per l'appunto rappresentati tali assi, disposti sul cartogramma a forma di croce. Questa è formata nel modo seguente: si fissa il baricentro della distribuzione; si calcola l'angolo α fra gli assi convenzionali e i due assi principali; su ciascuno di essi, a partire dal baricentro, per entrambi i versi si tracciano due segmenti di lunghezza $2\sigma_{x'}$ e $2\sigma_{y'}$ rispettivamente.

5. LA RAPPRESENTAZIONE GRAFICA DELLE DIFFERENZE DI DUE DISTRIBUZIONI DI POPOLAZIONE SUL TERRITORIO

L'uso dei cartogrammi razionali consente di andare al di là di una semplice rappresentazione delle serie originarie. Sul piano formale diventa possibile confrontare direttamente i cartogrammi di serie differenti operando sui valori associati alle zone e di procedere ad ulteriori rappresentazioni grafiche di rilevanza statistica. In questo paragrafo si studierà il cartogramma delle differenze fra le frequenze o ammontari delle due distribuzioni; nel prossimo paragrafo si tratterà del rapporto tra tali quantità.

Indichiamo con $\tilde{A}$ e $\tilde{B}$ le due distribuzioni di popolazione di riferimento; con A_i e B_i le frequenze od ammontari in valore assoluto, relativi alle celle esagonali i; con A e B i rispettivi ammontari complessivi; con a_i e b_i gli ammontari relativi, associati alle celle esagonali i. Indichiamo con (A) e (B) le due distribuzioni equivalenti corrispondenti.

Talvolta può essere interessante valutare le differenze delle frequenze od ammontari in valore assoluto. Ciò è possibile se le due distribuzioni territoriali, associate alle medesime zone, si riferiscono a due modalità, relative allo stesso carattere o caratteri differenti, che vengono misurate nella stessa unità di misura.

Ad esempio si possono confrontare le superfici misurate in ettari delle aziende agricole secondo una modalità di utilizzazione (ad esempio «superficie coltivata a seminativi») per due anni diversi; il carattere «superficie secondo la forma di utilizzazione» stabilisce, in questo caso, l'unità di misura delle grandezze da confrontare; il riferimento temporale costituisce il carattere, avente per modalità gli anni, che risulta differente fra le due distribuzioni territoriali.

È certamente possibile calcolare le differenze assolute fra le distribuzioni della popolazione residente in Italia fra due anni diversi, quando queste sono riferibili allo stesso sistema di zone.

Un altra coppia di distribuzioni per cui c'è la possibilità di calcolare tali differenze assolute sono le distribuzioni della popolazione residente, in un certo anno, dei maschi e delle femmine. In questo caso il carattere che assume modalità diverse fra le due distribuzioni è il sesso.

Negli esempi su riportati le differenze possono essere calcolate fra i valori originari delle due distribuzioni, zona per zona, e quindi successivamente rapportate ad un opportuno ammontare complessivo. Si può operare anche sulle distribuzioni equivalenti (A) e (B), riproporzionando opportunamente gli ammontari complessivi delle due distribuzioni; indicando con M_A e M_B gli ammontari delle distribuzioni riproporzionate, deve valere la relazione:

$$M_A : M_B = A : B$$

Per stabilire, comunque, il valore effettivo dell'ammontare complessivo delle diverse distribuzioni, occorre fissare l'ammontare complessivo di una popolazione di riferimento, che può essere anche diversa dalle due coinvolte nel confronto.

Con elevato grado di approssimazione le differenze possono essere calcolate direttamente tramite le due distribuzioni equivalenti (A) e (B), riproporzionate con i rispettivi fattori di scala.

I cartogrammi che rappresentano tali differenze richiedono l'uso di simboli diversi per le quantità negative e per quelle positive; si può a tal fine utilizzare in modo appropriato il colore.

Nella figura 4.9 si riprende, a mo' di esempio, il cartogramma del confronto di tipo assoluto fra le distribuzioni degli addetti alle industrie tessili ai due censimenti del 1981 e 1971; l'ammontare degli addetti nel 1971 viene rapportato a 10.000 unità, quello relativo al 1981 è ricondotto a 9.123 unità, dal momento che l'ammontare effettivo complessivo delle due distribuzioni è pari rispettivamente a 540 mila e 493 mila addetti.

La quantità rappresentata in ciascuna cella i è pari ($M_A \cdot a_i - M_B \cdot b_i$) in rosso se positiva, in verde se negativa; la scala del grafico è determinata dall'aver riproporzionato la popolazione del 1971 a 10.000.

La medesima procedura è stata utilizzata per la rappresentazione della serie storica della popolazione italiana dal 1871 al 1987, riportata nei primi grafici dell'Atlante Statistico Italiano 1988. La popolazione residente al 1.1.1987 è stata rapportata ad un totale pari a 10.000; gli ammontari delle altre distribuzioni sono nello stesso rapporto in cui si trovano le grandezze effettive delle rispettive popolazioni.

Mentre la possibilità di fare le differenze fra gli ammontari effettivi di due distribuzioni di popolazione incontra dei limiti concettuali, il calcolo delle differenze fra le distribuzioni statistiche

equivalenti (A) e (B) entrambe riportate ad M = 10.000[12] è formalmente sempre possibile, in quanto si tratta della differenza fra due numeri adimensionali, i valori delle rispettive frequenze relative nelle celle i.

Il modo più immediato di interpretare tale differenza è considerarla come una misura di quanto una distribuzione è più o meno concentrata rispetto all'altra nella zona i.

Supponendo che i valori relativi siano moltiplicati x 100 e che $a_i = 0,9$ e $b_i = 0,5$, ciò significa che nella zona i è concentrato lo 0,9 per cento della distribuzione (A) e lo 0,5 per cento della distribuzione (B); la loro differenza indica semplicemente, in questo caso, che nella zona i è concentrata una percentuale della distribuzione (A) dello 0,4 per cento maggiore di quella della distribuzione (B). Si può anche dire che nella zona i è concentrata una percentuale della distribuzione (B) dello 0,4 per cento minore di quella della distribuzione (A).

A tali valori, quando sono positivi, si dà il nome di «eccedenze di (A) su (B)», quando negativi il nome di «eccedenze di (B) su (A)».

Per arricchire l'interpretazione di tale operazione, si possono aggiungere alcune osservazioni. Conviene riferirci ad una ipotetica distribuzione territoriale del reddito (A) e la corrispondente popolazione residente (B). Con a_i e b_i rispettivamente indichiamo l'ammontare relativo di reddito prodotto nella zona i e quello della popolazione ivi residente; supponiamo che tali valori relativi siano moltiplicati per 100.

Il rapporto A/B indica il reddito medio pro-capite a livello nazionale; il rapporto A_i/B_i indica il reddito medio pro-capite nella zona i.

Le differenze positive $a_i - b_i$ possono essere interpretate, innanzitutto, in modo immediato e formale, come la differenza fra le percentuali di reddito e di popolazione residente nella zona i; tale differenza esprime quanto una distribuzione sia più o meno concentrata nella zona i rispetto all'altra; si può anche dire che tale differenza corrisponde alla percentuale di reddito nazionale prodotta nella zona i che risulta in eccesso rispetto alla percentuale di reddito che sarebbe stata ivi prodotta nell'ipotesi di un reddito medio pro-capite nella zona i pari a quello medio nazionale, (corrispondente ad una uguale concentrazione fra le due distribuzioni nella zona in questione).

(12) I valori delle frequenze od ammontari di tali distribuzioni vengono ottenuti moltiplicando le frequenze o gli ammontari relativi a_i e b_i per una quantità M, che consideriamo nel caso attuale un numero *costante ed adimensionale* e quindi approssimando tale valore all'intero più vicino. Si ricorda che nella rappresentazione grafica della distribuzione equivalente i simboli grafici possono essere letti come valori assoluti o relativi, in funzione della chiave di lettura data nella legenda.

Se ad esempio $a_i = 1{,}0$ e $b_i = 0{,}8$ allora $a_i - b_i = 0{,}2$; ciò vuol dire che nella zona i è stato prodotto l'1 per cento del reddito ed è presente lo 0,8 per cento della popolazione; che quindi la distribuzione del reddito è concentrata nella zona i con una percentuale maggiore (eccedenza di (A) su (B)) di 0,2 punti rispetto a quella della distribuzione della popolazione residente.

Ciò può essere interpretato nel senso che nella zona i è presente una percentuale di reddito dello 0,2 per cento del reddito complessivo che risulta in eccesso rispetto alla percentuale di reddito (pari allo 0,8 per cento), che sarebbe stata prodotta se ivi il reddito medio pro-capite fosse stato pari al reddito medio pro-capite a livello nazionale.

Se tale differenza assume un valore negativo è possibile una interpretazione simmetrica; la quantità $a_i - b_i$, in valore assoluto, verrebbe ad indicare la percentuale del reddito complessivo che risulterebbe mancante nella zona i, nell'ipotesi che il reddito medio pro-capite nella zona i fosse pari al reddito medio pro-capite a livello nazionale.

Per una trattazione formale, in termini di concentrazione reciproca fra due generiche distribuzioni, si rimanda alla nota riportata in fondo a questo paragrafo, che formalizza questa interpretazione.

Nella figura 4.10 è riportato il grafico delle differenze fra la distribuzione (A): addetti alle industrie tessili — 1981, e la distribuzione (B): unità locali nelle industrie tessili — 1981. Riesce A = 493.000 addetti e B = 60.000 unità locali; il valor medio nazionale di addetti per unità locale è pari perciò a 8,2 addetti per unità locale.

Il cartogramma rappresenta le differenze delle due distribuzioni: in rosso sono rappresentate le quantità M $(a_i - b_i)$, con M = 10.000, che risultano positive (eccedenze di (A) su (B)), ed in verde le quantità M $(a_i - b_i)$ che sono negative, indicate in valore assoluto (eccedenze di (B) su (A)). La legenda indica semplicemente i valori assoluti delle differenze da 1 a 100; (per ottenere i valori percentuali occorre dividere quelli riportati nella legenda per 100).

Una prima lettura del grafico indica le differenze fra la concentrazione delle due distribuzioni nelle varie zone.

Dal grafico si ricava l'informazione sulle caratteristiche della distribuzione territoriale della concentrazione di tipo tecnologico dell'industria tessile in Italia. Laddove il simbolo grafico è di color rosso la differenza fra la distribuzione degli addetti e quella delle unità locali è positiva ed il valore del simbolo indica la differenza assoluta fra le due distribuzioni; tanto più elevato è tale valore tanto maggiore è la concentrazione di addetti per unità locale. Viceversa laddove il simbolo grafico assume il colore verde, la differenza fra la distribuzione degli addetti e quella delle unità locali è negativa; si ha, ovvero, una concentrazione di addetti per unità locale inferiore al valor medio nazionale. Anche in questo caso il simbolo sta ad indicare il valore numerico assoluto di tale differenza; tanto più elevato è il valore rappresentato tanto minore è il numero di addetti per unità locale.

Ad esempio, esprimendo i valori in percentuali, se $a_i = 0{,}7$, $b_i = 0{,}2$ e $a_i - b_i = 0{,}5$, ciò significa che nella zona i è presente lo 0,7 per cento di addetti e lo 0,2 per cento di unità locali dell'industrie tessili; quindi nella zona i è concentrata una percentuale di addetti maggiore dello 0,5 per cento della percentuale di unità locali ivi presente nelle industrie tessili.

Quindi nella zona i il numero di addetti per unità locale, in tali industrie, è superiore al valor medio nazionale, che è 8,2 addetti per unità locale. Inoltre la differenza 0,5 può essere interpretata come la percentuale degli addetti alle industrie tessili in eccesso rispetto alla percentuale di addetti che si sarebbe avuta nel caso in cui, stante il numero di unità locali ivi presente, il numero di addetti per unità locale fosse stato pari a quello medio nazionale.

Viceversa un valor negativo della differenza fra le due distribuzioni, rese in valore assoluto con i simboli GRP di colore verde, indica che in tale zona il numero di addetti per unità locale è inferiore a quello medio nazionale; ed il valore numerico indicato esprime la percentuale di addetti che occorrerebbe ivi aggiungere per rendere il numero medio di addetti per unità locale uguale a 8,2, valor medio nazionale.

Sovente può essere interessante la rappresentazione delle differenze in un solo verso, le $a_i - b_i$ di segno positivo. Per evidenziare, ad esempio, le zone ove si concentrano le attività di un dato ramo è sufficiente rappresentare le differenze positive fra la distribuzione degli addetti a quel ramo e quella degli addetti in totale.

