12 LE MATRICI DI DATI

di Giampiero D'Alessandro

I. Introduzione

L'organizzazione dei dati in matrice è un passaggio preliminare, necessario per rispondere, attraverso opportune analisi, ai propri interrogativi di ricerca. Sia che si stia effettuando un'indagine quantitativa su dati primari (ad es. questionari più o meno strutturati, *cfr*. Cap. 6), sia che si utilizzino dati secondari (cfr. Cap. 18), i programmi di analisi statistica[1] hanno necessità di operare su *dati strutturati*. L'obiettivo dell'organizzazione dei dati è rendere le informazioni (raccolte direttamente o collezionate da altre fonti) analizzabili, ovvero *preparare* i dati in modo tale che essi possano essere analizzati (o interrogati se si preferisce) con il/i software di analisi statistica prescelto/i[2].

È possibile definire una matrice di dati come «una qualsiasi disposizione di informazioni (numeriche e non) ordinate per riga e per colonna» (Di Franco 2010, p. 31). Nella sua forma più comune (*cfr*. Par. 1.1), è possibile pensare alla matrice dei dati come a una griglia in cui vengono archiviate le informazioni relative ai *caratteri* (variabili) di determinate *unità statistiche* (casi).

L'unità statistica è «l'elemento di base della popolazione sul quale viene effettuata la rilevazione o la misurazione di uno più fenomeni oggetto di indagine» (Piccolo 2010, p. 39). È bene precisare subito che le unità statistiche possono essere

[1] R, SPSS e Stata sono, tra i programmi di analisi statistica, più noti e utilizzati. In questo capitolo si farà talvolta riferimento ad alcune funzioni di questi programmi attinenti alla gestione dei dati in matrice.

[2] Il singolare/plurale è dovuto al fatto che, solitamente, si effettua un utilizzo congiunto di più programmi di analisi dei dati. Se per inserire i questionari in matrice (si veda oltre) è sufficiente l'utilizzo del programma Excel di Microsoft, utile anche per le analisi mono, bi e tri variate tramite la funzione "tabelle pivot", per analisi più specifiche sarà necessario ricorrere a programmi di analisi statistiche completi (*comprehensive*), come quelli citati in precedenza, o specifici per particolari tipi di analisi dei dati.

singoli individui (nel caso di un'indagine con questionario i casi saranno i rispondenti al questionario) quanto altri tipi di unità quali, ad esempio, unità amministrative territoriali (gli stati europei, le regioni italiane, i comuni, ecc.) o aggregati di individui (famiglie, scuole, aziende, associazioni, università ecc.). Ma ancora unità di analisi possono essere beni materiali (abitazioni, computer, case, autovetture ecc.) o immateriali (come gli strumenti finanziari, i brevetti o le canzoni[3] ecc.).

Anche i caratteri, ossia le informazioni rilevate sulle unità statistiche, possono essere di diverso tipo. Le informazioni raccolte, relativamente a ciascuna unità statistica possono essere di tipo *qualitativo* (testuale) o di tipo *quantitativo* (numerico). Sono informazioni testuali, ad esempio, la professione oppure il genere (maschio o femmina) o il titolo di studio. Informazioni numeriche sono invece la sua età, il numero di fratelli e sorelle, gli anni di istruzione.

Affinché le informazioni afferenti a un certo numero di unità statistiche (casi) possano essere organizzate in una matrice dati sono necessarie due condizioni:

- unicità delle unità;
- identicità delle informazioni (Corbetta 1999).

La prima delle due condizioni fa riferimento al fatto che in una matrice dei dati tutte le informazioni raccolte[4] devono fare riferimento alle medesime unità statistiche. Ciò vuol dire che l'*oggetto dell'osservazione* (l'unità statistica) deve *necessariamente* essere lo stesso ossia tutte le informazioni che si intendono organizzare in matrice devono riferirsi *o* a individui *o* a regioni *o* a scuole ecc. In altri termini non è possibile avere un'unica matrice dati che contenga informazioni che riguardano unità statistiche differenti. La seconda condizione è che per ciascuna delle unità devono essere rilevate le stesse informazioni. Non è possibile cioè che per alcune unità vengano raccolte informazioni differenti da quelle raccolte per altre unità.

Il presente capitolo ha l'obiettivo di familiarizzare lo studente con i concetti inerenti le (il plurale è voluto) matrici dei dati (e alcune altre questioni) che con maggiore frequenza vengono utilizzate nelle scienze sociali. Si intende cioè illustrare i diversi tipi di organizzazione dei dati che possono essere progettati o nei quali ci si può facilmente imbattere durante le proprie esperienze di ricerca. Il capitolo vuole essere un documento utile ad aumentare la consapevolezza e che consenta di prevenire (o evitare) possibili problemi in fase di analisi dei dati.

Il capitolo prende avvio dalla definizione e dall'illustrazione della forma di matrice più comune: la matrice casi per variabili (*Cfr*. Par. 1.1) e due suoi formati. Il secondo paragrafo riguarda l'organizzazione dei dati in matrice e gli strumenti necessari per una corretta organizzazione, primo fra tutti il *codebook*. In esso si

[3] In caso le unità statistiche siano dei testi, come per le canzoni, le informazioni relative alle unità (le variabili), non strutturate in origine, dovranno necessariamente subire delle trasformazioni di strutturazione (*cfr*. Cap. 23 e Cap. 25).

[4] Siano esse raccolte per via diretta (survey) o per via indiretta (indagini secondarie).

illustrano i passaggi necessari, da un corretto inserimento ai controlli preliminari all'analisi vera e propria, che portano la ricerca verso la sua fase conclusiva, ossia la diffusione dei risultati (*cfr*. Par. 1.2).

1.1 La matrice casi per variabili

La matrice casi per variabili è la più ricorrente tra le forme di organizzazione dei dati[5]. Indicata in modo sintetico come *matrice C x V*, è una struttura di organizzazione del dato che prevede in riga le unità statistiche, ossia i casi (C), e in colonna le informazioni raccolte, ossia le variabili (V).

È possibile rappresentare la matrice C x V come una griglia (*cfr*. Tavola 11.1) in cui vengono organizzate le informazioni raccolte sull'insieme delle unità statistiche oggetto di indagine[6].

Tavola 11.1 Schema di matrice Casi per Variabili (C x V)

	V_1	V_2	V_3	V_4	…	V_j
C_1	S_{11}	S_{12}	S_{13}	S_{14}	…	S_{1j}
C_2	S_{21}	S_{22}	S_{23}	S_{24}	…	S_{2j}
C_3	S_{31}	S_{32}	S_{33}	S_{34}	…	S_{3j}
C_4	S_{41}	S_{42}	S_{43}	S_{44}	…	S_{4j}
…	…	…	…	…	…	…
C_i	S_{i1}	S_{i2}	S_{i3}	S_{i4}	…	S_{ij}

Ogni *cella* della matrice, cioè ogni incrocio fra ciascuna riga e ciascuna colonna, contiene una informazione sul singolo caso. Ciascuna cella della matrice è identificabile attraverso due indici di posizione, identificati convenzionalmente con i e j. Il primo dei due indici è relativo ai casi e varia da 1 a i, il secondo è relativo alle variabili, e varia da 1 a j.

Presa singolarmente, ciascuna riga della matrice dei dati costituisce un *vettore riga*: in questo vettore sono registrate tutte le informazioni relative al singolo caso.

---

[5] Sebbene la matrice Casi per Variabili (C x V) sia il tipo di matrice più diffuso e utilizzato, non è infrequente incontrare tipi di matrice differenti, tra queste la matrice Variabili per Variabili (V x V), di cui si tratterà nel Cap. 17, e la matrice Casi per Casi (C x C) utile alla network analysis (*cfr*. Cap. 26).

[6] In statistica questo insieme è definito *collettivo* o *popolazione statistica*. Un collettivo statistico è formato da un insieme di unità statistiche che sono omogenee rispetto ad alcuni caratteri (l'*unicità delle unità* di cui si diceva nell'introduzione a questo capitolo) per i quali si acquisiscono, ai fini di studio, determinate informazioni (*identiche* per tutte le unità, per riprendere quanto detto poc'anzi).

Quando oltre alle informazioni rilevate è contenuto anche l'indice di posizione, o identificativo[7], del singolo caso si ha un intero *record*.

Leggendo la matrice in verticale, ciascuna delle colonne costituisce invece un *vettore colonna*: in ognuno di questi sono registrate le informazioni di tutte le unità statistiche su una determinata variabile (cfr. figura 11.1).

La *dimensione* della matrice è data dal prodotto del numero delle righe per quello delle colonne.

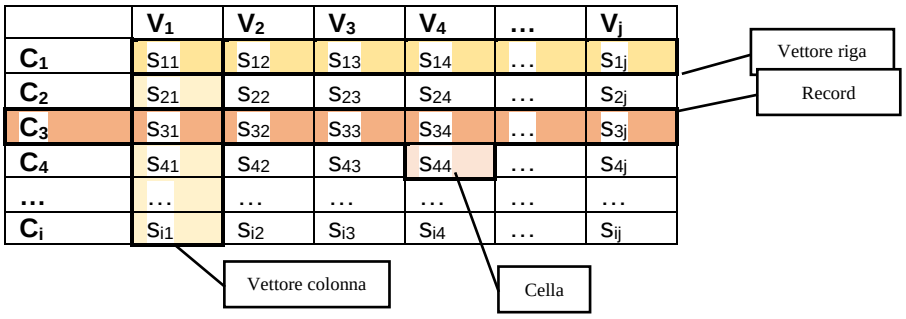


Figura 11.1 Composizione della matrice dati

Nel caso di un'indagine da questionario, ad esempio, in riga si avranno i casi raggiunti dal questionario (i rispondenti), in colonna le informazioni raccolte tramite lo strumento di rilevazione (il questionario).
Tutte le informazioni raccolte sul questionario somministrato al Caso 1 sono contenute nel primo *vettore riga* della matrice, identificata dalla riga C1, identificativo del record; (*cfr*. tabella 11.1), mentre le professioni di tutti gli intervistati sono registrate nel primo *vettore colonna* (variabile "Professione").