Nel cartogramma saranno evidenziate le zone ove il rapporto fra gli addetti a quel ramo e gli addetti in totale è superiore al valor medio nazionale. Il simbolo grafico esprimerà le differenze positive, in termini relativi, della distribuzione degli addetti a quel ramo rispetto a quella degli addetti alle varie attività economiche.

Ad esempio, se (A) è la distribuzione degli addetti ad un certo ramo e (B) quella degli addetti in totale, i valori, espressi in percentuale, $a_i = 0{,}7$ e $b_i = 0{,}4$ indicano che nella zona i è presente lo 0,7 per cento di addetti ad un certo ramo e lo 0,4 per cento di addetti in totale; c'è quindi una presenza relativa maggiore di tale attività, rispetto alle attività economiche in generale, pari allo 0,3 per cento. È sempre la medesima argomentazione, solo che in questo caso l'attenzione si concentra esclusivamente sulle zone che presentano differenze positive.

Si vuole osservare che le caratteristiche delle due distribuzioni all'interno delle singole zone sono ovviamente escluse da ogni possibile considerazione e i confronti effettuati suppongono una distribuzione uniforme di queste all'interno delle zone.

Moltiplicando le differenze fra le due distribuzioni per l'ammontare complessivo di una delle due, ovvero leggendo la legenda in modo diverso, si può dare una ulteriore interpretazione di tale valore. Ad esempio se si moltiplicano le differenze fra le distribuzioni del reddito e della popolazione per il reddito complessivo, nel caso di differenze positive, tale prodotto indica la quantità di reddito prodotta nella zona in eccesso rispetto a quella che sarebbe stata prodotta in una zona ugualmente popolata avente un reddito medio pro-capite pari a quello medio pro-capite a livello nazionale [13].

In questo caso si può dire che essa corrisponde alla quantità di reddito che andrebbe idealmente trasferita verso le zone aventi un reddito medio pro-capite inferiore al reddito medio pro-capite a livello nazionale per raggiungere una distribuzione caratterizzata da un reddito medio pro-capite uguale in ogni zona del territorio.

Immaginando di completare l'insieme costituito dalle celle esagonali i, *con* i=1,.. 1000, *con l'insieme vuoto* ø, *si può definire lo spazio misurabile* H *associando a tale insieme la sigma-algebra degli insiemi* J, *generata da operazioni di unione e complementazione, combinate in modo arbitrario, sulle celle esagonali e l'insieme vuoto.*

Ogni distribuzione $\bar{A}$ *è allora interpretabile come una misura non negativa e finita su* H.

È formalmente sempre possibile confrontare le due distribuzioni territoriali (A) *e* (B) *quando entrambe sono rapportate allo stesso valore* M.

Infatti, in tal caso, possiamo considerare (A) *e* (B) *come misure normalizzate definite sullo stesso spazio misurabile* H *(Tranquilli 1985a). Per esse è possibile calcolare l'indice relativo di dissomiglianza* $\Delta = \frac{1}{2}\Sigma \mid a_i - b_i \mid$, *interpretabile come distanza fra le due distribuzioni (Leti 1983) e che varia fra 0 e 1.*

Le differenze $a_i - b_i$, *eventualmente moltiplicate per* M, *misurano, prese in valore assoluto, la maggiore preponderanza relativa di una distribuzione sull'altra. Di tali quantità si può dare una ulteriore interpretazione, che viene esposta qui di seguito.*

L'indice su riportato appare anche nello studio della concentrazione fra variabili statistiche, come una misura sintetica della concentrazione reciproca.

Riferendoci ai concetti e alle definizioni date da Tranquilli (1985a), si può costruire sul piano (a,b) *degli ammontari relativi delle distribuzioni* (A) *e* (B), *associate sullo stesso spazio* H, *l'area e la relativa curva di concentrazione inferiore.*

Ad ogni insieme J, *elemento dello spazio misurabile* H, *resta associata una determinata coppia* (a(J), b (J)), *i cui valori rappresentano gli ammontari parziali relativi, cumulati sull'insieme* J *dalle distribuzioni* (A) *e* (B) *rispettivamente. Rappresentando sul piano* (a,b) *tutti i punti di coordinate* (a(J),b(J)), *corrispondenti agli*

(13) Quanto qui espresso verbalmente può essere simbolicamente reso tramite la seguente formula:

$$R\left(\frac{R_i}{R} - \frac{P_i}{P}\right) = \left[R_i - \left(\frac{R}{p}\right) P_i\right]$$

avendo indicato con $\bar{R}$ e $\bar{P}$ rispettivamente la distribuzione del reddito e della popolazione residente.

insiemi J *della sigma-algebra di* H, *e considerando la minima chiusura convessa dell'insieme di punti così ottenuto, si costruisce l'area di concentrazione, la cui frontiera inferiore costituisce la corrispondente curva di concentrazione (inferiore); si può vedere, a tal proposito, la figura 4.11.*

L'indice relativo di dissomiglianza misura la distanza massima fra la retta di equiconcentrazione b=a *e la curva di concentrazione, calcolata lungo l'asse* b.

Tale osservazione è presente già nel lavoro di Duncan, Duncan (1955), dove Δ *veniva studiato e proposto come possibile misura della segregazione fra due popolazioni* $\bar{A}$ *e* $\bar{B}$ *(neri e bianchi in Duncan, Duncan 1955).*

Nel caso che le due distribuzioni statistiche (A) *e* (B) *risultino assolutamente continue l'una rispetto all'altra* ($a_i = 0$ *se e solo se* $b_i = 0$), *si dimostra l'equivalenza dello studio della concentrazione reciproca fra* (A) *e* (B) *a quello della concentrazione classica secondo Lorenz-Gini della variabile statistica che associa ad ogni cella esagonale il valore* a_i/b_i, *ponderato con peso pari a* b_i *(Tranquilli 1985b).*

La rappresentazione grafica delle differenze $a_i - b_i$ ***ha perciò un significato di rilevante senso statistico; essa riconduce lo studio della concentrazione reciproca fra le due distribuzioni alla sua dimensione territoriale.***

Ritornando all'area di concentrazione nel piano (a,b), *ha un certo interesse considerare la proiezione dei punti associati alle singole celle* i, (a_i, b_i), *sulla retta di equiconcentrazione* b=a, *al modo seguente:*

$(a_i,b_i) \rightarrow (a_i^e,b_i^e)$ dove $a_i^e = b_i^e = \min(a_i,b_i)$;

(min *indica il minore fra i due valori dell'argomento).*

Per il generico punto (a_i,b_i) *possiamo porre:*

(1) $a_i = a_i^e + sa_i$, $b_i = b_i^e + sb_i$,

dove sa_i *e* sb_i *misurano rispettivamente la lunghezza dei segmenti che collegano il punto* (a_i^e,b_i^e) *sulla retta* b=a *al punto* (a_i,b_i), *rispettivamente lungo l'asse* a *e* b, *come risulta dalla figura 4.12.*

Riesce $sb_i = 0$ *se* $sa_i > 0$ *e viceversa. Se è* $sa_i = sb_i = 0$ *allora già in partenza* $a_i = b_i$. *È facile dimostrare che è* $\Sigma_i sa_i = \Sigma_i sb_i = \Delta$ e che $\Sigma\, a_i^e = \Sigma_i b_i^e = 1 - \Delta$.

L'indice Δ *ha acquisito negli studi sulla segregazione di popolazioni distribuite sullo stesso territorio una certa predominanza (White 1986); ciò è legato, come ricorda lo stesso White, alla pregnanza della sua interpretazione nei termini già posti dai Duncan nel '55.*

Esso è semplicemente interpretabile come la frazione di elementi di una distribuzione che è necessario spostare per arrivare ad una situazione di assenza di segregazione, tale che gli ammontari relativi di ciascuna distribuzione siano fra loro uguali in ciascuna cella i, *(Duncan 1957, Morgan 1982).*

L'interpretazione del Duncan in termini di segregazione fra due popolazioni è formalmente traducibile in termini di concentrazione fra le due distribuzioni statistiche generiche (A) *e* (B), *riferite al medesimo territorio. Nell'ipotesi che la modifica degli ammontari relativi di una distribuzione non influenzi gli ammontari relativi dell'altra, la rappresentazione cartogrammatica delle differenze* $a_i - b_i$ *può essere letta al modo seguente: la quantità* sa_i *rappresenta la parte di* a_i *che deve essere spostata dalla zona* i *perchè ci sia equiconcentrazione localmente; essa deve idealmente essere spostata, insieme con le altre frazioni relative alle zone con* $a_i > b_i$, *verso le zone con* $sb_i > 0$ *affinchè si raggiunga equiconcentrazione a livello globale.*

In questo contesto le quantità $a_i^e = b_i^e$ *possono essere interpretate come le frazioni della* (A) *e della* (B) *che, nelle zone* i, *risultano associabili in partenza secondo il rapporto medio di equiconcentrazione* A/B *e che, perciò, non risultano coinvolte negli spostamenti ideali necessari a raggiungere l'equiconcentrazione.*

Terminiamo il corrente paragrafo con alcune definizioni.

Si indica con (A > B) *e* (A < B) *le distribuzioni equivalenti determinate dai valori* sa_i, sb_i, *dopo che la loro somma è stata riportata al valore* M; *si indica con* (A = B) *la*

distribuzione equivalente determinata dalle quantità $a_i^* = b_i^*$, una volta che la loro somma è stata riportata al valore M.

Si definisce:

(A > B) come la distribuzione delle «eccedenze di (A) *su* (B)*» e (A < B) come la distribuzione delle «eccedenze di* (B) *su* (A)*»; (A = B) come «parte comune fra* (A) *e* (B)*».*

Le distribuzioni (A) *e* (B)*, i cui ammontari relativi risultano scomposti secondo la (1), possono considerarsi come misture della* (A = B) e delle (A > B) *e* (A < B) *rispettivamente; ciò può essere riscritto in forma simbolica al modo seguente:*

(2) $(A) = \Delta\,(A > B) + (1 - \Delta)\,(A = B)$; $(B) = \Delta\,(A < B) + (1 - \Delta)\,(A = B)$;

Tale analisi è suscettibile di ulteriori sviluppi; è possibile trovare una semplice ed elegante relazione fra i baricentri delle cinque distribuzioni equivalenti che appaiono nella (2); si può anche estendere lo studio dell'indice di Duncan fino alla proposta di un indice basato sulla distanza dotato di interessanti proprietà (Bachi 1984).

Nel caso del confronto di tipo assoluto fra due distribuzioni gli ammontari di queste sono rapportati a valori diversi; le considerazioni su riportate non sono chiaramente applicabili ed in particolare i valori complessivi delle differenze in un verso non coincidono con quelle nel verso opposto e la loro somma algebrica è uguale alla differenza fra gli ammontari dei due fenomeni.

6. LA RAPPRESENTAZIONE GRAFICA DEI QUOZIENTI DI DUE DISTRIBUZIONI DI POPOLAZIONE SUL TERRITORIO

Il confronto fra due serie territoriali può richiedere la considerazione dei rapporti fra le rispettive frequenze od ammontari, zona per zona, sul territorio di riferimento. Rifacendoci alle simbologie utilizzate nei paragrafi precedenti indicheremo con A_i e B_i gli ammontari assoluti delle due distribuzioni $\hat{A}$ e $\hat{B}$, rispettivamente, nella zona i e con A e B gli ammontari rispettivi complessivi. Si indicherà con $g_i = A_i / B_i$ il rapporto fra le due quantità; si indicherà con $\bar{g} = A/B$ il valore medio del rapporto a livello nazionale.

Il significato di tale rapporto dipende dalla natura delle due distribuzioni messe a confronto; può essere utile presentare alcuni esempi:

a) se A_i è la popolazione in i e B_i è l'area totale o l'area effettivamente abitata di i, detta area ecumenica, allora il rapporto g_i indica la densità totale o netta della popolazione in i;

b) se A_i è l'area coltivata a frumento in i, e B_i è l'area totale o agrario-forestale o l'area a seminativi in i, allora il rapporto $100 \times g_i$ indica la percentuale dell'area di i (definita in uno dei modi indicati sopra) coltivata a frumento;

c) se A_i è il numero delle nascite avvenute in i durante un anno e B_i la popolazione di i, il rapporto $g_i = 1000\ A_i / B_i$ indica il quoziente (annuo) di natalità in i; se B_i è la popolazione femminile in età feconda, g_i è il quoziente generico di fecondità;

d) se A_i è il numero delle nascite naturali in i e B_i è il numero delle nascite in i, il rapporto $g_i = 100\ A_i / B_i$ è la percentuale della filiazione naturale in i.

Tali quantità sono comunque direttamente interpretabili come il risultato della divisione fra i valori associati alla zona i, cioè del loro quoziente, e, quindi, come il valore medio, nella zona i, della grandezza a numeratore rispetto ad un valore unitario della grandezza posta a denominatore.