Tabella 11.1 Esempio di matrice Casi per Variabili (C x V)

	Sesso	**Età**	**Titolo di studio**	**Condizione occupazionale**	**Professione**	**…**	**V_j**
C1	M	22	2	Non occupato	Studente	…	s_{1j}
C2	F	41	4	Occupato	Impiegato	…	s_{2j}
C3	M	35	3	Occupato	9	…	s_{3j}

[7] È consigliabile abbinare ad ogni singolo questionario, in caso di rilevazione *paper and pencil*, un codice identificativo univoco da riportare sul questionario e in matrice dati. Questo codice consentirà un agevole ritorno sul questionario nel caso in cui durante le fasi di analisi dei dati si riscontrino delle incongruenze o degli errori effettuati durante la fase di inserimento dati.

C4	F	62	1	Occupato	Avvocato	…	s_{4j}
C5	M	18	2	Non occupato	8	…	S_{5j}
…	…	…	…	…	…	…	…
C_i	s_{i2}	s_{i3}	s_{i4}	s_{i5}	s_{i1}	…	s_{ij}

Le celle della matrice possono contenere valori testuali, numerici o codici. Ogni tipo di informazione contenuta nella matrice dovrà essere trattata (statisticamente) con le misure/tecniche più adeguate (*cfr*. resto del volume e, in particolare, *cfr*. Cap. 13). Nell'esempio sopra riportato la variabile sesso contiene dei codici (alfabetici in questo caso) che rappresentano sinteticamente l'informazione. L'informazione sottesa dal codice è evidente: il carattere "M" corrisponde a Maschio, "F" a Femmina. Mentre la variabile età è composta da informazioni di tipo numerico (anni compiuti), come per il sesso, l'informazione relativa al titolo di studio è codificata, questa volta attraverso codici numerici. In questo caso è necessario esplicitare il significato dei valori (codici) inseriti in matrice al fine di una univoca interpretazione delle modalità della variabile[8]. Ciò per consentire una corretta analisi[9] e interpretazione dei dati e, successivamente, la restituzione dei risultati.

Di tipo testuale sono le informazioni registrate per le variabili condizione occupazionale e professione. Per quest'ultima si noterà che un caso (C3), pur dichiarando di essere occupato alla domanda precedente, non ha indicato la professione. In questo caso è stato inserito nella cella della matrice il valore 9, convenzionalmente utilizzato per i mancanti non dovuti (l'intervistato avrebbe dovuto rispondere alla domanda ma non ha risposto). Parimenti nella cella relativa alla professione del quinto caso (C5) è presente il valore 8, utilizzato solitamente per i mancanti dovuti (l'intervistato ha risposto alla domanda perché non pertinente, in questo caso perché non occupato)[10].

Un altro esempio comune di matrice C x V è la matrice di dati ecologici. Solitamente con la locuzione "dati ecologici" si fa riferimento a informazioni (variabili) che hanno come unità statistiche (casi) un territorio o, meglio, un aggregato territoriale (Pintaldi 2003). Il livello di aggregazione può assumere diverse connotazioni, più o meno estese. Sono dati ecologici le informazioni riferite ai

[8] Quasi la totalità dei software statistici consente di attribuire delle etichette ai valori numerici. Ad ogni modo, è opportuno che le chiavi di interpretazione dei codici (testuali e numerici) utilizzati nella matrice dati vengano rese pubbliche attraverso il *codebook* (si veda oltre).

[9] Il ricorso a indagini online (*cfr*. Cap. 4) può comportare la necessità di trasformare etichette archiviate, automaticamente, come informazioni testuali in informazioni numeriche utili ad essere processate attraverso le opportune tecniche di analisi. È il caso, ad esempio, di una scala ancorata semanticamente (*cfr*. Cap. 7) in cui, nella stessa variabile, saranno presenti valori testuali per i poli positivo e negativo e valori numerici per i gradienti centrali.

[10] Trattandosi di codici convenzionali, è estremamente importante esplicitare il significato dei codici dei valori mancanti nel *codebook* (si veda oltre).

comuni o alle provincie, alle regioni o agli stati. Ma il livello di aggregazione può essere anche sub comunale (unità catastali, collegi elettorali o, per alcune città metropolitane come Roma, zone urbanistiche) o sovra nazionale (sub continentali o continentali)[11]. Nell'esempio riportato in Tabella 11., le unità statistiche sono le regioni italiane e le informazioni organizzate in matrice sono relative alla superfice (in ettari – *ha*) e alla popolazione (ultime tre colonne).

Quando si intraprende un percorso di ricerca, per rispondere ai propri obiettivi cognitivi il ricercatore può decidere di adottare due principali strategie: 1) costruire strumenti idonei a ottenere informazioni che, una volta organizzate e interrogate, possano rispondere alle domande di ricerca che si era posto; 2) utilizzare dati secondari già organizzati in matrice (anche provenienti da più fonti, *cfr*. Cap. 18). Nel primo caso è evidente il ricorso a strumenti quali il questionario, utili a raccogliere le informazioni necessarie sulle unità statistiche selezionate[12]. Nel secondo caso, invece, si fa ricorso, ad esempio, a indagini provenienti da statistiche ufficiali[13] oppure da banche dati pubbliche o di altro tipo[14].

---

[11] Diverse sono le ripartizioni territoriali comunemente utilizzate. Tra le più note in ambito statistico europeo, e dunque da inserire nel proprio bagaglio di conoscenze, vi è la classificazione NUTS (Nomenclature of Territorial Units for Statistic) utilizzata dal sistema statistico europeo (Eurostat). La classificazione attuale (2021) suddivide gerarchicamente il territorio europeo in 3 livelli. Il primo livello (NUTS 1) è il livello maggiore (gerarchicamente più alto, che comprende cioè i successivi) e identifica 92 regioni. Il secondo livello (NUTS2) è il livello medio e conta 242 unità territoriali. Al terzo livello (NUTS 3) troviamo il livello più piccolo che individua 1.166 regioni. Per maggiori informazioni e approfondimenti https://ec.europa.eu/eurostat/web/nuts/background

[12] Oltre che ai soggetti (i rispondenti, *cfr*. Cap. 4), uno strumento con la struttura simile ad un questionario può essere progettato anche per interrogare dei testi (in questo caso è denominato scheda d'analisi, *cfr*. Cap. 8).

[13] Sono definite statistiche ufficiali le statistiche che presentano «i caratteri dell'imparzialità, dell'affidabilità, dell'obiettività, dell'indipendenza scientifica, dell'efficienza economica e della riservatezza statistica» (TFE, art. 338). Per l'Italia il SISTAN (SIstema STAtistico Nazionale) fornisce al Paese e agli organismi internazionali l'informazione statistica ufficiale mettendo in relazione soggetti pubblici e privati. L'ISTAT (Istituto Nazionale di Statistica) ha un ruolo di indirizzo, coordinamento, promozione e assistenza tecnica. Per maggiori informazioni www.sistan.it.

[14] Vari sono gli enti, con diversa natura giuridica, che mettono a disposizione della comunità i propri dati. Quasi sempre l'unico vincolo richiesto all'utilizzatore è la citazione della fonte. Il concetto di Open Data (e più in generale di Open Government) richiederebbe una trattazione dedicata. Ad oggi, in rete sono disponibili diversi portali a carattere sovranazionale (https://index.okfn.org/; https://ourworldindata.org/; https://data.world/datasets/open-data; https://data.europa.eu/en) o nazionale (https://www.dati.gov.it/) che collezionano dati open provenienti da diverse fonti. Altre volte invece le banche dati sono private. È il caso dei dati raccolti da società di marketing, ad esempio, il cui interesse principale e quello del profitto, dei dati di profilazione, come i dati di navigazione o i dati di società la cui ricchezza primaria consiste proprio nei dati (ad esempio Google). In alcuni casi è possibile avere accesso a banche dati private (complete o parziali) facendone richiesta. Ad esempio, è possibile richiede di operare su microdati di ricerche condotte in precedenza da altri ricercatori. Alcune volte i dati vengono rilasciati a pagamento.

Tabella 11.2 Esempio di matrice ecologica casi per variabili (C x V)

Regione	Sup. (ha)	Bosco (ha)	Altre terre boscate (ha)	Pop. maschile	Pop. femminile	Pop. Totale
Abruzzo	1.083.150	391.492	47.099	625.585	655.427	1.281.012
Basilicata	1.007.311	263.098	93.329	267.989	277.141	545.130
Calabria	1.522.161	468.151	144.781	907.985	952.616	1.860.601
Campania	1.367.060	384.395	60.879	2.739.038	2.885.222	5.624.260
…	…	…	…	…	…	…
Lazio	1.723.172	543.884	61.974	2.767.173	2.963.226	5.730.399
Liguria	541.615	339.107	36.027	728.845	789.650	1.518.495
…	…	…	…	…	…	…
Puglia	1.954.052	145.889	33.151	1.913.253	2.020.524	3.933.777
Sardegna	2.409.945	583.472	629.778	778.110	811.934	1.590.044
Sicilia	2.583.255	256.303	81.868	2.346.759	2.486.946	4.833.705
…	…	…	…	…	…	…
Veneto	1.834.537	397.889	48.967	2.391.165	2.478.665	4.869.830

Fonti: Istat e Crea

1.2 L'unione di più matrici

Si può dare il caso di avere la necessità di unire le informazioni, riferite alle medesime unità statistiche, organizzate in due o più matrici di dati. In Tabella 11., ad esempio, sono state unite fonti provenienti da due banche dati pubbliche: per le variabili relative alla popolazione si è utilizzato il Censimento Permanete della popolazione[15]; per i dati sulla superficie dati dell'Inventario Nazionale delle Foreste e dei Serbatori Forestali di Carbonio (INFC) del Consiglio per la Ricerca in agricoltura e l'analisi dell'Economia Agraria (CREA)[16]. Di seguito si illustreranno brevemente le principali casistiche.
Prendiamo, ad esempio, due matrici di dati (M_1 e M_2) ciascuna con tre casi e due variabili, di cui la prima (V_1) identifica il caso e ha funzioni di chiave.