Generalmente si usa rappresentare la quantità g_i, trasformata e suddivisa in classi di valori, distribuendo una variabile visiva, come il valore ed anche il colore, sull'intera area della zona di riferimento. Ciò equivale ad effettuare una ponderazione che, a livello percettivo, risulta proporzionale all'area.

Si ricorda a questo proposito quanto argomentato da Schmid (1983) nell'individuare a tal proposito un «area bias» ed il lavoro di Bachi del 1978.

Una considerazione utile per approfondire questo aspetto è che il valore del rapporto da attribuire alla zona ottenuta per aggregazione di due zone contigue i e j è una media ponderata dei rapporti delle zone iniziali, con pesi $B_i/(B_i + B_j)$, $B_j/(B_i + B_j)$; può essere importante perciò fornire al lettore una indicazione delle quantità poste a denominatore.

È ovvio che nel caso in cui B_i venga a coincidere con l'area della zona i (ed il rapporto è perciò una densità territoriale), la rappresentazione grafica tramite il cartogramma che, nella sua costruzione, utilizza un sistema regolare di zone, non richiede alcuna accortezza ulteriore.

Negli altri casi può essere utile fornire al lettore, insieme con la rappresentazione del rapporto, anche una indicazione del valore B_i associato alle varie zone.

Per quanto riguarda la rappresentazione del valore numerico del rapporto g_i, si può utilizzare il sistema dei simboli GRP; come in termini generali è stato indicato nel paragrafo 7 del capitolo 2 può talvolta essere utile scegliere una opportuna funzione di trasformazione delle g_i, che renda più evidenti le caratteristiche della distribuzione territoriale.

La scelta di tale funzione di trasformazione è anche legata al significato del rapporto; descriviamo alcune possibili semplici funzioni di trasformazione.

Se la distribuzione $\hat{A}$ si può considerare una parte di $\hat{B}$, come nel caso degli addetti ai servizi e gli addetti in totale, o nel caso della popolazione di una classe di età e il totale della popolazione, potremmo essere interessati a mostrare il rapporto degli ammontari A_i e B_i come una percentuale; in questo caso si tratta di scegliere l'intero più vicino al valore $100xg_i$ e rappresentarlo con il simbolo grafico razionale "s_i" corrispondente.

Nel caso di quozienti di natalità o di mortalità si moltiplicano generalmente tali valori per 1.000; si usa anche, quando il numero dei casi in rapporto alla popolazione di riferimento è ancora più esiguo, un fattore pari a 100.000.

Talvolta si può essere interessati, volendo confrontare fenomeni diversi, ad eliminare da g_i l'effetto dovuto all'ordine di grandezza del fenomeno, come viene misurato dal valore medio $\bar{g}$. Si può talvolta ritenere necessario eliminare nella rappresentazione l'effetto dovuto alla variabilità dei rapporti. Ciò può essere necessario a scopo di confronto o per ampliare o, al contrario, per ridurre le differenze dei valori assunti dai quozienti, che possono risultare numericamente molto simili o, al contrario, essere distribuiti su un intervallo molto ampio di valori. Si può, in questo caso, procedere alla standardizzazione della variabile statistica g_i.

Strettamente legato a tale problema è quello della scelta della scala della rappresentazione grafica; per rendere nel modo più evidente le differenze territoriali, conviene utilizzare per ciascuna distribuzione l'intero intervallo dei GRP da 1 a 100. Concettualmente ciò è analogo a quanto si fa nei diagrammi cartesiani quando si sceglie di troncare la scala effettivamente rappresentata. Per poter fare confronti semplici fra grafici diversi conviene invece utilizzare la medesima scala per tutti i fenomeni da rappresentare; a tal proposito si possono considerare i cartogrammi dei quozienti di natalità e mortalità riportati nell'Atlante Statistico Italiano 1988.

La discussione precedente ha affrontato il problema della rappresentazione del valore numerico del quoziente tramite i simboli grafici GRP. Rimane la questione relativa all'indicazione del denominatore del quoziente g_i. Si sottolinea che nel caso di «variabili statistiche rapporto» l'informazione consiste in due quantità numeriche; essa può essere disponibile secondo la coppia (A_i, B_i) degli ammontari parziali, o nella forma, ad esempio, (g_i, B_i) che dà direttamente il valore del rapporto e del suo denominatore; nel contesto dell'attuale discussione si suppone che $B_i \neq 0$ e che le due forme siano equivalenti.

Per l'indicazione di B_i distinguiamo due casi, a seconda che il dato sia noto per ciascuna cella esagonale o solo a livello di zone come le Regioni o le Province.

Nel primo caso l'indicazione di B_i può essere data disegnando i simboli GRP con gradazioni diverse, dal chiaro allo scuro, di uno stesso colore in funzione del suo valore. Per far ciò, occorre ordinare i quozienti in funzione delle classi in cui vanno suddivisi i valori B_i. Nel cartogramma 4.13 viene fornito un esempio di tale rappresentazione; in esso la distribuzione a denominatore è la popolazione media del periodo 1981-1986 ed è stata suddivisa in quattro classi di ampiezza i cui limiti sono stati scelti in modo che i 1.000 esagoni venissero ugualmente distribuiti fra di esse.

In questa rappresentazione si dà l'informazione del tasso in termini quantitativi, tramite il sistema dei GRP e si fornisce ulteriormente al lettore l'indicazione dell'esistenza del peso nella definizione della variabile statistica rappresentata, tramite l'uso della

variazione di valore [14], che svolge in questo caso una utile funzione di arricchimento dell'informazione.

Nell'altro caso il dato viene fornito a livello di un sistema di zone ampie, come le Province, e si può adottare una procedura del tipo seguente. All'interno di ciascuna zona (Provincia), viene distribuito in modo uniforme un numero di punti proporzionale al valore di B_i; su ciascuno di questi punti viene riprodotto il simbolo grafico corrispondente al valore g_i. Tale sistema pare fondamentalmente corretto in quanto fornisce separatamente entrambe le informazioni, rispettando le caratteristiche quantitative dell'informazione di entrambi gli elementi del dato. Esso inoltre si può avvalere di applicazioni informatiche di tipo standard che lo rendono di facile realizzazione, una volta che si disponga dei dati relativi alle poligonali chiuse che delimitano le varie zone.

Come esempio riportiamo nel grafico 4.14 la rappresentazione del quoziente di natalità relativo al periodo intercensuario 1971-81, ripreso dall'Atlante Statistico Italiano 1988.

I metodi grafici proposti risolvono alcuni importanti problemi posti dalla rappresentazione di quozienti; tuttavia esistono altri problemi oltre a quelli qui considerati; solo accumulando ulteriori esperienze si potranno migliorare i risultati presentati in questo lavoro.

7. LA RAPPRESENTAZIONE GRAFICA DI SERIE TERRITORIALI SECONDO UN SISTEMA NON REGOLARE DI ZONE

In questo paragrafo si considera la possibilità che la rappresentazione grafica faccia riferimento ad un sistema non regolare di zone, a diversi livelli possibili di risoluzione spaziale.

Se l'informazione della serie territoriale consiste nell'associare alle varie zone che costituiscono una partizione del territorio una modalità di un carattere qualitativo e se la scala della rappresentazione è sufficientemente grande da permettere una chiara identificazione dei limiti delle singole zone, si può senz'altro pensare a rappresentare in modo congiunto le varie modalità qualitative con l'uso di variabili visive quali il valore o il colore, a seconda che il carattere sia qualitativo ordinabile o sconnesso; a tal proposito è opportuno tener presente gli argomenti presentati nel paragrafo 6 del Capitolo 2.

Alcuni esempi di tali applicazioni, si possono ritrovare nella rappresentazione dei Comuni classificati secondo le zone altime-

(14) Nel caso che si desideri fissare l'attenzione esclusivamente sul valore del quoziente, si può evitare di rappresentare il valore dei B_i, il che, se semplifica la rappresentazione, comporta una perdita di informazione.

triche della classificazione Istat; oppure nella rappresentazione della classificazione dei Comuni italiani secondo il grado di sismicità, come fissata dal Ministero dei Lavori Pubblici.

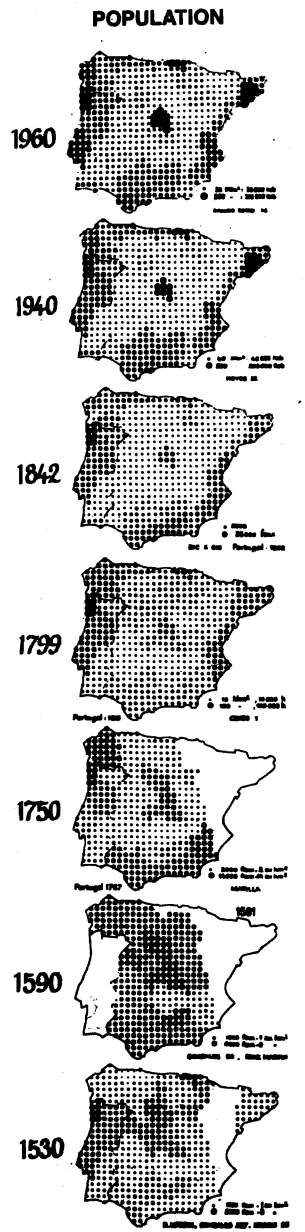
In altri casi ciò che interessa è la semplice individuazione e differenziazione di ampie zone sul territorio, come nel caso della rappresentazione dei bacini idrografici a livello di primo affluente, dove è quanto mai opportuno l'uso del colore per differenziare le diverse zone.

Si vuole terminare con un accenno alle rappresentazioni di quantità, come frequenze od ammontari, relative a suddivisioni ampie del territorio come le Regioni o le Province; anche per tale tipo di rappresentazione si può far riferimento alla suddivisione dell'Italia in 1000 celle esagonali, distribuendo il dato relativo a ciascuna zona in modo uguale fra gli esagoni che vi appartengono, per passare infine alla rappresentazione delle quantità tramite i simboli GRP.

La flessibilità del sistema può consentire, tuttavia, di distribuire in modo differenziato l'informazione fra le celle appartenenti alla stessa zona, utilizzando informazioni ausiliarie di dettaglio spaziale maggiore; talvolta nella redistribuzione del dato fra le celle si può tener conto del fatto che esse contengano o meno il centro di capoluoghi di Provincia.

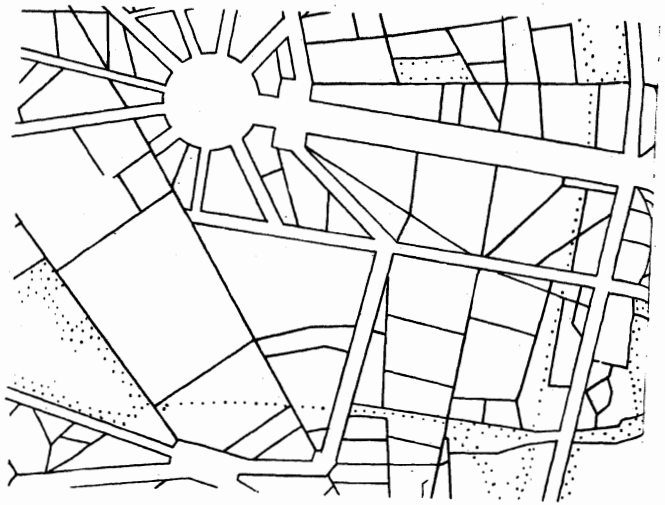
Nel caso della rappresentazione dei quozienti, di cui si è discusso nel precedente paragrafo 6, è possibile l'uso di algoritmi che, in riferimento a zone di cui sono noti i contorni chiusi, distribuiscono in modo uniforme un numero di punti proporzionale alla grandezza del denominatore della variabile rapporto, sui quali si rappresenta poi, tramite i simboli grafici GRP, il valore numerico del quoziente.

Figura 4.1 — Serie di cartogrammi relativi alla popolazione della penisola iberica



Fonte: Bertin (1983)

Figura 4.2 — Popolazione in un quartiere di Parigi (1946)



III. — RESIDENTIAL POPULATION AND GEOGRAPHIC SECTORS (1946)

1/10 000

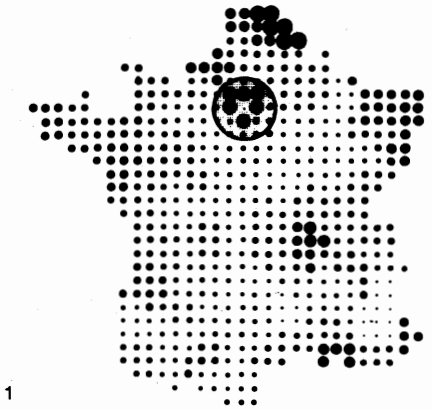
1

2

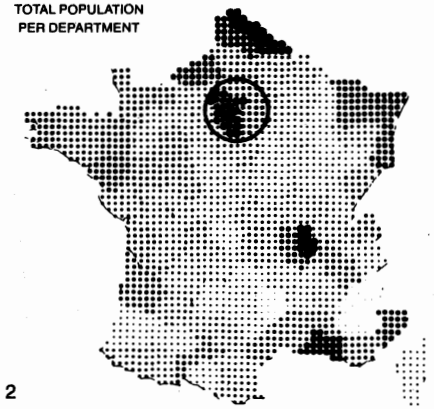
1/25 000

Fonte: Bertin (1983)

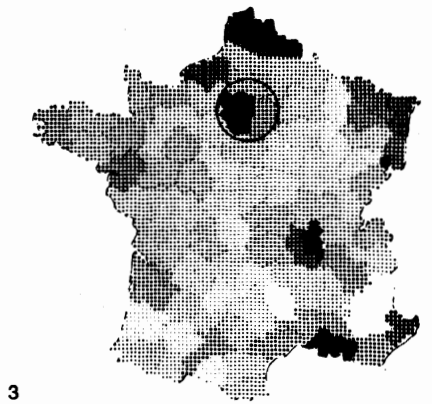
Figura 4.3 — Cartogrammi che utilizzano nella rappresentazione un numero crescente di punti



1



2

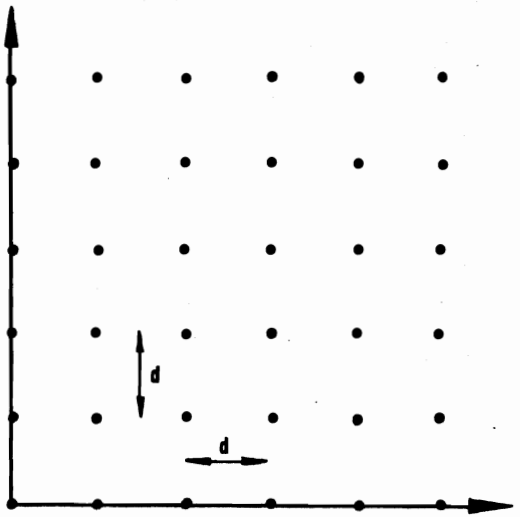


3

Fonte: Bertin (1983)

Figura 4.4 — Esempi di sistemi regolari di punti

(a) sistema di celle quadrate



(b) sistema di celle esagonali

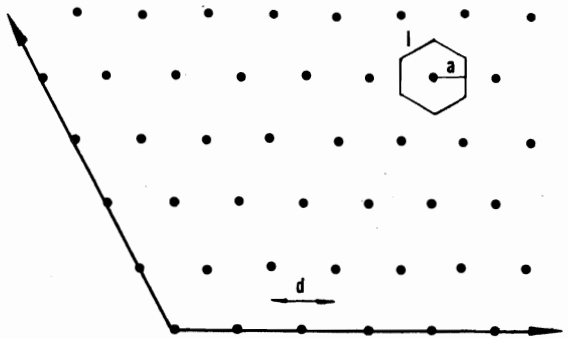
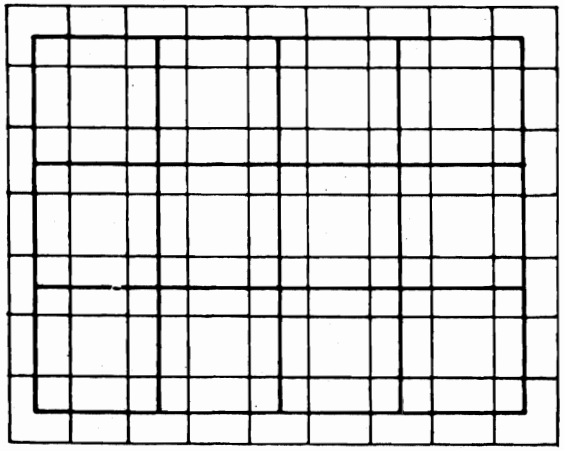


Figura 4.5 — Possibili aggregazioni delle celle in due sistemi regolari di punti

(a)



(b)

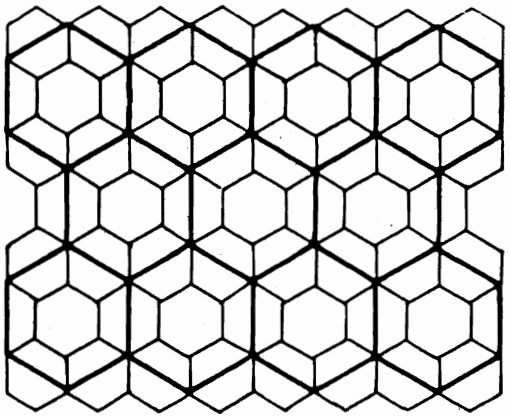


Figura 4.6 — Suddivisione del territorio italiano in 1000 celle esagonali

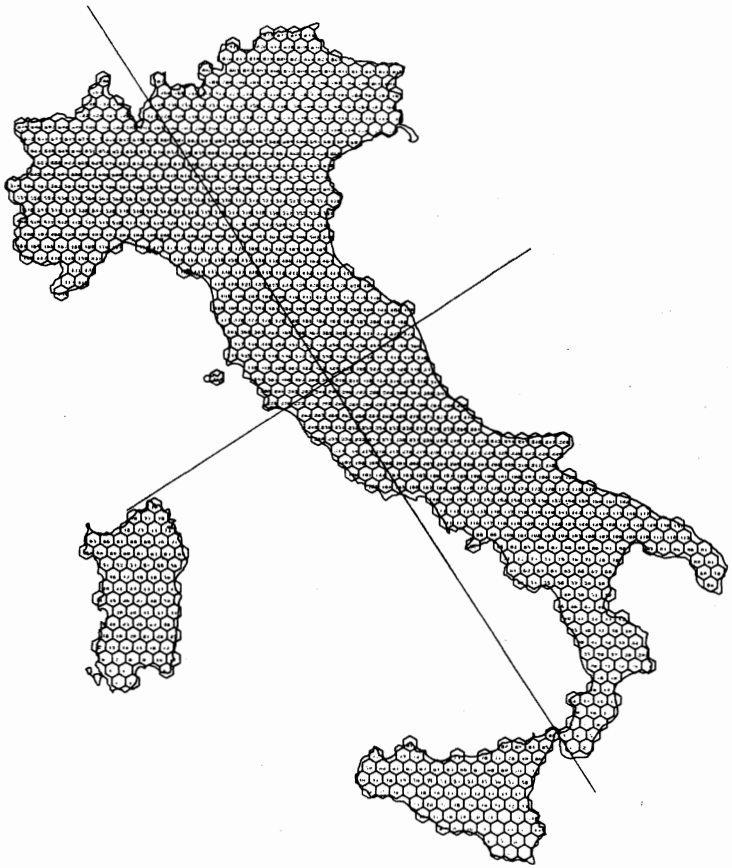
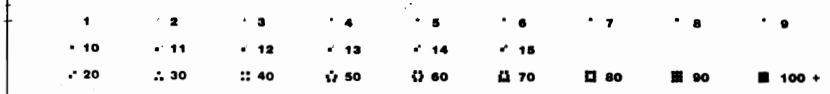
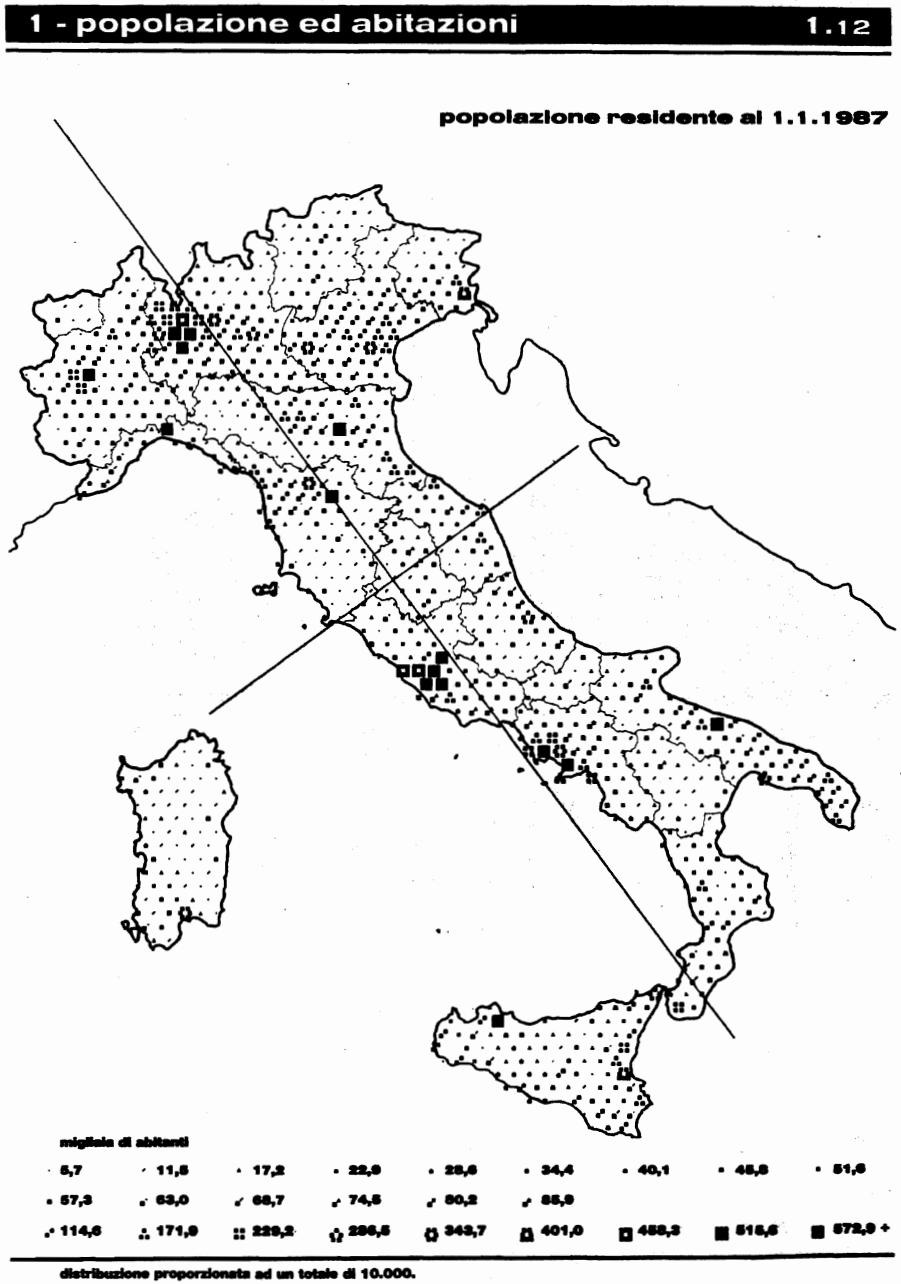


Figura 4.7 — Scala ridotta di simboli grafici razionali (GRP)