[15] http://dati-censimentipermanenti.istat.it/
[16] https://www.sian.it/inventarioforestale/jsp/01tabelle_superficie.jsp

M_1

V_1	V_2
C_1	a
C_2	b
C_3	c

M_2

V_1	V_3
C_1	X
C_2	Y
D_3	Z

Nelle matrici, due dei tre casi (C_1 e C_2) hanno il medesimo identificativo mentre altri due casi (C_3 e D_3) appaiono esclusivamente in una delle due matrici. Ciascuna delle due matrici contiene, per i propri casi, variabili differenti (V_2 e V_3). Avendo come obiettivo quello di unire le informazioni presenti nei due insiemi di dati per venire incontro alle proprie esigenze cognitive, è possibile effettuare diverse operazioni di combinazioni.
Stabilendo quale delle due matrici è l'*insieme di dati attivo* si può decidere se aggiungere informazioni (variabili) alla prima matrice (M_1) unendo per ciascuno dei casi della matrice le informazioni contenute nella seconda matrice (M_2) per i soli casi che hanno la medesima chiave (V_1) identificativa.

V_1	V_2	V_3
C_1	a	X
C_2	b	Y
C_3	c	ND

In modo speculare, decidendo che l'insieme di dati attivo è la seconda matrice (M_2) si possono unire alle informazioni in essa contenute quelle contenute nella prima (M_1) per i soli casi corrispondenti.

V_1	V_3	V_2
C_1	X	a
C_2	Y	b
D_3	Z	ND

Alternativamente, si può poi decidere di mantenere nel nuovo insieme di dati i soli casi che sono presenti in tutte e due le matrici (M_1 e M_2)

V_1	V_2	V_3
C_1	a	X
C_2	b	Y

oppure di effettuare un'unione completa tra i due insiemi, dando luogo in questo modo a una nuova matrice che contiene tutti i casi e tutte le variabili presenti nelle due matrici originarie

V_1	V_2	V_3
C_1	a	X
C_2	b	Y
C_3	c	ND
D_3	ND	Z

I programmi di analisi più noti contengono funzioni specifiche per l'unione di matrici (file) di dati attraverso chiavi. SPSS, ad esempio, mette a disposizione dell'utente la funzione *aggiungi variabili* del menu data che consente di unire alla matrice di dati attiva un'altra matrice che contiene gli stessi casi (righe) ma variabili (colonne) differenti[17] (IBM 2021, p.109). STATA invece offre la funzione *merge* (Data > Combine datasets > Merge two datasets) per l'unione di due matrici di dati (STATA 2021). La funzione merge consente di effettuare unioni uno-a-uno, uno-a-molti, molti-a-uno e molti-a-molti lavorando con chiave o per ordinamento[18]. Il software R invece, nella sua funzione base, utilizza la funzione *merge* per l'unione di due o matrici di dati (R 2021) ma molto numerosi sono i pacchetti applicativi che consentono di effettuare numerose operazioni di combinazione tra due insieme di dati, in relazione alle proprie esigenze di analisi. Fra tutti la libreria *dplyr*[19]. è tra le più ricche di funzionalità per le operazioni di unione di matrici di dati.

Infine, un accenno all'eventualità in cui si vogliano unire due banche dati prive di identificativi univoci. Qualora non si disponga nelle due matrici di una chiave univoca, è impossibile effettuare un'unione dei due insiemi di dati per via deterministica. È possibile, tuttavia, ricorrere ad altre tecniche più complesse (denominate tecniche di Record Linkage probabilistico) che, trattando congiuntamente più variabili presenti nelle due matrici, consentono di individuare (per via probabilistica) l'identicità di due casi in insiemi di dati differenti.

[17] SPSS consente di utilizzare diversi metodi unione: uno-a-uno in base all'ordinamento del file; uno-a-uno in base a valori chiave; uno-a-molti in base a valori chiave.

[18] Per STATA da riportare anche la funzione *joinby* che combina orizzontalmente due matrici formando tutte le combinazioni a coppie all'interno di un gruppo e la funzione *frlink* che consente di mettere in relazione due matrici di dati senza unirli fisicamente.

[19] Gli esempi riportati nel testo sono ripresi dal cheat sheet del pacchetto applicativo (https://dplyr.tidyverse.org/) e riprendono, in ordine, le funzioni *left_join*, *right_join*, *inner_join* e *full_join*.

1.3 Formati delle matrici

Convenzionalmente (e anche secondo quanto sin qui esposto) siamo abituati a pensare la matrice dei dati in un formato largo, ampio (*wide format*). La matrice dei dati è infatti spesso definita come insieme rettangolare di numeri organizzati in modo tale che a ciascuna riga corrisponda un unico caso (unità di analisi) e a ciascuna colonna corrisponda un'unica variabile.

V_1	V_2	V_{t3}	V_{t4}	V_{t5}
C_1	A	d	g	j
C_2	B	e	h	k
C_3	C	f	i	l

Nell'esempio sopra riportato abbiamo tre casi (C_1, C_2 e C_3) per i quali, oltre a informazioni di carattere generale (V_2) sono contenute misurazioni di uno stesso concetto ripetute in momenti differenti (V_{t3}, V_{t4} e V_{t5}). Diversi programmi di analisi dei dati (o pacchetti applicativi) richiedono una organizzazione differente della matrice dati. Nel formato esteso (*long format*) l'unità di analisi non è più il soggetto, ma la singola occasione di misurazione.

V_1	V_2	V_3	V_4
C_1	A	V_{t3}	d
C_2	B	V_{t3}	e
C_3	C	V_{t3}	f
C_1	A	V_{t4}	g
C_2	B	V_{t4}	h
C_3	C	V_{t4}	i
C_1	A	V_{t5}	j
C_2	B	V_{t5}	k
C_3	C	V_{t5}	l

Questa forma di organizzazione del dato in matrice consente di avere (rispetto alla precedente) un'unica variabile (V_4) relativa alla proprietà oggetto di analisi e una nuova variabile (V3) che identifica il momento di misurazione. Va da sé che questo formato è ottimale per tutte quelle che possiamo definire *misurazioni ripetute* nel tempo.
Con questo formato, tutte le variabili caratterizzate da stabilità temporale (ossia le variabili che non subiscono modifiche nelle diverse rilevazioni – V_2), avranno lo stesso valore per ciascuna unità statistica (V_1). Le informazioni contenute nelle due matrici sono le medesime; stiamo solo impostando i dati in modi diverso.

Con il formato long ma troviamo l'informazione che varia nel tempo (V_3) in un'unica colonna (V_4).

La scelta di un formato piuttosto che un altro è dettata esclusivamente dal tipo di analisi che si vuole effettuare. La maggior parte delle analisi richiede un formato wide. Tale formato è sicuramente più agevole anche per attività preliminari all'analisi dei dati quali la pulizia dei dati in matrice e le eventuali ricodifiche. Analisi più specifiche, quali ad esempio l'analisi di sopravvivenza (*survival analysis*) richiedono un'impostazione dei dati in formato long. In pratica, utilizzando l'unità temporale di rilevazione come unità di analisi è più semplice utilizzare una variabile come covariata di un'altra variabile. I principali pacchetti di analisi dei dati offrono strumenti, più o meno assististi, che consentono di passare dal formato wide al formato long o viceversa[20].

2. L'organizzazione dei dati in matrice

Progettare una matrice dei dati è uno passaggio fondamentale quando si ha intenzione di risolvere uno specifico problema di ricerca utilizzando i propri strumenti di raccolta delle informazioni. Si fa riferimento non solo alla *survey research* (*cfr*. Capp. 4 e 6), con annesse o meno informazioni circa gli atteggiamenti (*cfr*. Cap. 7), ma anche a specifiche attività di analisi di dati testuali, come l'analisi del contenuto (cfr. Cap.7). Il disegno della ricerca prevede, dopo la fase di formulazione e concettualizzazione del problema, altre due fasi che precedono la redazione del rapporto di ricerca e dunque la diffusione dei risultati (*cfr*. Cap.2). Sebbene le operazioni di organizzazione e trattamento dei dati, propedeutiche alla vera e propria analisi dei dati, si inseriscano di diritto nella penultima fase del disegno della ricerca, quella appunto del trattamento e analisi (*ibidem*), è opportuno se non necessario iniziare a pensare all'organizzazione dei dati nella terza fase, ossia durante la costruzione della base empirica. Durante la progettazione degli strumenti di rilevazione, infatti, è conveniente rivolgere uno sguardo al futuro, ossia a come i dati raccolti per mezzo di questi saranno poi organizzati (in matrice) per il trattamento statistico che porterà alla redazione del rapporto di ricerca. Come già osservato (*cfr*. Cap.2), non esiste una successione netta e lineare tra le diverse fasi del disegno della ricerca. A volte è fondamentale ritornare sui propri passi al fine di affinare le scelte effettuate durante le fasi precedenti. Tuttavia, quando la raccolta delle informazioni è stata avviata, è difficile (se non

[20] SPSS, ad esempio, propone la procedura guidata "Ristruttura"; in R esistono le funzioni pivot_longer() and pivot_wider() del pacchetto tidyr che, assieme al pacchetto dplyr, ha ottime funzionalità per il Data Wrangling; con SAS è possibile utilizzare la funzione proc transpose; con STATA la funzione reshape.

impossibile nella maggior parte dei casi) intervenire con modifiche sugli strumenti di rilevazione (i questionari). Pertanto, è bene progettare la matrice dati parallelamente alla costruzione degli strumenti di rilevazione onde prevenire disallineamenti tra quanto si intende rilevare per rispondere alle proprie domande di ricerca (dalle proprietà più semplici ai concetti più astratti) attraverso l'analisi dei dati.

Sia che si utilizzino schede di analisi del contenuto (è il caso dell'analisi del contenuto come inchiesta), sia che si utilizzino questionari rivolti a soggetti rispondenti (è il caso della *survey*), dunque, la progettazione degli strumenti utili a custodire, in maniera organizzata, le informazioni raccolte (la matrice dati) dovrebbe avvenire contestualmente all'impaginazione del questionario.

Quando si costruisce un questionario (stiamo dunque parlando di analisi primaria di dati, *cfr*. Cap. 4) è buona prassi pensare a come l'informazione che intendiamo raccogliere sarà poi archiviata in matrice per poi essere analizzata.
In questo paragrafo si farà, per semplicità, riferimento al caso più comune di matrici dati (la matrice casi x variabili) e gli esempi faranno principalmente riferimento a indagini condotte attraverso questionario cartaceo[21], siano esse somministrate da un intervistatore o autosomministrate.

Nell'esempio di matrice effettuato nel Par. 1.1 (*cfr*. Tabella 11.), abbiamo visto che le informazioni che raccogliamo possono essere riportate fedelmente all'interno della matrice dei dati ma (ed è questo il caso più comune) possono anche essere codificate. Per ogni informazione cioè possono (devono in alcuni casi) essere previsti dei codici, numerici, alfabetici o, a volte, alfanumerici. Ciò per una serie di motivi fra i quali: la semplicità dell'inserimento dei dati in matrice (ossia la trasmigrazione delle informazioni dal questionario cartaceo alla matrice dei dati[22]), il risparmio di spazio di archiviazione[23] e la necessità, per alcuni tipi di analisi di dati, di lavorare con un determinato tipo di variabili.

---

[21] Altre forme di somministrazione computer assistita, come la CATI (Computer Assisted Telephone Interview) e la CAPI (Computer Assisted Personal Interview) prevedono l'utilizzo di strumenti digitali in vece di quello cartaceo. Questi strumenti, solitamente, utilizzano software che consentono di progettare in modo automatico la matrice dati contestualmente alla progettazione del questionario.

[22] La fase di inserimento dei dati in matrice è una fase del processo di ricerca importante ma spesso sottovalutata, e oggi ancora di più dato il sempre più frequente ricorso al web per condurre survey (*cfr*. Cap. 6). Nei casi di questionari online l'inserimento dei dati in matrice è effettuato automaticamente dai software che registrano, organizzandole, le risposte fornite ai questionari direttamente in matrice dati (nota bene: ciò non elimina la necessaria fase di controllo e pulizia delle informazioni raccolte di cui si tratterà in seguito). Nei casi in cui si effettui una ricerca con questionario per via tradizionale (cartaceo) i dati registrati dall'intervistatore (o dall'intervistato se il questionario è autosomministrato) sulla carta andranno digitalizzati. Da ricordare che il progetto di inserimento (manuale) dei dati in matrice è estremamente *time comsuming*, e che la qualità/completezza/correttezza delle informazioni registrate dipende dall'attenzione con cui questa attività di *data input* è condotta.

[23] Ogni carattere in codice ASCII occupa un byte di memoria (unità minima di misura dell'informazione). Così, ad esempio, l'informazione "F" della variabile sesso occupa 1 byte mentre la forma

Per tali motivi, progettando un questionario, una buona pratica può essere quella di incorporare nello stesso la codifica dei dati. Il questionario con *codifica incorporata* propone, per ciascuna domanda che prevede più di una alternativa di risposta (anche a scelta multipla), accanto a ciascuna modalità un codice che sarà poi riportato in matrice dati durante la fase di inserimento dati (*cfr*. Tavola 11.2).

Tavola 11.2 Esempio di domande di un questionario con e senza codifica incorporata

Codifica incorporata	**Senza codifica incorporata**
Qual è il tuo titolo di studio?	Qual è il tuo titolo di studio?
1. □ Licenza Elementare	○ Licenza Elementare
2. □ Licenza Media	○ Licenza Media
3. □ Diploma	○ Diploma
4. □ Laurea o superiore	○ Laurea o superiore

A differenza del questionario *senza codifica incorporata*, questa soluzione riporta diversi vantaggi. Primo fra tutti, semplifica le operazioni di inserimento dati in matrice e secondo, ma non per ordine di importanza, il questionario stesso costituisce (parte) del libro codice (o codebook).

2.1 Il codebook e la definizione delle variabili in matrice

Il *codebook* è un elemento fondamentale che deve necessariamente accompagnare la matrice dei dati. In esso sono contenute tutte informazioni utili alla lettura e interpretazione delle informazioni contenute in ciascuna delle variabili (colonne) che compongono la matrice dati. Per ciascuna variabile il codebook deve contenere informazioni sul nome delle variabili, l'etichetta, il tipo di codifica, la scala di riferimento e i codici mancanti. Possono essere presenti poi altre informazioni ausiliarie.

Il *nome della variabile* solitamente è un codice che corrisponde al numero (progressivo) della domanda del questionario. Poiché convenzionalmente i numeri interi positivi vengono associati ai casi (le righe) della matrice dati, i principali programmi di analisi dei dati impongono che il nome della variabile inizi con un carattere testuale[24]. È consigliabile utilizzare come nome delle variabili codici non eccessivamente lunghi. Si può pensare, ad esempio, di utilizzare un codice

estesa (femmina) occuperebbe 7 byte di memoria. Minore è la quantità di spazio occupata dall'informazione minore saranno le capacità computazionali richieste all'hardware per la fase di analisi dei dati.

[24] Esistono anche altri vincoli, specifici per ciascun programma di analisi. Ad esempio, R, oltre a non accettare un numero come carattere inziale, non accetta spazi. In caso di conversione/importazione da altri formati, R sostituisce lo spazio con un punto (.) e trattini alti (-).

alfanumerico dove il primo sia testuale (ad esempio *d* per domanda o *v* per variabile) e i secondi, numerici, riportino l'ordine di domanda nel questionario. Ad esempio, la variabile nominata *d01* indicherà la variabile che fa riferimento alla prima domanda del questionario, *d10* la variabile della matrice contenente le informazioni della decima domanda del questionario.

Sebbene sia opportuno che l'ordine delle variabili della matrice dati segua l'ordine delle domande del questionario, è necessario tenere presente che alcune domande possono produrre più di una variabile. È il caso, ad esempio, delle domande a scelta multipla che prevedono più di una risposta o delle domande in batteria (cfr. Cap. 4).

Domanda 10. Da quale fonte/sorgente ascolti musica prevalentemente? *(Dopo averle lette tutte, indicare al massimo tre alternative di risposta)*		
1.	□	Radio analogica (FM)
2.	□	Radio digitale (DAB, DAB+)
3.	□	CD
4.	□	Super Audio CD, DVD o Blu Ray Audio
5.	□	Vinile
6.	□	Supporto magnetico (audio cassette, Super 8)
7.	□	Digitale non hi-res (MP3, AAC, OGG ecc.)
8.	□	Digitale hi-res (FLAC, MQA ecc.)
9.	□	Web radio non hi-res
10.	□	Web radio hi-res
11.	□	Altro _____
12.	□	Non ascolto musica

Figura 11.2 Esempio di domanda a scelta multipla con più di un'alternativa di risposta

L'esempio di domanda riportato in Figura 11. prevede che il rispondente, lette le alternative di risposta, ne selezioni da un minimo di una a un massimo di tre. Questa singola domanda occuperà in matrice quattro colonne: le prime tre (che è possibile denominare secondo quanto detto sopra *d10_1*, *d10_2* e *d10_3*) conterranno le informazioni codificate (secondo il codice posto a sinistra delle modalità della domanda) delle scelte effettuate dal rispondente; la quarta variabile (*d10_a*) sarà una variabile testuale che conterrà informazioni soltanto qualora il rispondente abbia selezionato la modalità 11 ("altro") della domanda. In caso di domande in batteria, a ciascun item sarà dedicata una colonna della matrice dati. Se, ad esempio, la domanda 21 (*d21*) è composta da 10 item, questa occuperà dieci colonne della matrice dati che è possibile nominare con codici alfanumerici da *d21_01* a *d21_10*. Con questa nomenclatura i primi tre codici indicano il numero della domanda, gli ultimi due (dopo un trattino basso utilizzato come separatore al posto dello spazio) indicato il numero dell'item della batteria.

Al nome della variabile va poi associata un'*etichetta*, secondo elemento da riportare nel codebook. I programmi di analisi dei dati di comune impiego prevedono la possibilità di inserire, per ciascuna variabile un'etichetta estesa. È bene non utilizzare la formulazione della domanda come etichetta ma effettuare una parafrasi della stessa in quanto l'etichetta associata a ciascuna variabile sarà poi restituita negli output tabellari e grafici del programma di analisi. Per lo stesso motivo è opportuno che le etichette non siano eccessivamente estese e, soprattutto, che non modifichino il senso e il significato della domanda, ossia rispecchino in maniera più fedele possibile il concetto/la dimensione sotteso/a alla domanda del questionario. Ad esempio, per la domanda "Da quante persone, escluso te, è composto il tuo nucleo familiare?", un'etichetta possibile è "Numero componenti del nucleo familiare (escluso il rispondente)" oppure "Numero di altri componenti del nucleo familiare" ma non "Numero di componenti del nucleo familiare" o "Dimensioni del nucleo familiare".

Una terza informazione di cui tenere conto nel codebook è il *tipo di variabile*. I principali programmi di analisi dei dati riconoscono (almeno[25]) due tipi di variabili: numeriche e testuali (note anche come *stringhe* o *stringhe di testo*). Con la locuzione "tipo di variabile" non si fa qui riferimento alla distinzione tra variabile nominali, ordinali o cardinali (di cui si è già discusso nel Cap. 3), ma al tipo di informazione che viene registrata in matrice. Nell'esempio riportato in precedenza (Figura 11.) si fa riferimento a un questionario con codifica incorporata di tipo numerico[26]. In questo caso, dunque, le informazioni contenute nelle prime tre variabili della matrice di dati (le scelte indicate dal rispondente) sono di tipo numerico. La variabile *d10_a* (riferita alla stringa di testo della modalità altro) sarà invece di tipo testuale.

In tabella 11.3 si propone l'esempio dell'organizzazione in matrice delle informazioni registrate con la domanda 10 del questionario (Q1-Q5). Piero (Q1) ascolta musica prevalentemente via Radio analogica, Digitale hi-res e Web radio non hires; Mario invece esclusivamente tramite Radio analogica (modalità 1 della domanda 10 del questionario).