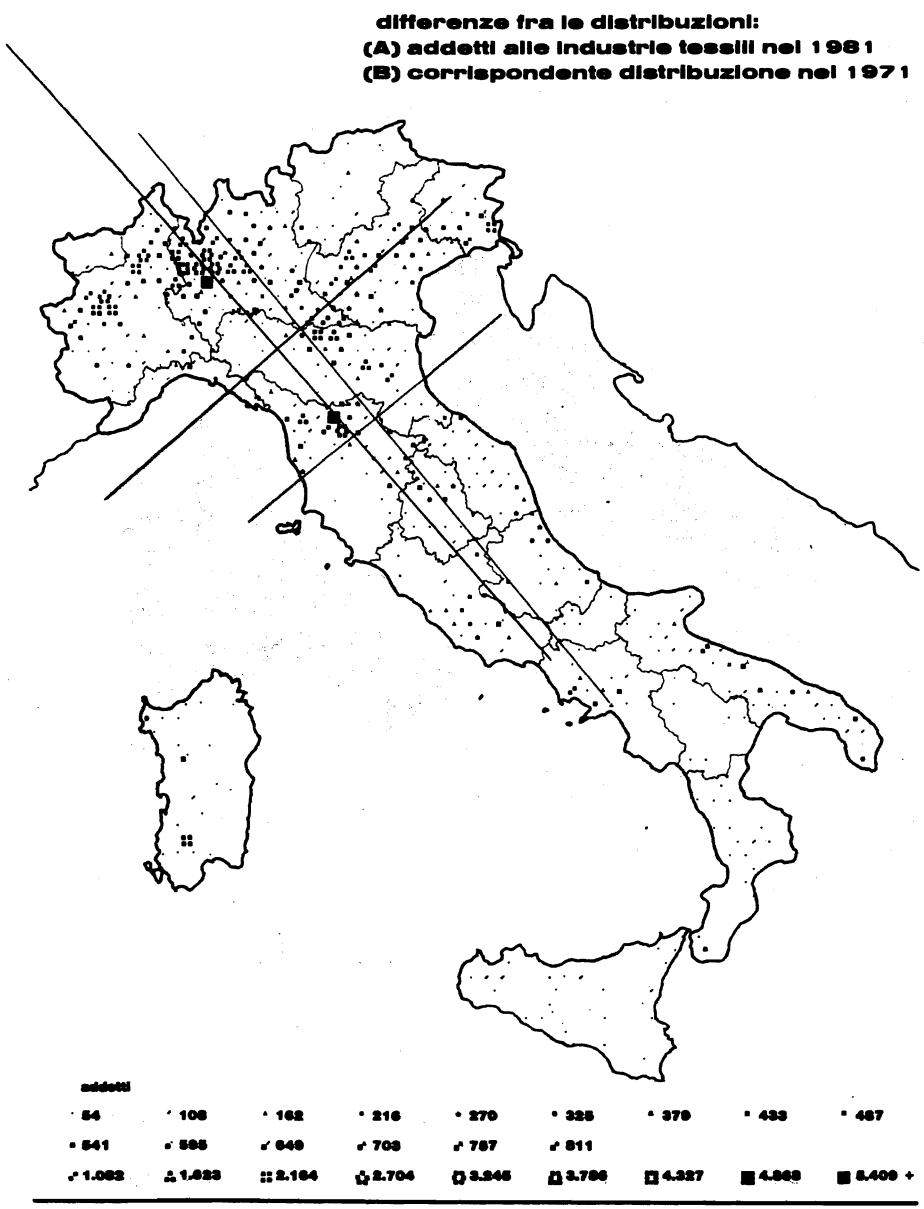
Fonte: Atlante Statistico Italiano (1988)

Figura 4.8 — Esempio di cartogramma razionale di una distribuzione semplice



Fonte: Atlante Statistico Italiano (1988)

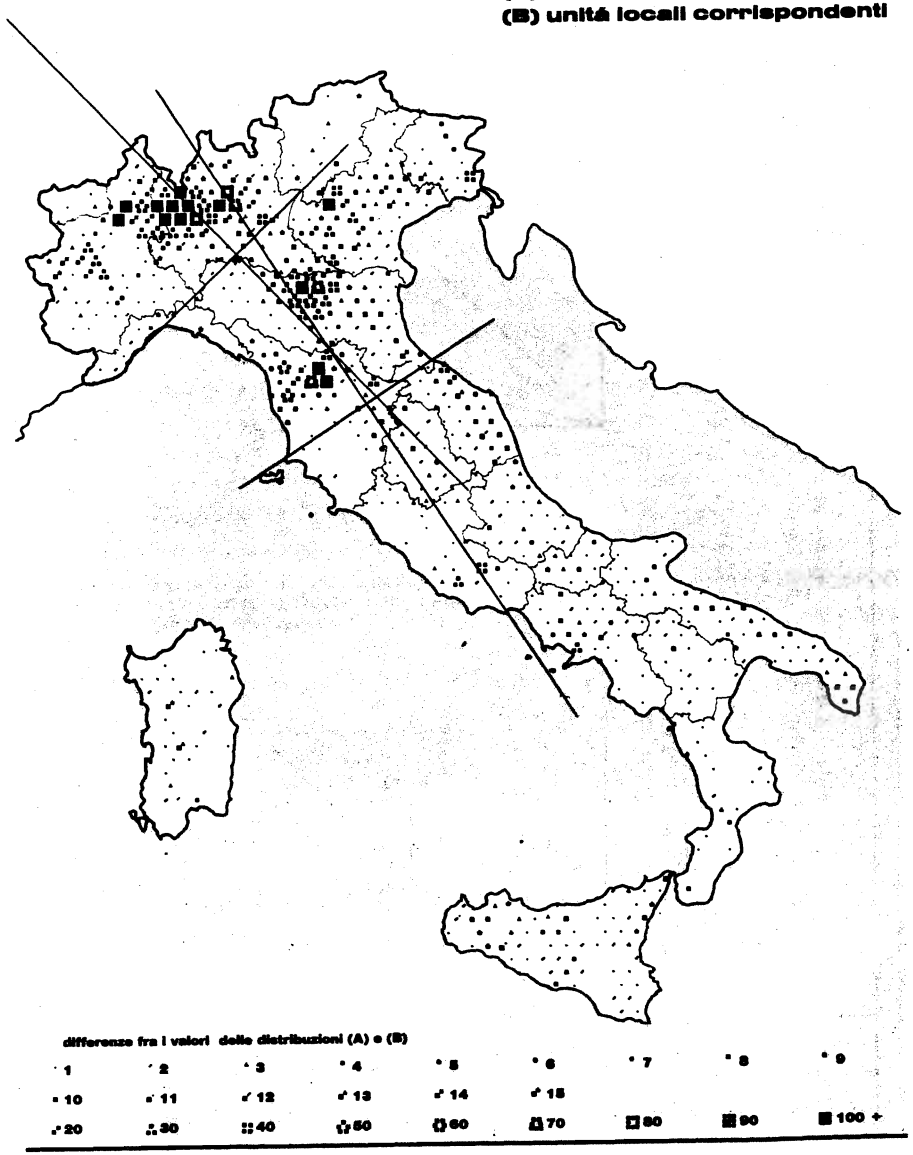
Figura 4.9 — Esempio di cartogramma delle differenze di tipo assoluto fra due distribuzioni



la distribuzione (A) è proporzionata ad un totale di 9.123. la distribuzione (B) è proporzionata ad un totale di 10.000.
eccedenze di (A) su (B) in rosso. eccedenze di (B) su (A) in verde.
totale delle eccedenze di (A) su (B) pari a 1.001. totale delle eccedenze di (B) su (A) pari a 1.878.

Fonte: Atlante Statistico Italiano (1988)

Figura 4.10 — Esempio di cartogramma delle differenze di tipo relativo fra due distribuzioni



le distribuzioni (A) e (B) sono proporzionate ad un totale di 10.000.
eccedenze di (A) su (B) in rosso. eccedenze di (B) su (A) in verde.
totale delle eccedenze di (A) su (B) pari a 3.363.

Fonte: Atlante Statistico Italiano (1988)

Figura 4.11 — Curva di concentrazione inferiore

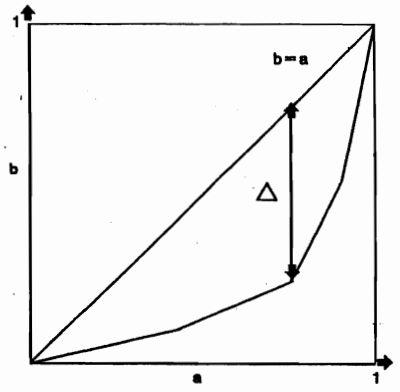


Figura 4.12 — Scomposizione degli ammontari relativi in parte comune ed eccedenze

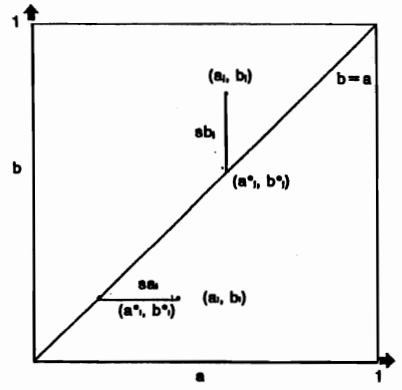
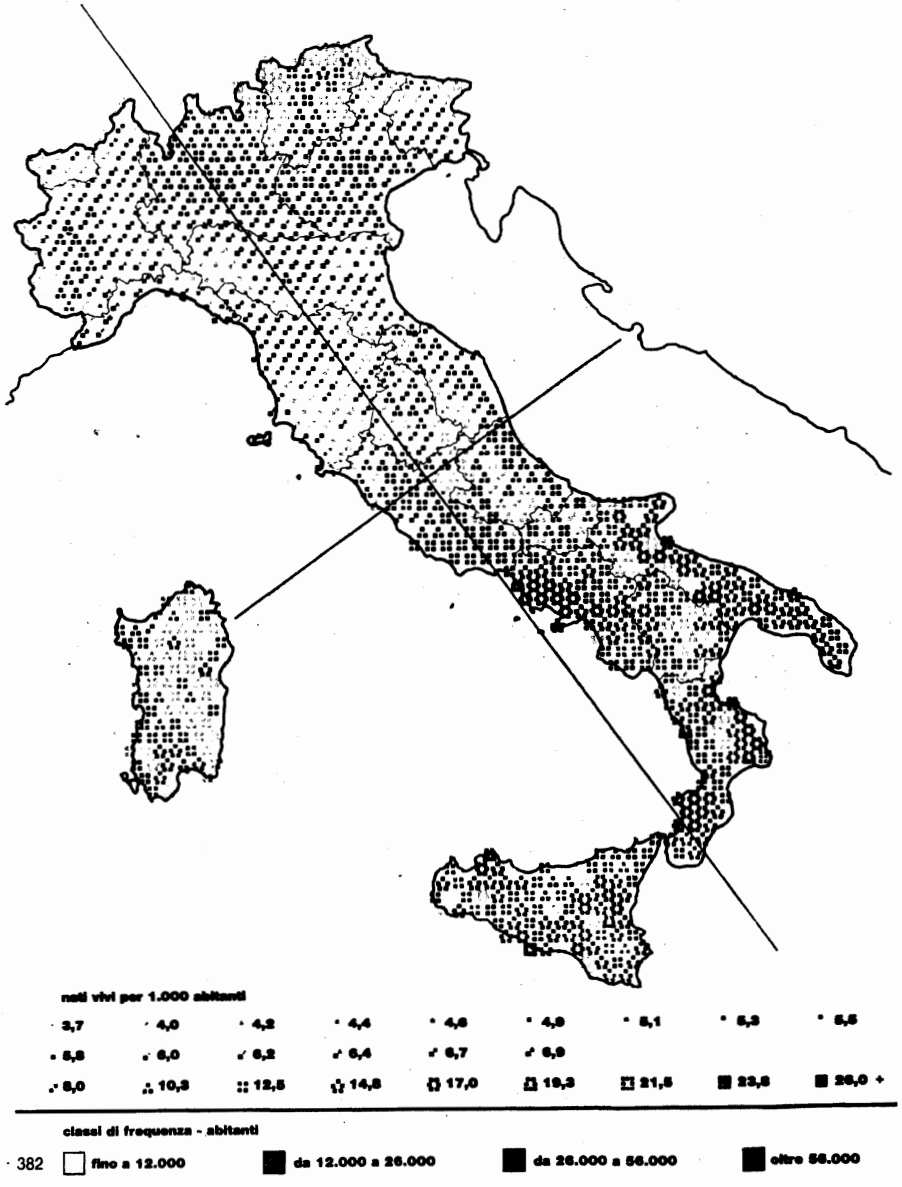
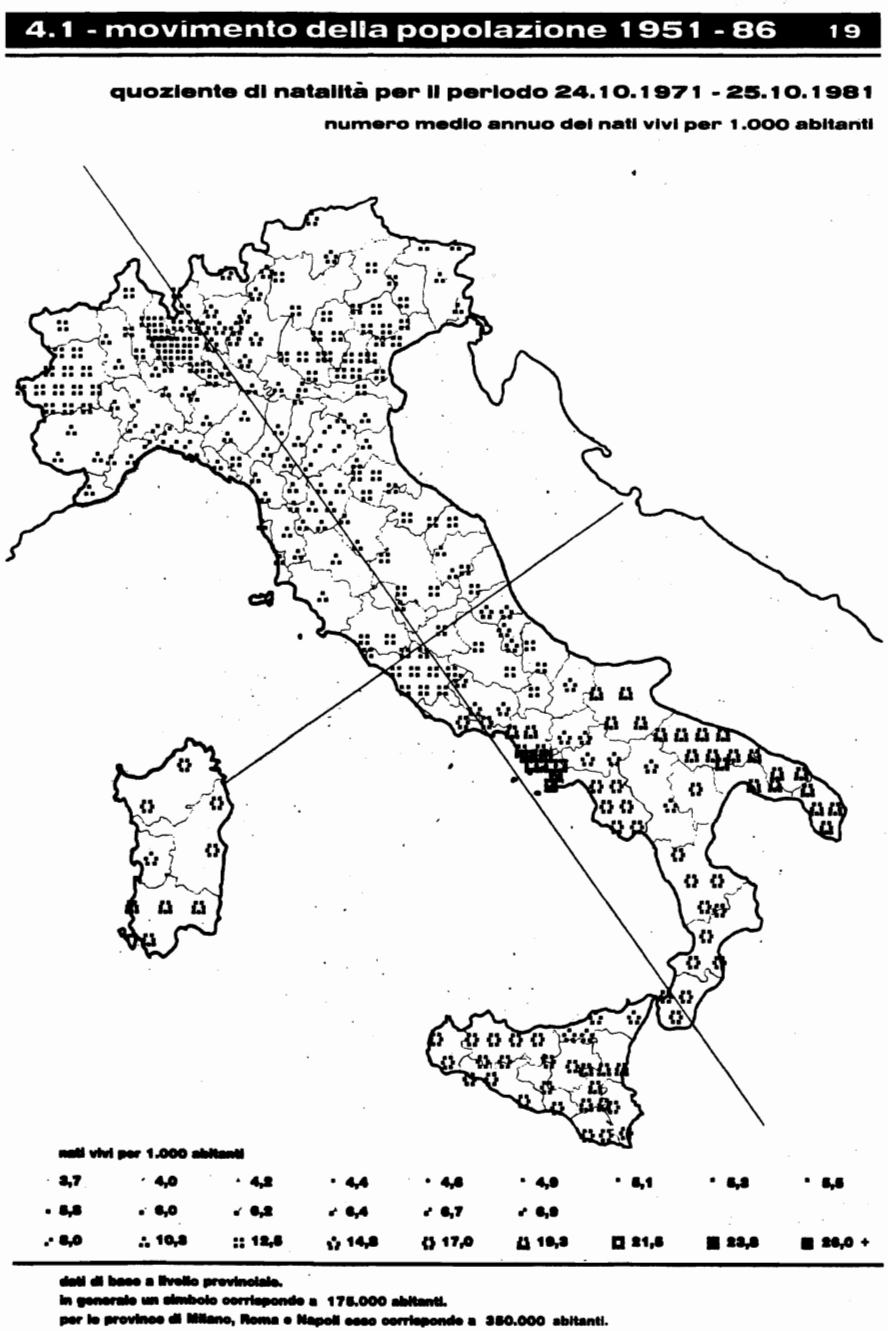