---

[25] R, ad esempio, distingue tra sei classi di oggetti (valori) basilari (atomici): reale (double), intero, carattere, logico, complesso e raw. Altre numerose classi possono essere importate con i diversi pacchetti applicativi o impostate dall'utente. SPSS consente di scegliere tra un insieme finito di nove tipi di variabile: numerica, virgola, punto, in notazione scientifica, data, dollaro, valuta personalizzata, stringa, valore numerico ristretto (intero preceduto da zeri). STATA consente l'utilizzo di variabili di tipo testuale (character) e numerico distinguendo in questo caso tra numeri interi (distinti in byte, int, long in base alle dimensioni – bits – necessarie alla loro archiviazione) e numeri reali (double e float dove il secondo ha una dimensione doppia – 64 bits – rispetto al primo).

[26] Alternativamente si sarebbe potuto optare per una codifica testuale associando a ciascuna modalità di risposta delle lettere (nell'esempio dalla *a* alla *l*, dodicesima lettera dell'alfabeto inglese). Il codebook può contenere informazioni anche relative al *tipo di codifica* utilizzata (numerica o non numerica) per ciascuna variabile della matrice dati.

Tabella 11.3 **Esempio di organizzazione in matrice della Domanda 10 in Figura 11.**

	Nome	**Sesso**	**Età**	*d10_1*	*d10_2*	*d10_3*	*d10_a*	**…**	$\mathbf{V_j}$
Q1	Piero	M	22	1	8	9		…	s_{1j}
Q2	Lina	F	41	3	8	10		…	s_{2j}
Q3	Mario	M	35	1				…	s_{3j}
Q4	Agata	F	62	3	6			…	s_{4j}
Q5	Andrea	M	18	7	9	11	TV Radio	…	S_{5j}
…	…	…	…	…	…	…		…	…
$\mathbf{Q_i}$	s_{i1}	s_{i2}	s_{i3}	s_{id10_1}	s_{id10_2}	s_{id10_3}		…	s_{ij}

Una forma alternativa per raccogliere un'informazione non sulla prevalenza di utilizzo ma sull'utilizzo, in un arco temporale di riferimento limitato (si veda l'istruzione alla compilazione), di tutte le fonti di ascolto può essere una domanda in batteria.

Domanda 10. Quale fonte/sorgente utilizzi per ascoltare musica? *(Facendo riferimento all'ultimo mese, indicare una risposta per ciascuna riga)*				
			La utilizzo	Non la utilizzo
1.	□	Radio analogica (FM)		
2.	□	Radio digitale (DAB, DAB+)		
3.	□	CD		
4.	□	Super Audio CD, DVD o Blu Ray Audio		
5.	□	Vinile		
6.	□	Supporto magnetico (audio cassette, Super 8)		
7.	□	Digitale non hi-res (MP3, AAC, ecc..)		
8.	□	Digitale hi-res (FLAC, MQA ecc.)		
9.	□	Web radio non hi-res		
10.	□	Web radio hi-res		
11.	□	Altro _____		
12.	□	Non ascolto musica		

Figura 11.3 Esempio di domanda in batteria

In questo caso la domanda produrrà 12+1 variabili (colonne) nella matrice dati (da *d10_1* a *d10_12* cui si aggiunge da variabile stringa d10_11a contente

l'informazione qualora il rispondente abbia selezionato "La utilizzo" nella modalità 11 – Altro della domanda).

Anche in questo caso le variabili saranno di tipo numerico ma lo stato del rispondente su ciascuna modalità della domanda potrà assumere soltanto due valori (0 oppure 1) indicanti l'utilizzo o il mancato utilizzo di ciascuna delle fonti/sorgenti musicali. Le 12 variabili saranno dunque delle variabili dummy (*cfr.* nota 14) e la matrice, con riferimento a questa variabile (*cfr*. tabella 11.4), assumerà la forma di una matrice disgiuntiva[27] (per approfondimenti *cfr*. Cap. 17).

Tabella 11.4 Esempio di organizzazione in matrice della Domanda 10 in tab. 11.3

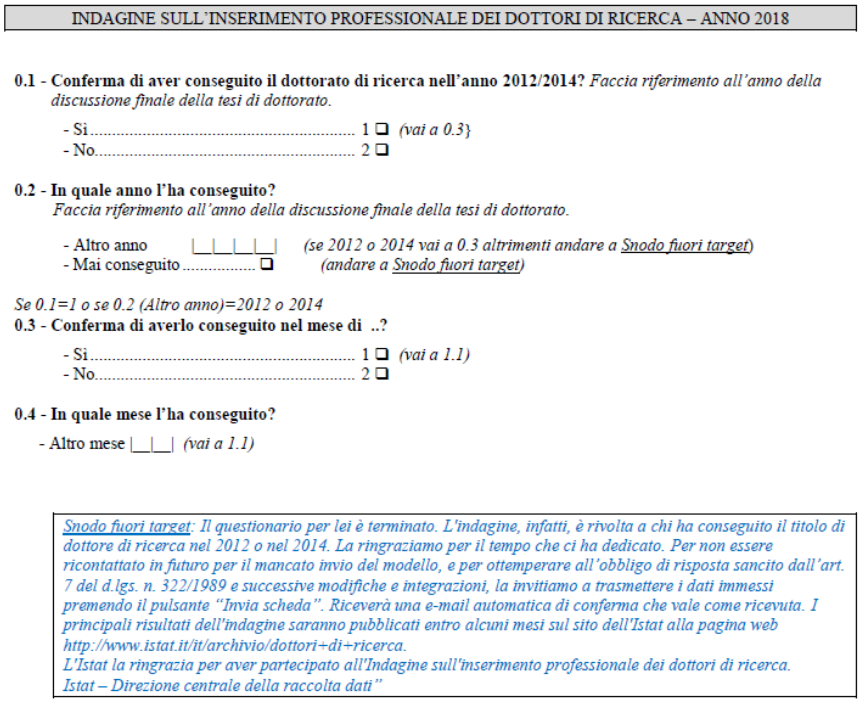
	d10_1	*d10_2*	*d10_3*	*d10_4*	*d10_5*	*d10_6*	*d10_7*	*d10_8*	*d10_9*	*d10_10*	*d10_11*	*d10_12*	V_j
Q1	1	0	0	0	0	0	0	1	1	0	0	0	S_{1j}
Q2	0	0	1	0	0	0	0	1	0	1	0	0	S_{2j}
Q3	1	0	0	0	0	0	0	0	0	0	0	0	S_{3j}
Q4	0	0	1	0	0	1	0	0	0	0	0	0	S_{4j}
Q5	0	0	0	0	0	0	1	0	1	0	1	0	S_{5j}
…	…	…	…	…	…	…	…	…	…	…	…	…	…
C_i	S_{id10_1}	S_{id10_2}	S_{id10_3}	S_{id10_4}	S_{id10_5}	S_{id10_6}	S_{id10_7}	S_{id10_8}	S_{id10_9}	S_{id10_10}	S_{id10_11}	S_{id10_12}	S_{ij}

Ancora un'informazione contenuta nel codebook è relativa alla *scala di riferimento* della variabile. L'esempio di domanda appena riportato ha dato luogo a variabili nominali (siano esse con codici numerici o testuali). Per ciascuna variabile è necessario conoscere a quale scala faccia riferimento, se cioè è nominale, ordinale o categoriale. Per ciascun tipo di variabili, come si vedrà dal prossimo capitolo del volume, sono infatti possibili tipi diversi di operazioni.

Infine, è necessario che per ciascuna variabile vengano stabiliti i *codici mancanti* (*missing value*), ossia i codici attribuiti per ciascuna variabile in caso in cui il

[27] Si noti che anche al momento dell'inserimento dei dati in matrice si può decidere di inserire le informazioni raccolte tramite la domanda 10 proposta in Figura 11. in forma disgiuntiva. In questo modo ciascuna modalità della domanda sarà tradotta in matrice con una variabile, cui si aggiunge la variabile stringa. Questa decisione solitamente è presa dal ricercatore in base alle intenzioni di utilizzo futuro delle informazioni raccolte.

rispondente non abbia fornito informazioni (o abbia fornito informazioni palesemente incongrue in un questionario autosomministrato). Generalmente si di distingue tra *missing dovuto* e *missing non dovuto*. Il primo caso si presenta qualora nel questionario siano presenti delle domande filtro (*cfr*. Capp. 4 e 6) che escludono a una o più sezioni della popolazione di riferimento alcune domande o intere parti del questionario. Ad esempio, nell'indagine sull'inserimento professionale dei dottori di ricerca l'ISTAT richiede all'intervistato, prima di iniziare il questionario vero e proprio, alcune informazioni di conferma sul conseguimento del titolo[28]. Qualora il rispondente confermi, alla domanda 0.1 (*cfr*. Figura 11.), di aver conseguito il titolo nell'anno 2012 o 2014, saltando la domanda 0.2, l'intervistato passa alla domanda 0.3 dove si chiede conferma del mese di conseguimento del titolo.



INDAGINE SULL'INSERIMENTO PROFESSIONALE DEI DOTTORI DI RICERCA – ANNO 2018

0.1 - Conferma di aver conseguito il dottorato di ricerca nell'anno 2012/2014? *Faccia riferimento all'anno della discussione finale della tesi di dottorato.*

- Si ..... 1 ❑ *(vai a 0.3)*
- No ..... 2 ❑

0.2 - In quale anno l'ha conseguito?
Faccia riferimento all'anno della discussione finale della tesi di dottorato.

- Altro anno |__|__|__|__| *(se 2012 o 2014 vai a 0.3 altrimenti andare a Snodo fuori target)*
- Mai conseguito ..... ❑ *(andare a Snodo fuori target)*

Se 0.1=1 o se 0.2 (Altro anno)=2012 o 2014
0.3 - Conferma di averlo conseguito nel mese di ..?

- Si ..... 1 ❑ *(vai a 1.1)*
- No ..... 2 ❑

0.4 - In quale mese l'ha conseguito?