Figura 4.13 — Esempio di rappresentazione di un quoziente basato su dati comunali



Fonte: Atlante Statistico Italiano (1988)

Figura 4.14 — Esempio di rappresentazione di un quoziente basato su dati provinciali



Fonte: Atlante Statistico Italiano (1988)

RIFERIMENTI BIBLIOGRAFICI

Ad Hoc Commitee on Statistical Graphics, ASA (1979), *Final Report*, a cura di Biderman, A.D.

Andrews, D.F. (1972) *Plot of high dimensional data*, Biometrics, 28

Bachi, R. (1962) *Standard distance measures and related methods for spatial analysis*, Regional Science Association, Paper X, Zurich Congress

Bachi, R. (1968) *Graphical rational patterns: a new approach to graphical presentation of statistics*, Israel University Press, Jerusalem

Bachi, R. (1975) *Graphical methods: achievements and challanges for the future*, ISI, IX sessione, Varsavia

Bachi, R. (1978) *Proposals for the development of selected graphical methods*, Technical paper n. 43, Bureau of census, Washington D.C.

Bachi, R. (1981) *Mapping the main characteristics of distributions of populations over territories*, 43 Sessione dell'ISI, Buenos Aires

Bachi, R. (1984) *Parametri delle caratteristiche generali di distribuzioni di popolazioni sul territorio*, Atti delle XXXII riunione scientifica SIS, Sorrento

Beniger, J.R., Robyn, D.L. (1978) *Quantitative Graphics in Statistics: a brief history*, American Statistician, 32

Benini, R. (1905) *I diagrammi a scala logaritmica*, Il Giornale degli Economisti 2-da serie

Bertin, J. (1981) *La grafica ed il trattamento grafico dell'informazione*, ERI

Bertin, J. (1983) *Semiology of graphics*, Wisconsin University Press

Boldrini M. (1968) *Statistica, teoria e metodi*, V ediz., Milano

Bonfatti, F. (1985) *Basi di dati geografici;* in *Basi di dati*, a cura di Tiberio, P., Milano

Bonin, S. (1975) *Initiation a la graphique*, EP1 editeurs, Paris

Carr, D.B., Littlefield, R.J., Nicholson, W.L., Littlefield, J.S. (1987) *Scatterplot matrix techniques for large N*, J.A.S.A. 82, 398

Caselli, C., Lombardo, E., (1988) *Grafici ed analisi dei fenomeni demografici*, XXXIV Riunione scientifica SIS, Siena

Castellano V. (1988) *Prospettive di una collaborazione fra produttori e utilizzatori della informazione statistica;* Genus, vol XLIV, n. 1-2.

Chambers, J.H., Cleveland, W.S., Kleiner, B., Tukey, P.A. (1983) *Graphical methods for data analysis*, Boston, Mass. Duxbury Press

Chernoff. H., (1973) *The use of faces to represent points in k-dimensional space graphically*, J.A.S.A. 68, 342

Chernoff, H., Rizvi, M.H. (1975) *Effect on classification of random permutations of features in representing multivariate data by faces*, J.A.S.A. 70, 351

Cleveland, W.S. e McGill, M. (1984) *Graphical perception theory, experimentation and application to the development of graphical methods*, J.A.S.A. 79, 387

Cleveland, W.S. e McGill, M. (1987) *Graphical perception: the visual decoding of quantitative information on graphical displays of data*, Journal of Royal Statistical Society, series A, 150, 3

Cleveland, W.S. (1987) *Research in statistical graphics*, J.A.S.A., 82, 398

Cox, D.R. (1978) *Some remarks on the role in statistics of graphical methods,* Applied Statistics 27

Dajoz, R. (1972) *Manuale di Ecologia,* Milano

Davis, J.C., McCullagh, M.J. (1975) *Display and analysis of spatial data,* Wiley

De Meo, G. (1975) *Corso di Statistica economica,* Roma

Diethelm, W. (1984) *Signet, Signal, Symbol,* Zurigo

Duncan, O.T., Duncan, B. (1955) *A methodological analysis of segregation indexes,* American Sociological Review

Duncan, O.T. (1957) *The measurament of population distribution,* Population studies, vol XI

Feinberg, B.M., Franklin, C.A. (1975) *Social Graphics Bibliography,* Bureau of social science research, Washington D.C.

Foley, J.D., Van Dam, A. (1983) *Fundamentals of interactive computer graphics,* Addison-Wesley

Franchi N. (1950) *Elementi di cartografia,* Firenze

Frova A. (1984) *Luce, Colore, Visione,* Roma

Funkhouser, H. G. (1937) *Historical development of graphical representation of statistical data,* Osiris, 3

Gabaglio, A. (1988) *Teoria generale della statistica,* 2 vol. 2-da ed., Milano

Gini, C. (1933) *Due osservazioni a proposito delle rappresentazioni grafiche,* XXI sessione ISI, Città del Messico

Gini, C. (1949), *Pregi ed inconvenienti delle rappresentazioni statistiche,* XXVI sessione ISI, Berna

Graunt, J. (1987) *Osservazioni naturali e politiche fatte sui bollettini di mortalità,* traduzione di Lombardo, E., La Nuova Italia

Grimaldi, R. (1983) (a cura di): *La cartografia ed i sistemi informativi per il governo del territorio,* Franco Angeli

Hilbert, D., Cohn-Vossen, S. (1960) *Geometria intuitiva,* Boringhieri, Torino

Hoaglin, D.C., Mosteller, F., Tukey, J.W. (1983) *Understanding robust and exploratory data analysis,* Wiley

Istat (1927) *Compendio Statistico Italiano,* dal 1927, Roma

Istat (1963) *Cromografie statistiche d'Italia,* Roma

Istat (1967) *Sintesi grafica della vita economica italiana,* dal 1967 al 1970, Roma

Istat (1965) *Grafici dell'Annuario Statistico Italiano,* Roma

Istat (1988) *Atlante Statistico Italiano 1988,* Roma

Istat (1988) *Annuario Statistico Italiano 1988,* Roma

Itten, J. (1982) *Arte del colore,* Il Saggiatore

Leti, G., (1983) *Statistica descrittiva* , Bologna

Lindbeck, A. (1980) *Inflazione,* Giuffré, Milano

Lombardo, E. (1984) *Analisi esplorativa dei dati: materiali per una introduzione,* Roma

Lombardo, E. (1986) *Un carteggio rilevante per la demografia: la corrispondenza fra Christiaan e Lodewijk Huygens del 1669,* in AA.VV. Pagine in ricordo di Gianni Bellei, Dipartimento di studi geoeconomici, statistici e storici per l'analisi regionale, Roma

Luzzatto-Fegiz, P. (1934) *Alcune regole pratiche sulle rappresentazioni grafiche,* Trieste

Luzzatto, L., Pompas, R. (1980) *Il linguaggio del colore,* Milano

McDonald, J.A. (1986) *Orion I: interactive graphics in statistics,* in Wegman, E.J., DePriest, D.J. ed., Statistical image processing and graphics, D.J., Marcel Dekker Inc.

Morgan B.S. (1982) *The properties of a distance-based segregation index,* Socio-economic planning sciences

Napolitano P. (1985) *Esposizione di alcune tecniche per l'investigazione dei dati,* Quaderni di discussione Istat

Napolitano P. (1987) *Metodi grafici per la presentazione di dati statistici,* Atti della giornata di studio SIS 9 aprile 1987: l'informazione statistica nei mezzi di comunicazione di massa, Roma

Neurath O. (1936) *International picture language - The first rules of isotype* Londra

Openshaw, S. (1987) *The aggregation problem in the statistical analysis of spatial data,* Convegno SIS 1987, Perugia 5-6 ottobre

Pedroni, F. (1968) *Rappresentazioni statistiche* III ed., Milano

Salvemini, T., Girone, G. (1981) *Lezioni di statistica,* Cacucci, Bari

Samuelson, P.A., Nordhaus, W.D. (1987) *Economia,* Zanichelli

Schmid, C.F. (1978) *The role of the standards in graphic presentation,* Technical paper n. 43, Bureau of census Washington D.C.

Schmid, C.F. (1983) *Statistical Graphics,* Wiley

Shiskin, J., Young, A.H., Musgrave, J.C. (1965) *The X-11 variant of the Census method II seasonal adjustment program,* Technical paper n. 15, Bureau of the Census U.S. Department of Commerce

Simkin, D. e Hastie, R. (1987) *An information-processing analysis of graph perception,* J.A.S.A. 82, 398

Tranquilli, G.B. (1985a) *Concentrazione descrittiva tra due distribuzioni multiple ed estensione a coppie di misure normalizzate qualsiasi,* Metron, XLIII, 3-4

Tranquilli G.B. (1985b) *Concentrazione semplice tra due variabili ponderate congiunte non negative. Caso assolutamente continuo,* in AA. VV. Pagine in ricordo di G. Bellei, Dipartimento di studi geoeconomici, statistici e storici per l'analisi regionale, Roma

Tukey, J.W. (1962) *The future of data analysis,* Annals of Mathematical Statistics, 33

Tukey, J.W. (1977) *Exploratory data analysis,* Addison - Wesley, Reading MA

Tukey, J.W. (1979) *Methodology and the statistician's responsability for both accuracy and relevance,* J.A.S.A., 74, 368

Unwin D. (1986) *Analisi spaziale, un'introduzione geocartografica,* Franco Angeli

White, M.J. (1986) *Segregation and diversity measures in population distribution,* Population index, 52

Wyszecki, G., Stiles, W.S. (1982) *Color science: concepts and methods, quantitative data and formulae,* 2º edition Wiley

1 SINTESI DELLE PRINCIPALI NOZIONI ED APPLICAZIONE PRATICA DEI METODI

1. CRITERI CUI DEVONO ATTENERSI I GRAFICI RAZIONALI

I criteri cui devono attenersi i grafici razionali possono formularsi in vari modi. Indichiamo qui appresso una possibile formulazione:

a) il metodo grafico prescelto per rappresentare una serie di dati deve corrispondere alla natura della serie di dati rappresentati;

b) il grafico deve essere semplice e facile da percepire;

c) la grandezza o intensità di simboli rappresentanti dati quantitativi deve essere, per quanto possibile, proporzionale ai dati rappresentati o ad una loro funzione chiaramente definita;

d) attributi qualitativi possono essere rappresentati da simboli grafici o da colori che esprimono differenze qualitative, ma occorre evitare che i simboli qualitativi introducano impressioni erronee di differenze quantitative;

e) nei cartogrammi che rappresentano serie territoriali i simboli grafici devono essere collocati in modo adatto su punti, linee o zone cui i dati rappresentati si riferiscono;

f) serie di dati dello stesso tipo prodotte dall'Istat per scopi analoghi vanno rappresentate graficamente con lo stesso metodo, anche se compaiono in pubblicazioni differenti.

Rispetto al problema della percezione visiva dei simboli grafici, è noto che l'occhio umano è imperfetto: simboli indicanti grandezze o intensità in modo correttamente proporzionale ai dati quantitativi rappresentati, possono essere percepiti in modo inesatto e diverso dai diversi utenti dei grafici. Tuttavia i simboli grafici GRP possono permettere in futuro correzioni, nel caso in cui «leggi» relative alle proprietà della percezione possano essere esattamente definite.

Ai criteri precedenti va aggiunto il suggerimento generale che le opzioni relative ad aspetti del grafico come legenda, scala geografica, scala statistica, formato, tipo di dettaglio per circoscrizioni desiderato nei cartogrammi, impiego di colori, ecc. vanno prese a ragion veduta, utilizzando al massimo l'esperienza passata e i suggerimenti dati nel fascicolo 7.

2. NECESSITÀ DI ADATTARE IL METODO GRAFICO ALLA NATURA DELLA SERIE DI DATI DA RAPPRESENTARE

Abbiamo indicato nel paragrafo precedente che il metodo grafico scelto deve adattarsi alla natura della serie dei dati da rappre-

sentare. Le serie statistiche possono essere definite e classificate in vario modo. Qui adottiamo una definizione e classificazione avente importanza pratica dal punto di vista della rappresentazione grafica.

Indichiamo come: a) *serie statistica* una successione o insieme di dati statistici ordinati secondo le modalità di uno o più caratteri; b) *serie semplice* una serie di dati ordinati secondo le modalità di un carattere (per es. popolazione italiana al censimento 1981, secondo classi di età; produzione del frumento nel 1984 per provincia); c) *serie multipla* una serie di dati ordinati secondo le modalità di più caratteri (per es.: 1) popolazione italiana al censimento 1981, secondo sesso ed età: qui si hanno due caratteri; 2) popolazione italiana al censimento 1981, secondo sesso, età e stato civile: qui si hanno tre caratteri).

Partiamo dal concetto che uno degli scopi fondamentali del grafico è di permettere al lettore di apprezzare rapidamente le relazioni fra modalità del carattere e dati riferentesi a tali modalità: nelle serie citate sopra vogliamo far vedere per mezzo del grafico: come varia con l'età la frequenza degli appartenenti a ciascuna classe di età; come varia con l'età e il sesso la struttura per stato civile, ecc.

È quindi di importanza fondamentale, dal punto di vista grafico, considerare la natura dei caratteri secondo cui sono presentati i dati e l'ordine secondo cui si susseguono le modalità di questi caratteri.

Una classificazione dei caratteri spesso adoperata che possiamo adottare con qualche modifica è la seguente.

I caratteri secondo i quali sono classificati i dati sono:
a) quantitativi (variabili);
b) qualitativi (mutabili);
c) di tempo (cronologici);
d) territoriali (spaziali o geografici).

Si può notare che nelle serie a) e c) le modalità si susseguono in un ordine naturale da un valore minimo ad un valore massimo.

Questo facilita molto la presentazione grafica perché in generale è possibile dare tale rappresentazione su un diagramma e indicare le modalità del carattere nel loro ordine naturale lungo uno degli assi delle coordinate e le grandezze dei dati lungo l'altro asse e scoprire perciò rapidamente le relazioni fra carattere e dati. Dal punto di vista grafico le serie a) e c) possono considerarsi come serie aventi caratteri e modalità ordinate linearmente (in breve: serie lineari).

Si possono talora assimilare alle serie lineari anche «serie lineari qualitative» in cui le modalità sono espresse verbalmente o per mezzo di simboli, ma si susseguono in un ordine naturale (per es.: i gradi della gerarchia militare).

Invece gli altri caratteri qualitativi sono «sconnessi» e non hanno un ordine naturale: questa proprietà deve essere presa in considerazione nella loro presentazione grafica.

Un altro tipo di carattere (di minore importanza) è costituito da caratteri «circolari» o «ciclici» in cui le modalità si susseguono secondo un ordine naturale, ma non hanno valori estremi: per esempio la direzione dei venti (N, NE, EST, ... NW); o i giorni della settimana (lunedì, ... domenica). Per caratteri di questo tipo la presentazione naturale è circolare.

Indichiamo, per comodità, col nome di *serie parallele* serie di dati classificati nello stesso modo. Per esempio, due tabelle indicanti rispettivamente per il censimento del 1971 e del 1981 la distribuzione della popolazione per sesso, età, stato civile, contengono due serie parallele composte di tre caratteri.

3. PRESENTAZIONE GRAFICA E PRESENTAZIONE NUMERICA DEI DATI

Scopo fondamentale della rappresentazione grafica è di dare una visione generale, rapida e facile da percepire, delle relazioni fra i dati e i caratteri, secondo cui questi sono presentati. Riferendoci ad alcuni esempi, si può voler cogliere: come vari la frequenza della popolazione col crescere dell'età; come tali variazioni differiscano fra i due sessi; come vari la ripartizione per stato civile secondo il sesso e l'età; come vari la densità della popolazione o la frequenza della coltivazione del frumento sopra il territorio italiano, ecc.

Per ottenere questo scopo, il grafico deve essere semplice e facile da percepire.

Scopo sussidiario del grafico è di permettere la identificazione e valutazione dei dati singoli aventi speciale interesse. Per raggiungere questo scopo è necessario che il grafico riporti le scale dei valori e le necessarie indicazioni in una accurata legenda.

La tabella ha per compito fondamentale di fornire agli utenti informazioni precise (dettagliate o generali) di cui possono avere bisogno: è perfettamente legittimo produrre anche tabelle contenenti decine di migliaia di dati (per es.: distribuzione per classi di professione della popolazione attiva di ciascun Comune d'Italia). Una tabella di questo tipo può essere utilizzata da molti utenti ognuno dei quali ha interesse in un Comune o gruppo di Comuni. Tuttavia la «lettura» di una tabella di questo tipo per ottenere una visione generale della distribuzione geografica delle classi professionali sarebbe impraticabile. Qui solo il grafico può venire in soccorso dell'utente. Ma un utente non deve aspettarsi di «leggere» sul grafico informazioni numeriche su uno specifico Comune che lo interessa. Il grafico non può e non deve sostituire una tabella dettagliata.

4. LE RAPPRESENTANZIONI GRAFICHE DELLE SERIE STORICHE

Il metodo migliore per rappresentare le serie storiche è il diagramma cartesiano, sul quale si rappresenta la variabile tempo lungo l'asse x e la variabile statistica associata lungo l'asse y.

Tali serie vanno distinte, innanzitutto, in funzione della loro periodicità, che potrà essere generalmente: mensile, trimestrale o annuale.

In generale si tende a rappresentare per le serie mensili un periodo pari al più a 5 anni (per un numero complessivo di 60 dati): per le serie annuali si danno al più venti anni utilizzando la pagina in tutta la sua larghezza ed al più 10 anni per metà pagina.

Nella scelta della scala dell'asse y, nell'ambito di un dato argomento, per serie caratterizzate da valori dello stesso ordine di grandezza, si sceglierà una scala uguale, al fine di consentire il confronto fra i diversi fenomeni.

Nel caso che la scala del fenomeno da rappresentare è di tipo rapporto, ovvero se lo zero ha un significato intrinseco, se si sceglie di rappresentare graficamente solo l'intervallo in cui la serie assume i valori, si deve comunque indicare lo zero come origine della scala e va quindi segnalata con un simbolo standard l'interruzione dell'asse delle y.

Se si deve rappresentare una singola serie relativa ad un fenomeno di flusso, si può optare per una rappresentazione tramite istogrammi, ovvero per un diagramma a scalini, in cui la base di ogni rettangolo rappresenta la durata di tempo cui il dato si riferisce. Se si devono rappresentare più serie storiche contemporaneamente, pur trattandosi di dati di flusso, si tornerà alla rappresentazione congiunta di poligonali. Va comunque evitata la rappresentazione congiunta di serie di stato con serie di flusso.

Quando si rappresentano insieme più serie sullo stesso diagramma è importante scegliere il tipo di linea ed il colore in modo da rendere ben distinguibili fra loro le diverse serie.

Se si vuole mettere in evidenza la variazione percentuale del fenomeno ad ogni istante di tempo rispetto all'istante precedente, va usato il diagramma semilogaritmico. In questo caso le stesse serie relative a dati di flusso si presentano in modo generale tramite poligonali.

Dal momento che il logaritmo di un numero esiste solo se esso è strettamente positivo e tende a meno infinito al tendere del numero a zero, non si pone il problema dell'interruzione della scala. In tali diagrammi è importante che l'indicazione dei valori posti in corrispondenza delle suddivisioni dell'asse y faccia riferimento ai valori effettivi della serie.

Se si hanno diagrammi diversi che riportano serie differenti raggruppate in modo appropriato, gli assi dei tempi vanno allineati accuratamente.

Quando il numero dei dati è limitato si può rappresentare la serie con un diagramma a colonne, nel qual caso si immagina la serie come una successione finita di tempi distinti per i quali si vuol confrontare gli ammontari o le intensità corrispondenti.

Nel caso di diagrammi a barre non è consigliabile interrompere le barre, e se la scala parte da zero essa va rappresentata nella sua interezza.

Nei diagrammi a colonne non conviene rappresentare più di due, tre serie contemporaneamente tramite barre affiancate; è preferibile rappresentare giustapposti o uno sotto l'altro più diagrammi a colonne, ciascuno relativo ad un singolo fenomeno.

Un numero limitato di linee di riferimento è utile per guidare l'utente alla lettura del grafico.

5. LE RAPPRESENTAZIONI GRAFICHE DI FREQUENZE, AMMONTARI E QUANTITÀ DERIVATE SECONDO LE MODALITÀ DI UNO O PIÙ CARATTERI

Si tratta inizialmente di distribuzioni connesse a caratteri quantitativi.

Per le distribuzioni relative a caratteri quantitativi continui, la rappresentazione grafica corretta è l'istogramma; per la sua realizzazione occorre definire preliminarmente con precisione i limiti delle classi che coprono il campo dei valori assunti dalla variabile. Della eventuale diversa ampiezza delle classi si deve tener conto nel tracciare le basi dei rettangoli che costituiscono l'istogramma, ricordando che le aree dei rettangoli devono essere proporzionali alle frequenze di ciascuna classe.

Per la rappresentazione di serie doppie si può pensare ad istogrammi contrapposti come la piramide delle età, o costruzioni simili; per le serie multiple non conviene interrompere l'omogeneità della rappresentazione e, perciò, le diverse distribuzioni vanno affiancate orizzontalmente o verticalmente, ponendo in corrispondenza fra loro le suddivisioni in classi.

Se le serie da confrontare sono numerose, conviene sostituire l'istogramma con la poligonale con cui si interpolano i valori delle densità di frequenza. Le osservazioni sviluppate per le serie storiche sono estendibili al caso suddetto.

Per i caratteri quantitativi discreti le varie modalità si rappresentano con colonne distanziate in funzione del valore ad esse associato.

Per quanto riguarda le distribuzioni relative a caratteri qualitativi, distinguiamo il caso dei caratteri ordinabili e non.

Le distribuzioni relative a mutabili ordinabili sono rappresentate

in modo corretto da diagrammi a colonne, in cui colonne successive sono poste ad uguale distanza l'una dall'altra; le varie colonne sono associate alle modalità in modo che queste siano ordinate coerentemente con l'ordine intrinseco, da sinistra verso destra sull'asse x.

Si potrebbe pensare in questo caso a campire le varie colonne con disegni a grana, o tessitura, differente, da associare alle varie modalità.

Per le serie multiple è, in generale, scarsamente efficiente interrompere la sequenza naturale delle misure relative alle modalità del carattere con quelle omologhe di un'altra grandezza. Se si dovesse scegliere questa rappresentazione è importante selezionare i colori o la grana per ben distinguere le misure relative alla stessa modalità.

Per le distribuzioni di frequenze o quantità relative a caratteri qualitativi sconnessi, la rappresentazione che, nella generalità dei casi, si presenta come la più conveniente consiste nel diagramma a barre, sia a colonne che a nastri. Il diagramma a nastri consente una più semplice lettura delle didascalie associate alle varie modalità. Per la campitura delle barre si possono utilizzare colori diversi o linee ad orientamento diverso.