- Altro mese |__|__| *(vai a 1.1)*

Snodo fuori target: Il questionario per lei è terminato. L'indagine, infatti, è rivolta a chi ha conseguito il titolo di dottore di ricerca nel 2012 o nel 2014. La ringraziamo per il tempo che ci ha dedicato. Per non essere ricontattato in futuro per il mancato invio del modello, e per ottemperare all'obbligo di risposta sancito dall'art. 7 del d.lgs. n. 322/1989 e successive modifiche e integrazioni, la invitiamo a trasmettere i dati immessi premendo il pulsante "Invia scheda". Riceverà una e-mail automatica di conferma che vale come ricevuta. I principali risultati dell'indagine saranno pubblicati entro alcuni mesi sul sito dell'Istat alla pagina web http://www.istat.it/it/archivio/dottori+di+ricerca.
L'Istat la ringrazia per aver partecipato all'Indagine sull'inserimento professionale dei dottori di ricerca.
Istat – Direzione centrale della raccolta dati"

Figura 11.4 ISTAT, parte introduttiva del questionario Indagine sull'inserimento professionale dei dottori di ricerca – Anno 2018

In questo caso, il rispondente che risponde sì alla domanda 0.1 confermando l'anno di conseguimento, passando alla domanda 0.3 avrà un valore mancante

[28] L'ISTAT ottiene queste informazioni incrociando diverse banche dati locali (atenei) e nazionali (CINECA). Maggiori info su: https://www.istat.it/it/archivio/224302.

alla domanda 0.2. Questo mancante però è un missing dovuto in quanto l'intervistato non era tenuto a rispondere a questa domanda.
Allo stesso modo, nell'indagine statistica multiscopo sulle famiglie, l'ISTAT chiede informazioni circa il possesso di un'automobile (domanda 8.4) chiedendo poi, ai possessori, se si dispone di un luogo di proprietà per il ricovero della stessa (*cfr*. figura 11.5).

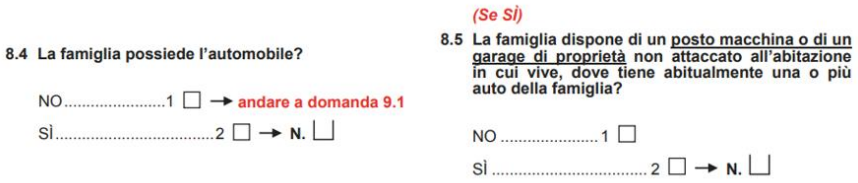


Figura 11.5 ISTAT, Indagine statistica multiscopo sulle famiglie, aspetti della vita quotidiana, Questionario Famiglie – Anno 2021

Nel caso in cui la famiglia rispondente non disponga di un'automobile, la compilazione del questionario prosegue saltando la domanda 8.5 e passando a quella successiva 9.1 (sulla quantità di libri posseduti in famiglia[29]). Evidentemente coloro che non dispongono di un'automobile non devono essere conteggiati fra i rispondenti alla domanda 8.5. In altre parole, la popolazione di riferimento della specifica domanda del questionario 8.5 saranno soltanto le famiglie che hanno risposto "sì" alla domanda (filtro) 8.4.

Questione differente è invece se l'intervistato si è rifiutato di rispondere alla domanda 8.4. Per questa domanda è infatti prevista una risposta (non essendo preceduta da nessun filtro). In caso di mancata risposta a questo tipo di domande si parla di *missing (o dato mancante) non dovuto*. Le cause del missing non dovuto possono essere le più varie: dalla semplice distrazione o fretta nella compilazione (nel caso di un questionario autosomministrato) al rifiuto alla risposta. Un'elevata quantità di mancanti non dovuti su una o più domande del questionario costituisce un *alert* importante per il ricercatore. Tra le cause principali vi è la cattiva progettazione della domanda (mancano ad esempio alcune alternative di risposta e/o non è stata prevista la modalità residuale altro), la percezione della domanda come troppo personale (può essere il caso, ad esempio, della richiesta diretta e puntuale del reddito del rispondente[30]), la collocazione della domanda all'interno del questionario (*cfr*. Capp. 4 e 6) o, nel caso di un questionario somministrato, l'influenza dell'intervistatore. Questi problemi possono essere affrontati in diversi modi. Se individuati nel durante la fase di pretesting degli

[29] Per maggiori informazioni: https://www.istat.it/it/archivio/91926.

[30] È prassi comune, proprio per evitare l'elevato numero di non risposte, ricorrere ad indicatori indiretti di reddito (*cfr*. Cap. 14).

strumenti (Mauceri, 2003) o durante le prime rilevazioni si può pensare di intervenire sugli strumenti di rilevazione[31]. Se invece vengono riscontrati solo al termine della rilevazione, portano necessariamente all'eliminazione della domanda da tutte le analisi per le quali era stata progettata.

Attribuire un codice ai valori mancanti, siano essi dovuti o non dovuti, ha come obiettivo quello di effettuare le operazioni di analisi esclusivamente sui *casi validi*[32]. Un numero eccessivo di mancate risposte dovute può portare il ricercatore alla decisione di escludere l'intero caso dalla popolazione raggiunta.
I programmi di analisi statistica di comune impiego prevedono soluzioni specifiche per l'impostazione dei casi mancanti. SPSS, ad esempio, consente all'utente di definire valori discreti (fino a tre) o consecutivi (un intervallo) per identificare i dati mancanti[33].

Il codebook può poi contenere altre informazioni ausiliarie. Ne è un esempio la schermata *visualizza variabili*[34] del software SPSS che consente, oltre le informazioni sulle variabili elencate in precedenza, di inserire altre informazioni quali la lunghezza (ossia il numero di campi della variabile), il numero di decimali (in caso di variabile numerica), il ruolo[35]. Altre due informazioni (l'allineamento e le

---

[31] Per tale motivo avendo progettato la matrice dei dati parallelamente alla progettazione del questionario, è buona pratica procedere all'inserimento dei dati in matrice in maniera immediatamente successiva alla rilevazione in modo tale da effettuare i primi controlli già durante la stessa fase di somministrazione. Ovviamente nel questionario online (CAWI o CAMI) o comunque nelle forme computer assisted (CATI o CAPI) non essendo presente la fase di inserimento dati, questa attività può essere condotta in maniera più agevole.

[32] Si considera valida la risposta di un caso se l'intervistato ha risposto alla domanda utilizzando in maniera opportuna lo strumento di rilevazione. Non è valida, ad esempio, una risposta fornita aggiungendo manualmente una modalità di risposta a quelle già previste nel questionario, oppure selezionando più opzioni dove era espressamente richiesto di indicare una sola risposta, o ancora (è il caso delle domande in batteria) se l'intervistato segue "schemi" preconfigurati di risposta (selezionando ad esempio sempre l'ultimo, il primo o il gradiente centrale di una scala Likert, *cfr.* Cap. 7). Un altro caso frequente di risposta non valida è quando al momento dell'inserimento dei dati in matrice troviamo nel questionario un dato implausibile. Ad esempio, se al rispondente viene richiesto con una domanda aperta di indicare il titolo/livello di istruzione e troviamo valori irrealistici (ad esempio: "scuola della vita" oppure "la strada") è opportuno inserire in matrice dati il codice utilizzato per i missing non dovuti.

Nei casi di indagini online, un utile strumento per valutare la bontà di un questionario è il tempo medio dedicato alla compilazione: questionari compilati con tempi di molto inferiori al tempo medio di compilazione indicano una possibile mancata attenzione nella compilazione; tempi molto superiori possono indicare una difficoltà di comprensione dello strumento o una compilazione a intermittenza.

[33] SPSS consente anche di inserire un intervallo di valori consecutivi e un valore discreto.

[34] L'ambiente di lavoro di SPSS è distinto in due sotto ambienti: il primo (*visualizza dati*), similmente ad un foglio di lavoro Excel, è il luogo dove sono inseriti i dati secondo lo schema Casi x Variabili; il secondo (*visualizzazione variabili*) è un editor di variabili in formato matriciale dove in riga sono elencate tutte le variabili della matrice dei dati e in colonna sono specificate determinate loro caratteristiche (Di Franco 2010). Le informazioni contenute nel visualizzatore delle variabili costituiscono, in sostanza, il Codebook.

[35] Il campo ruolo, recentemente inserito nelle ultime versioni del programma, consente di semplificare il richiamo delle variabili in fase di elaborazione dati. Il campo consente di distinguere le variabili in base all'utilizzo che se ne intende fare. Si può scegliere tra: input (se si tratta di una variabile

colonne) sono invece riferite soltanto alla visualizzazione delle stesse nella schermata visualizzazione dati del programma.

2.2 L'inserimento dei dati in matrice

Se, come si diceva, la costruzione della matrice è avvenuta contestualmente alla costruzione del questionario (o comunque in una fase immediatamente prossima), il passo successivo alla raccolta delle informazioni consiste nell'inserimento delle stesse nella matrice dei dati strutturata. L'inserimento dei dati raccolti nella matrice è una fase estremamente delicata in cui possono insorgere errori che, se non accuratamente individuati durante la fase di pulizia (si veda oltre), si trascineranno sino all'analisi dei dati per invalidare, nei casi più gravi, i risultati della ricerca. Tra tutte le fasi descritte, l'inserimento dei dati è una di quelle maggiormente *time-consuming*. È bene ricordare che questa fase è presente prevalentemente qualora si stia conducendo una ricerca che preveda l'utilizzo di dati primari (sulla distinzione tra dato primario e secondario *cfr.* Cap. 18) e si sia fatto ricorso a una raccolta delle informazioni cartacea[36]. Le risposte alle domande registrate su supporto cartaceo devono poi essere digitalizzate, trasformate cioè in valori, codici e testo al momento dell'inserimento nella matrice dati. Anche per questo motivo oggi si fa sempre più ricorso a indagini online e forme di raccolta dei dati computer assistita (CAPI, CAWI, CATI, CAMI). In tutti questi casi, infatti, la fase di inserimento dei dati viene condotta contestualmente alla compilazione del questionario (nel caso di un questionario online) o alla registrazione delle domande da parte dell'intervistatore (nel caso di raccolta computer assistita)[37].
La fase di inserimento dati in matrice inizia con la numerazione del questionario. È buona norma assegnare un codice identificativo al questionario, da riportare

utilizzata come indipendente o un predittore in modelli di regressione), obiettivo (variabile dipendente), entrambi, nessuno, partizione (se la variabile è utile a suddividere la popolazione raggiunta come nel caso di test a campioni separati), suddivisione (IBM 2021).