Talvolta può essere interessante ordinare le modalità del carattere qualitativo sconnesso in funzione dei valori della grandezza rappresentata.

Se lo scopo della rappresentazione è visualizzare la composizione percentuale di una singola distribuzione o della suddivisione in parti di un tutto, se il numero delle modalità non supera le 4 o 5 unità, si può utilizzare il diagramma a torta, in cui le diverse modalità siano ben distinte graficamente.

Per la rappresentazione di serie parallele associate a un carattere qualitativo sconnesso si utilizzano sistemi di rappresentazione del tipo a barre affiancate o contrapposte.

Per distribuzioni di frequenze relative a più caratteri si può costruire una sequenza di diagrammi a barre affiancati o posti uno sotto l'altro in funzione allo sviluppo delle modalità di uno dei caratteri.

Se si vuol rappresentare delle frequenze o degli ammontari che si suddividono in ammontari parziali, si può utilizzare il grafico a barre suddivise.

6. LE RAPPRESENTAZIONI GRAFICHE DI SERIE TERRITORIALI RELATIVE A DATI COMUNALI, PROVINCIALI E REGIONALI

Per le serie territoriali di tipo areale l'informazione statistica risulta associata a zone ben definite.

Generalmente le serie territoriali si presentano come ammontari, frequenze e quantità derivate associate ad un certo sistema di zone; la diversa ampiezza e forma di tali zone costituisce una grave difficoltà al fine di una semplice e corretta rappresentazione.

Il metodo dei GRP, dall'inglese «Graphical rational Patterns», costituisce un sistema razionale di rappresentazione cartogrammatica; esso, da una parte, risolve razionalmente il problema della irregolarità del sistema di zone sostituendolo con un reticolo regolare di celle esagonali, e, dall'altra, preserva il carattere quantitativo dell'informazione, che consiste per lo più di frequenze, ammontari e quantità derivate.

La possibilità di disporre di dati comunali e il fatto che un cartogramma costruito sulla base di dati comunali occupa il medesimo spazio di un cartogramma a dati regionali, al contrario di quello che accade per le tabelle, consigliano in generale l'uso di tale dettaglio territoriale.

Diamo qui di seguito una breve descrizione della metodologia GRP:

a) il territorio italiano è stato suddiviso secondo un reticolo di 1.000 celle esagonali di area uguale, pari a circa 300 km^2;

b) i dati dei Comuni vengono assegnati alle celle esagonali in base alla posizione ed alla superficie; i dati provinciali e regionali sono distribuiti in modo uniforme tra le celle associate a ciascuna provincia e regione rispettivamente; essi vengono quindi trasformati in frequenze relative e riportati ad un ammontare pari a 10.000 unità;

c) essi sono infine rappresentati mediante simboli chiamati GRP, dall'inglese «Graphical Rational Patterns»; con questi simboli le unità sono rappresentate dal numero corrispondente di quadrati di area unitaria, le decine da quadrati di area 10 volte maggiore, ecc.

La scala data in calce ad ogni grafico indica, per ognuno dei simboli effettivamente utilizzati, il valore delle frequenze assolute corrispondenti.

Per semplicità di rappresentazione si utilizza una scala abbreviata, che invece di utilizzare tutti i valori interi compresi fra 1 e 100, fa uso dei valori 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 30, 40, 50, 60, 70, 80, 90, 100; per la giustificazione teorica dell'uso di questa scala semplificata si rimanda all'introduzione dell'Atlante Statistico Italiano 1988; per una descrizione dettagliata del metodo si può far riferimento al capitolo 4 del fascicolo 7.

Si sottolinea che il simbolo 100 rappresenta i valori corrispondenti a 100 ed oltre.

Nella metodologia GRP esiste la possibilità di rappresentare le differenze fra le frequenze relative delle distribuzioni dei due fenomeni indicate con (A) e (B) rispettivamente; in tali cartogrammi

vengono rappresentate in rosso l'eccedenze delle frequenze della distribuzione (A) rispetto a quelle della distribuzione (B); in colore verde si rappresentano le eccedenze nel verso opposto. In questo modo è possibile eliminare da un distribuzione l'influenza di un fenomeno, o si può, in termini più generali, studiare la concentrazione di una coppia di fenomeni.

Nel caso della rappresentazione di quozienti si utilizza una metodologia speciale, diversa se i dati sono a base provinciale o associati alle 1.000 celle esagonali.

Nei cartogrammi dei quozienti a base provinciale il valore rappresentato da ciascun simbolo è proporzionale al quoziente della provincia cui esso si riferisce; il numero dei simboli all'interno dei confini provinciali è proporzionale all'ammontare della popolazione a denominatore del tasso nella provincia medesima.

Per le applicazioni relative alla popolazione si è arrivati ad una soluzione standard, per cui un simbolo corrisponde in generale a 175.000 abitanti; per le province di Roma, Milano e Napoli un simbolo corrisponde a 350.000 abitanti.

Nei cartogrammi dei quozienti costruiti sulla base delle 1.000 celle esagonali il valore rappresentato da ciascun simbolo è proporzionale al quoziente della cella cui esso si riferisce; il «valore» dei simboli di ciascuna cella è proporzionale all'ammontare della popolazione a denominatore del tasso nella cella medesima. Si utilizzano quattro classi, i cui estremi sono determinati dalla mediana e dai quartili della distribuzione territoriale relativa alle 1.000 celle.

Per mettere in chiara evidenza la diversità delle distribuzioni territoriali dei quozienti si possono utilizzare scale differenti per i fenomeni, in modo che i valori siano distribuiti in ogni cartogramma sull'intera scala da 1 a 100 dei simboli GRP.

Questo procedimento è concettualmente analogo a quello talora usato nei diagrammi cartesiani, in cui si indica che la scala non parte da 0 ma comprende solo i valori inclusi fra il minimo ed il massimo dei dati.

Nei cartogrammi sono sovrapposti due assi che sintetizzano le caratteristiche generali della distribuzione e si intersecano perpendicolarmente nel baricentro. La lunghezza degli assi misura la dispersione nelle due direzioni in cui essa è, rispettivamente, massima e minima.

Nei cartogrammi dei confronti fra due distribuzioni, nel caso che siano rappresentate le eccedenze dei due tipi, di (A) su (B) e di (B) su (A), vengono riportate le due coppie di assi relative alle due distribuzioni delle eccedenze.

Il metodo dei GRP è certo applicabile al caso di dati provinciali e regionali.

Per la rappresentazione di dati a livello regionale, i dati assoluti possono essere trasformati in valori rapportati ad altra grandezza, per evitare che l'effetto della diversa estensione delle varie regioni prevalga nel grafico, (ad esempio quozienti, valori di densità, etc..); per la loro rappresentazione si possono utilizzare barre ben posizionate sulle rispettive regioni.

Se lo scopo della rappresentazione è dare una rappresentazione semplificata in classi di valori, si può utilizzare un sistema di campiture con «valore» di intensità crescente; conviene utilizzare un numero di classi al più pari a 7, i cui limiti si determinano facendo sì che, in linea di massima, le varie classi contengano un numero uguale di zone; è anche utile che i valori delle classi siano multipli di 5 e 10.

Per valori associati a punti sul territorio si utilizzano cartogrammi a punti sui quali in corrispondenza a ciascuna località di presenza del fenomeno si sovrappone l'informazione, che può essere di semplice localizzazione o della sua intensità; in questo ultimo caso per rappresentare le diverse quantità si possono utilizzare i simboli GRP o cerchi di area proporzionale all'intensità del fenomeno.

2 BREVE DESCRIZIONE DI ALCUNI PRINCIPALI TERMINI USATI NEL FASCICOLO 7

Box-plot: diagramma adatto alla rappresentazione sintetica di osservazioni individuali di tipo quantitativo: lungo la direzione orizzontale, con una scala determinata, vengono indicate la posizione della mediana, quella dei due quartili, inferiore e superiore, della distribuzione; con simboli speciali vengono altresì evidenziati valori «eccezionali».

Cartogramma: rappresentazione grafica che associa ad una componente spaziale una componente descrittiva; in esso a punti, linee, superfici di un territorio determinato vengono associati i valori di una variabile statistica.

Componente: per arrivare a definire l'informazione da rendere graficamente, si individuano le componenti ovvero i costituenti concettuali dell'informazione che vanno posti in relazione; nei diagrammi si pongono graficamente in relazione due componenti; nei grafici statistici, in generale, una delle componenti è costituita dal carattere e dalle sue modalità, l'altra dagli individui o dalle frequenze od altre quantità associate alle varie modalità.

Diagramma: rappresentazione che associa graficamente una coppia di componenti al fine di evidenziarne le relazioni; se le due componenti sono entrambe quantitative si hanno i diagrammi cartesiani; se per rappresentare una delle componenti quantitative si utilizzano delle barre, si hanno i diagrammi a barre; se per rappresentare la componente quantitativa si utilizzano i settori circolari si hanno i diagrammi a torta; etc...

Diagramma a barre: diagramma in cui l'informazione relativa alle frequenze o altre quantità è data tramite la lunghezza delle barre (nastri o colonne); ciascuna barra è associata alle diverse modalità del carattere.

Diagramma a colonne: variante del diagramma a barre nel quale queste sono disposte verticalmente; la sua utilizzazione è particolarmente indicata per serie relative a caratteri quantitativi discreti e qualitativi ordinabili; viene anche utilizzato per serie storiche composte di pochi dati, se si vuole facilitare il confronto fra valori contigui piuttosto che dare l'andamento generale della serie.

Diagramma a nastri: variante del diagramma a barre nel quale queste sono disposte orizzontalmente; il suo uso è indicato per la rappresentazione di distribuzioni di caratteri di tipo qualitativo

sconnesso in quanto facilita l'associazione fra la descrizione verbale delle modalità e le rispettive barre.

Diagramma areale: rappresentazione grafica che usa l'area per rappresentare frequenze ed ammontari.

Diagramma cartesiano: rappresentazione grafica che utilizza due assi ortogonali associati a due variabili quantitative; entrambi gli assi hanno proprietà metriche ben definite.

Diagramma semilogaritmico: variante del diagramma cartesiano, in cui, mentre l'asse x mantiene la scala lineare, l'asse y è caratterizzato dalla scala logaritmica; la sua utilizzazione è particolarmente indicata quando si vuole evidenziare il tasso di crescita del fenomeno al tempo t rispetto al livello raggiunto dal fenomeno al tempo immediatamente precedente.

Distribuzione territoriale equivalente: data una serie territoriale, che associa ad un sistema di zone come Province e Comuni valori come frequenze od ammontari, questa viene trasformata nella distribuzione equivalente attraverso le seguenti operazioni:

a) le frequenze od ammontari iniziali vengono associati a 1.000 celle esagonali di area uguale, pari a circa 300 km²;

b) le frequenze od ammontari sono riproporzionati in modo che la loro somma sia pari ad M = 10.000 unità;

c) i valori così ottenuti sono approssimati al valore intero più vicino.

Grafico: tecnicamente per grafico si intende una delle seguenti costruzioni: diagramma, cartogramma, grafo ed ideogramma. Si utilizza, perciò, tale nome per indicare una costruzione grafica composta da uno o più di questi elementi che risulti ad un tempo completa dei suoi riferimenti esterni.

Grafo: grafico costituito da punti e da linee che li collegano; i punti, detti nodi, sono in corrispondenza con entità, ad esempio degli oggetti, le linee indicano l'esistenza di relazione fra le entità.

GRP: acronimo delle parole inglesi Graphical Rational Patterns; si traduce in italiano in simboli grafici razionali (s.g.r.); indica un sistema di simboli basato su semplici criteri che consente la rappresentazione degli interi da 1 a 100.

Ideogramma: tipo di grafico che riporta e trasmette l'informazione tramite l'uso di forme simboliche, che possono essere di tipo

generico o specifico; esso è di tipo specifico se con la sua forma si cerca di richiamare direttamente un oggetto o un concetto.

Istogramma: diagramma adatto alla rappresentazione di distribuzioni di frequenza o di quantità relative a caratteri quantitativi continui; il suo uso è raccomandato quando i dati sono noti rispetto a classi di valori della variabile.

Piramide delle età: diagramma formato da una coppia di istogrammi posti specularmente rispetto all'asse verticale associato alla variabile età; le due distribuzioni si riferiscono ai due sessi e le età sono generalmente raggruppate in classi di valori.

Stem-and-leaf: diagramma raccomandato per la rappresentazione di osservazioni individuali di tipo quantitativo; in esso si fa un uso grafico delle cifre che compongono i valori delle singole misure.