[36] Un atro caso in cui è necessario un inserimento manuale dei dati in matrici è, come detto in precedenza, il caso in cui si stia svolgendo un'analisi del contenuto come inchiesta.

[37] Queste ultime due forme di somministrazione degli strumenti di rilevazione consentono anche di ridurre considerevolmente i possibili errori in matrice (*cfr*. Par. successivo), soprattutto se si è investito in un adeguato addestramento dei rilevatori (*cfr*. Cap. 4). Tuttavia, assieme alle forme computer assisted, anche nel caso del questionario online sarà necessario un investimento di tempo maggiore per le ricodifiche (*cfr*. Par. successivo). Gli strumenti di rilevazione digitali offrono, solitamente, poca possibilità di gestione (*ex ante*) delle modalità con le quali l'informazione viene digitalizzata e dunque di personalizzazione della matrice dei dati. Riprendendo un esempio effettuato in precedenza con riferimento alla Figura 11., è lo strumento utilizzato che definisce se l'informazione viene registrata in maniera disgiuntiva o meno. Ovviamente se ciò non converge con le proprie necessità di analisi, sarà possibile intervenire durante la fase di ricodifica dei dati.

poi in matrice dati, in modo da poterlo recuperare in maniera agevole qualora, nella fase di controllo prima o nella fase di analisi, poi, si riscontrino valori anomali (si veda oltre), dati mancanti (non dovuti), incompletezze o errori nella compilazione, ad esempio, delle stringhe di testo. Nel caso il lavoro di inserimento dati sia svolto da più persone è anche utile dedicare un campo inziale della matrice al nome (o codice identificativo) di chi inserisce effettivamente il dato in matrice, sempre per le ragioni di ritorno sul cartaceo di cui sopra.

Si è detto in precedenza che è importante costruire la matrice dati con le variabili che seguono l'ordine delle domande del questionario. Ciò anche per agevolare il lavoro di inserimento dei dati. Poiché spesso alle modalità di risposta da inserire nel questionario corrispondono dei codici numerici (si veda come esempio il questionario con codifica incorporata in Tavola 11.2) è bene dotarsi di un computer con tastierino numerico. Utilizzare la parte superiore della tastiera, quelli in sequenza da 1 a 0 sopra le lettere per intendersi, solitamente ingenera un numero maggiore di errori[38].

L'inserimento dei dati in matrice è un lavoro ripetitivo. Come tutti i lavori ripetitivi, con la pratica aumenta la velocità di esecuzione. Se l'inserimento del primo questionario porterà via diverso tempo, una volta acquisita manualità e memorizzati i codici, le operazioni saranno via via più veloci. Ma è proprio questa velocità che porta a una falsa sicurezza e al rischio di facili distrazioni che possono comportare errori nell'inserimento.

I principali tipi di errore sono quattro. L'*errore di trascrizione* tra tutti è il più comune. Include errori di digitazione quali le ripetizioni, le eliminazioni o altri errori tipografici. L'*errore di trasposizione*, più comune nelle stringhe di testo, è determinato dallo scambio di posizione di un carattere con un altro che, nei casi più gravi, può alterare completamente il significato di una parola: si trascrive in matrice, ad esempio, *brodo* invece della parola *bordo* riportata nel questionario. Con i valori numerici, l'errore di trasposizione consiste nell'invertire i numeri che compongono un codice di una modalità (inserire ad esempio il codice *21* al posto del codice *12*) o di un campo numerico aperto[39]. L'*errore di formattazione* o di incongruenza tra le unità rilevate e la rappresentazione in matrice avviene quando si inserisce in matrice un valore che non corrisponde al tipo e/o alla scala di riferimento della variabile definita nel codebook. Se ad esempio si inserisce un valore numerico dove è previsto un valore testuale o viceversa. Questo tipo di errore è molto comune nel caso di variabili che necessitano di formati particolari quali, ad esempio, le date (GG/MM/AAAA) o le coordinate geografiche (che hanno specifici sistemi di riferimento) ma anche indirizzi o codici indentificativi (come il codice fiscale). L'*errore di salto*, dovuto al salto involontario di una o più colonne

[38] Una discreta quota di errori nell'inserimento dei dati in matrice è dovuta a quello che è definito *fat-finger error*, ossia ad un errore di involontaria digitazione di un tasto sbagliato sulla tastiera.

[39] Questo è uno degli errori più difficili da individuare durante la fase di pulizia della matrice, soprattutto quando la nuova stringa o il nuovo codice numerico assumono un significato all'interno della frase, nel primo caso, o tra gli altri codici numerici, nel secondo.

della matrice. Questo errore, dettato dagli automatismi che nascono nell'inserimento dei questionari, comporta l'inserimento di un dato in una colonna successiva della matrice. Non ci sarà corrispondenza, dunque, tra valore rilevato e campo in matrice.

2.3 La pulizia dei dati

Il *data cleaning* è il processo di preparazione dei dati per la successiva fase di analisi. L'obiettivo è correggere (o, in alcuni casi, eliminare) qualsiasi elemento distorsivo, dovuto agli errori visti nel paragrafo precedente, che inficerebbe i risultati della ricerca.

Una prima procedura da seguire è quella dei *controlli a vista*. Questa procedura, opzionale perché approssimativa, consente comunque in taluni casi di individuare errori di trascrizione, trasposizione o formattazione ma ha una portata limitata in quanto non è possibile visualizzare analiticamente e contemporaneamente ciascuna delle variabili della matrice. Questi controlli, come principale finalità, hanno quella di individuare le celle vuote in una matrice dati. Se abbiamo impostato per ciascuna variabile i codici per i valori mancanti (dovuti e non dovuti) nessuna cella della matrice dovrebbe, infatti, risultare vuota.
I controlli da effettuare a seguito dell'inserimento dei dati in matrice si possono distinguere in controlli di plausibilità e controlli di congruenza.

I *controlli di plausibilità* hanno la finalità di individuare i *wild code* (letteralmente, codici selvaggi), ossia verificare che in ciascuna colonna della matrice (dunque per ciascuna variabile) siano inserite soltanto informazioni pertinenti e nel formato definito nel codebook. Oltre al controllo a vista, la strategia da seguire per effettuare questi indispensabili controlli è effettuare un'analisi monovariata (*cfr*. Cap. 13) di ciascuna delle variabili che compongono la matrice dati. Un'attenta lettura dei risultati consentirà di individuare i valori anomali. Ci si riferisce qui a valori *out of range* per le variabili cardinali o quasi cardinali (ad esempio un'età pari a 200 anni o una popolazione eccessivamente numerosa per una determinata area geografica); a valori anomali per le variabili categoriali, ossia valori per i quali non è stato predisposto un adeguato codice in matrice e dunque nel codebook (riprendendo l'esempio della domanda riportata Figura 11. che prevede 12 modalità di risposta, un valore anomalo è il codice 15 o 22, ad esempio); nel caso di variabili testuali una attenta lettura dei risultati di un'analisi monovariata consente di individuare errori di battitura, elisioni, ripetizioni ecc. In tutti questi casi, individuato un errore, è necessario riprendere il questionario cartaceo (individuabile tramite il codice questionario riportato in matrice di cui si è detto sopra) e sostituire il *wild code* con il codice (o la porzione di testo) corretto. Da sottolineare infine che questo controllo tramite distribuzioni monovariate non consente di individuare tutti gli errori (materiali) di cui si è discusso nel

paragrafo precedente. Si pensi ad esempio al caso in cui, invece di inserire un codice (ad esempio il codice 2 riferendosi alla domanda in Figura 11.) si è inserito un altro codice previsto come modalità di risposta della domanda (nello stesso esempio, il codice 5). La maggior parte di questi errori può essere evitata soltanto garantendo la dovuta attenzione alla fase di inserimento dati in matrice (*cfr*. paragrafo 2.3).

Una volta terminata questa attività di pulizia delle informazioni registrate in matrice tramite analisi monovariata è necessario passare ai *controlli di congruenza*. Attraverso questa attività si vuole fare emergere eventuali incongruenze tra coppie di variabili, ossia tenendo conto congiuntamente della distribuzione di frequenza di due variabili attraverso una tabella di contingenza (come nell'analisi bivariata *cfr*. Cap. 14). Incrociando, ad esempio, l'informazione sull'età con quella sul possesso o meno della patente di guida, è incongruente che un dodicenne disponga di una qualsiasi licenza. Oppure, mettendo in relazione la variabile genere con la variabile professione non potrà logicamente darsi il caso di un sacerdote cattolico donna (almeno sino ad oggi). Anche in questo caso sarà necessario ritornare sul questionario cartaceo e individuare l'errore in matrice per correggerlo o, nei casi estremi, eliminare l'informazione non congruente (impostando il valore meno plausibile, o entrambi, con il codice utilizzato per il missing non dovuto). I controlli di congruenza sono estremamente utili anche per verificare la corretta impostazione dei *missing value*. Se il questionario prevede una domanda filtro che esclude una domanda successiva, i controlli di congruenza consentono di individuare i casi che non avrebbero dovuto rispondere ma hanno ugualmente risposto (valori che andranno corretti in matrice con il codice dedicato ai mancanti non dovuti) e i casi che avrebbero dovuto rispondere ma non hanno risposto (mancante dovuto).

Ad esempio, ponendo che alla domanda 0.1 in Figura 11. abbiano risposto positivamente 200 persone su un totale di 300 intervistati, la variabile relativa alla domanda 0.2 dovrà contenere soltanto 100 celle con valori differenti da quelli impostati per il missing. Allo stesso modo per la domanda successiva, la 0.3, dovranno essere presenti in matrice informazioni per i 200 casi che hanno risposto sì alla domanda 0.1 più i casi che hanno risposto 2012 o 2014 alla domanda 0.2 (*cfr*. Figura 11.)

Molto spesso, per carenza di risorse temporali o economiche o per la volontà di passare immediatamente alla vera e propria analisi dei dati, non si riserva l'adeguata attenzione alla fase di pulizia della matrice. Anche in questa eventualità – assolutamente non incoraggiata né consigliabile – si dovrebbe, tuttavia, aver cura di correggere in matrice le incongruenze che emergono durante la fase di analisi dei dati (Marradi 1993).

2.4 I metadati

Secondo la definizione più utilizzata in ambito internazionale[40], i metadati sono «dati che definiscono e descrivono altri dati e processi». L'etimologia del termine stesso esplicita chiaramente il suo significato. Il prefisso *meta-* (derivato dal termine greco μετὰ) indica appunto ciò che *va oltre* il singolo dato (dal latino *datum*). Più specificatamente i metadati statistici sono «dati su dati e altra documentazione che descrivono oggetti in modo formalizzato»[41].
In maniera più estesa possiamo dunque definire come metadati tutte quelle risorse, siano esse numeriche e/o testuali, che esplicitano, in maniera più o meno dettagliata, informazioni ritenute rilevanti per la *progettazione* degli strumenti di indagine, la *raccolta* e l'*elaborazione* dei dati, la *diffusione* degli stessi e dei risultati di ricerca.

Pensando ad esempio alla *survey research* (*cfr*. Cap. 4) possono essere considerati metadati tutte le informazioni utilizzate per la *progettazione* (del questionario strutturato cfr. Cap. 6) come, ad esempio, la letteratura sull'argomento oggetto di indagine, i risultati derivati da indagini condotte in precedenza su tematiche affini, le informazioni circa la popolazione di riferimento su cui poi, eventualmente, operare il campionamento (*cfr.* Cap. 5). Possono essere ancora considerati metadati informazioni circa le scelte effettuate dal/i ricercatore/i come, ad esempio, la scelta dell'utilizzo di una scala a sette gradienti invece di una a cinque (*cfr.* Cap. 7). È sempre opportuno tenere traccia di queste informazioni di selezione/scelta. Molte volte queste scelte vengono argomentate nei report di ricerca, altre vengono sottaciute[42].

Durante la fase di *raccolta* delle informazioni, nel caso ad esempio avessimo optato per una rilevazione tramite somministrazione guidata di questionario, sono considerati metadati le informazioni utili all'intervistatore (chi somministra il questionario) al fine di ottenere risposte quanto più attendibili/affidabili/fedeli. Sono metadati anche le informazioni a margine che l'intervistatore annota di sua sponte o in spazi appositamente dedicati. Possono essere informazioni di questo tipo la percezione di dinamiche intercorse durante l'intervista (fretta dell'intervistato, difficoltà di comprensione della lingua, necessità di riproposizione di alcune domande) o informazioni contestuali (orario dell'intervista, luogo ecc.). In caso di indagini condotte online sono metadati le informazioni circa l'orario in cui si è risposto al questionario, l'indirizzo IP del rispondente (da cui è

[40] Dall' inglese «data that defines and describes other data and processes» (SDMX 2009, p. 84)

[41] Dall' inglese «data and other documentation that describes objects in a formalized way» (UNECE 2000, p. 20).

[42] In questo secondo caso il rischio è di perdere informazioni utili alla ripercorribilità e alla pubblicità (dunque alla trasparenza) delle scelte effettuate in fase di progettazione. Come già sottolineato (*cfr*. Cap. 6) la fase di progettazione degli strumenti di indagine è una fase estremamente importante durante la quale vengono effettuate delle scelte irreversibili dalle quali dipendono i risultati dell'indagine.

desumibile, con una certa approssimazione, il luogo), il tempo impiegato per concludere il questionario o, a volte (se il software utilizzato lo consente), per rispondere alle singole domande, la piattaforma attraverso la quale il rispondente si è collegato (nel caso, ad esempio, di un campionamento a valanga con link pubblicati su diversi siti online). Nel caso di un disegno di indagine quasi-sperimentale (*cfr*. Cap. 9), rivestiranno un ruolo di non secondaria importanza le informazioni di carattere contestuale in cui è stato portato avanti l'intervento sperimentale. Queste informazioni, anch'esse definibili come metadati, spesso fanno riferimento al contesto o al clima in cui è stato svolto l'esperimento[43].

Anche durante la fase di *elaborazione* si producono metadati che sarebbe opportuno conservare e accludere ai risultati della propria ricerca. Ad esempio, sono metadati le informazioni circa le scelte operate dal ricercatore/analista dati per effettuare le ricodifiche di cui si è discusso sopra. Anche l'esplicitazione delle procedure adottate per la costruzione degli indici (*cfr*. Cap. 14) sono da considerare metadati in quanto esplicitano informazioni di *data handling* (manipolazione dei dati) che altrimenti rimarrebbero tacite e dunque non riproducibili mentre la loro pubblicazione rende queste attività pienamente controllabili e riproducibili. Ad esempio, la costruzione dell'indice di capitale culturale familiare dell'unità statistica di riferimento partendo dal titolo di studio del padre e della madre può assumere diversi valori in base alle scelte effettuate in sede di composizione.

Infine, di estrema importanza sono le risorse che accompagnano la *diffusione dei dati* e la *diffusione dei risultati* di ricerca. Sono sempre più numerose le attività di ricerca per le quali si mettono a disposizione di altri (adeguatamente anonimizzati) i dati raccolti. L'Istat, ad esempio, rende disponibili diversi tipi di microdati[44], pubblicamente o su richiesta per diversi scopi (*cfr*. Cap. 18). Di per sé anche diffondere il dato utilizzato è un metadato non solo in quando rende riproducibili i risultati della ricerca ma perché attraverso una matrice di dati già strutturata, se adeguatamente interrogata, può rispondere a obiettivi cognitivi diversi da quelli per i quali è stata progettata. Sempre più numerosi sono poi i portali open data (governativi, sovrannazionali e nazionali, ma anche di istituzioni di ricerca, prevalentemente pubbliche) che mettono a disposizione della comunità scientifica raccolte di informazioni strutturate in forma di matrice. Per quanto riguarda la diffusione dei risultati, gli stessi portali consentono a volte di effettuare elaborazioni semplici o complesse sulle matrici di dati in essi contenute. Sono poi da considerarsi metadati, oltre alle pubblicazioni direttamente derivate dalla ricerca, tutte le attività ad essa collegate quali i comunicati stampa, grafici ed elaborazioni, presentazioni a convegni e seminari ecc.

---

[43] Nel caso di disegni sperimentali o quasi sperimentali è particolarmente rilevante progettare sin dal principio gli strumenti utili a raccogliere, registrare e associare ai soggetti questo tipo di informazioni.

[44] Per informazioni: https://www.istat.it/it/dati-analisi-e-prodotti/microdati.

3. Per riassumere

In conclusione, riprendendo le fasi del processo della ricerca quantitativa (*cfr.* Cap. 2), quanto si è argomentato in questo capitolo si colloca, prevalentemente, lungo l'asse temporale che va dalla fase di progettazione degli strumenti di rilevazione alla fase di analisi dei dati e interpretazione dei risultati di ricerca.
Con il diagramma proposto in Figura 11. si vuole evidenziare come tutte le fasi siano interconnesse tra di loro, ossia la successione delle fasi non è così lineare come potrebbe apparire inizialmente. In ciascuna delle fasi è possibile, se non probabile, individuare elementi migliorativi del proprio percorso che portano alla revisione di alcune decisioni prese o scelte effettuate negli step precedenti. Sono queste continue attività di affinamento, la cui argomentazione dovrebbe essere compresa tra i metadati (processo che segue l'intero processo di ricerca), che rendono particolarmente delicata l'attività di ricerca in campo sociale.

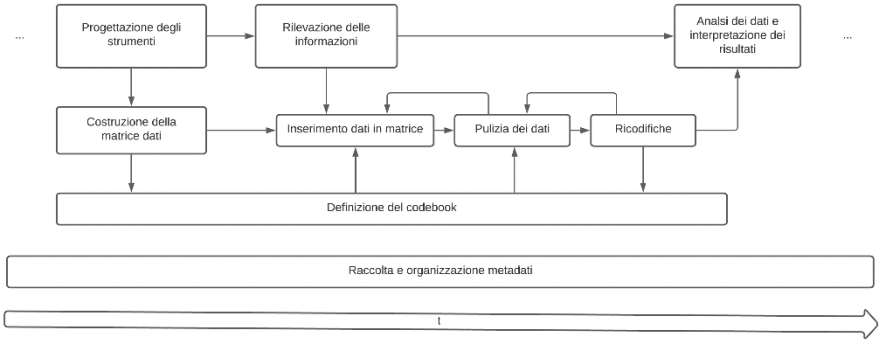


Figura 11.6 Fasi dell'organizzazione dati in matrice

Nel diagramma si evince come la *costruzione della matrice* dati sia una fase da avviare contestualmente alla progettazione degli strumenti di rilevazione e come questa debba essere pronta al momento della partenza dell'attività di rilevazione delle informazioni. È bene infatti prevedere immediatamente la fase di *inserimento dei dati in matrice* al fine non solo di contenere i tempi di questa delicata (e dispendiosa, in termini di tempo ed energie) fase ma anche di verificare la funzionalità degli strumenti stessi (il questionario e la matrice dati). Una volta terminata questa fase – o *anche* in itinere per sottoinsiemi predefiniti di casi (scuole, comuni, ecc. a seconda del piano della rilevazione) – si procederà ai controlli di qualità delle informazioni in matrice, effettuando le diverse operazioni di *pulizia*

che, in alcuni casi, possono comportare il ritorno al materiale cartaceo (errori di inserimento).

Solo quando si otterrà una base di dati pulita, si procederà alle *ricodifiche* delle variabili (*cfr*. Cap. 13), anche in relazione al tipo di analisi o al tipo di output tabellari e grafici da inserire nel rapporto di ricerca. Come la *raccolta e l'organizzazione dei metadati*, la *definizione del codebook* è un'attività che corre parallelamente a tutte le altre fasi. Il codebook costruito inizialmente sarà utilizzato durante la fase di inserimento dei dati in matrice e durante i controlli di congruenza e plausibilità; successivamente verrà arricchito con altre informazioni sulle variabili aggiuntive, realizzate ricodificando le variabili inizialmente presenti in matrice.