istat
working
papers

N. 6
2011

Rappresentazione sintetica di indicatori di qualità per i dati amministrativi

Antonio Bernardi

Istat
Istituto Nazionale
di Statistica

istat
working
papers

N. 6
2011

Rappresentazione sintetica di indicatori di qualità per i dati amministrativi

Antonio Bernardi

Istat

Istat Working Papers

Rapresentazione sintetica di indicatori di qualità per i dati amministrativi

N. 6/2011

ISBN 978-88-458-1686-9

Istituto nazionale di statistica
Servizio Editoria
Via Cesare Balbo, 16 – Roma

Rappresentazione sintetica di indicatori di qualità per i dati amministrativi

Antonio Bernardi

Sommario

La ricerca presenta un'applicazione dell'Analisi delle Corrispondenze di tipo non parametrico con cui si rappresenta sinteticamente un insieme di indicatori qualità calcolati sulle variabili di un archivio amministrativo. La metodica è poi estesa alla rappresentazione di indicatori di qualità calcolati su un insieme di archivi amministrativi integrati.

Parole chiave: Indicatori di qualità, dati amministrativi, analisi delle corrispondenze.

Abstract

The research presents an application of non-parametric Correspondence Analysis giving a synthetic representation about a set of quality indicators calculated on the variables of an administrative archive. Then the method extends to the description of quality indicators calculated on an integrated set of administrative sources.

Keywords: Quality indicators, administrative data, correspondence analysis.

Introduzione

Molti Istituti Nazionali di Statistica di paesi dell'UE impiegano sempre di più informazioni tratte da archivi amministrativi per migliorare la loro attività di produzione dati. Due i motivi che spiegano tale tendenza. Primo, in un'ottica finanziaria, anche gli Istituti Nazionali di Statistica, come ogni altro ente della pubblica amministrazione, rispondono pro quota alla richiesta di diminuzione delle spese pubbliche impiegando i dati amministrativi al posto dei dati rilevati con indagini correnti, riducendo così i costi di rilevazione. Secondo, in un'ottica statistica, poiché i dati amministrativi sono estesi su tutto il tessuto socio-economico di un paese a un livello molto dettagliato, il loro uso può effettivamente consentire un forte ampliamento della base informativa. In altre parole, i dati amministrativi possono svolgere una funzione importante sia verso le indagini in corso, tramite un'attività di sostituzione e/o di controllo, sia verso le indagini in fieri, mediante un ampliamento quantitativo e qualitativo della base conoscitiva che evidentemente produrrà in un secondo momento una migliore comprensione delle dinamiche socio-economiche territoriali.

Una volta che un singolo archivio amministrativo sia stato acquisito presso un Istituto Nazionale di Statistica, esso, in accordo alle raccomandazioni dell'UE più oltre esposte, dovrà essere corredato da una serie di indicatori di qualità relativi alle sue variabili e sorge la necessità di descrivere i primi e le seconde. Ciò può essere fatto in due modi: con report disaggregati per ogni specifica combinazione variabili-indicatori, oppure sintetici e compatti. Nell'ambito dei secondi, una possibile soluzione è fornita dall'analisi delle corrispondenze di tipo non parametrico, e il fine di questa ricerca[*] è la presentazione dei risultati ottenuti con tale metodo, rilevandone gli aspetti positivi e alcune criticità.

Questa l'articolazione del documento: nel primo paragrafo si presentano le principali linee guida indicate dall'UE sul trattamento degli archivi amministrativi; nel secondo paragrafo è esposta la metodologia con cui si propone di descrivere il comportamento delle variabili di un archivio rispetto agli indicatori di qualità; nel terzo paragrafo si applicano tali metodiche all'archivio degli Studi di settore (SDS), quindi facendo riferimento ad un singolo archivio; nel quarto paragrafo la metodiche vengono utilizzate su un complesso di archivi amministrativi utilizzati dal Central Bureau voor de Statistiek, l'Istituto di Statistica Olandese; infine, alcune osservazioni concludono la ricerca.

1. Alcune linee guida dettate dalla UE in materia di archivi amministrativi finalizzati ad uso statistico

L'UE ha finanziato il progetto "BLUE-ETS", acronimo di Best Linear UnbiasedEstimator - Enterprise and Trade Statistics,[1] che mira all'acquisizione di informazioni statistiche robuste e di alta qualità sulle imprese per migliorare la politica economica e la ricerca socio-economica a sostegno della rinnovata strategia di Lisbona.[2] Il progetto è molto importante e l'Istat è il paese coordinatore.[3] Si tratta in particolare di un progetto di R & S finalizzato a modernizzare e riorganizzare i metodi per la produzione di statistiche e la raccolta dei dati, a semplificare e definire le priorità, a ridurre l'onere di risposta e i costi per le imprese che derivano dalla burocrazia, dall'eccesso di regolamentazione e dalle duplicazioni, rendendo loro meno oneroso il fornire maggiori informazioni.

(*) L'autore ringrazia un anonimo Referee per le sue utili osservazioni. Le elaborazioni, concluse nel mese di dicembre 2010, sono state svolte con i software Stata (versione 11.1) e Miner 3D (versione 7.2.14).

[1] Il progetto è finanziato dalla Commissione Europea nel Settimo programma quadro dell'UE - Tema 8 - Scienze socioeconomiche e scienze umane - SSH-CT-2010 -244.767.

[2] Nel trattato di Lisbona, entrato in vigore il 1 dicembre 2009, sono previste "nuove regole che disciplinano la portata e le modalità della futura azione dell'Unione. Il trattato di Lisbona consente pertanto di adeguare le istituzioni europee e i loro metodi di lavoro", UE 2010.

[3] L'Istat è stato uno dei primi INS ad inaugurare la strada dell'uso dei dati amministrativi per fini statistici già dalla seconda metà degli anni '90, con l'Archivio Statistico delle Imprese Attive (ASIA).

Il progetto Blue-Ets è strutturato in vari working project e uno di essi ha il compito di studiare "la possibilità di aumentare l'uso dei dati amministrativi (ad esempio registri) a fini statistici",[4] un fatto questo oggetto anche di inquadramenti teorici in letteratura.[5]

"Poiché la produzione di statistiche di elevata qualità dipende in larga misura dalla qualità dei dati di input, è di vitale importanza che gli Istituti Nazionali di Statistica abbiano una procedura disponibile in grado di determinare la qualità dei dati amministrativi - per uso statistico - in maniera rapida, semplice e in modo standardizzato",[6] per cui "l'obiettivo principale di questo pacchetto di lavoro è quello di sviluppare nuovi strumenti comprensivi di indicatori di qualità, una Quality Report Card per i dati amministrativi che possa essere generalmente applicata a diverse fonti di dati amministrativi in diversi paesi europei. Il lavoro farà progredire la comprensione della misurazione della qualità delle fonti dei dati in dati generali e amministrativi, in particolare, affrontando le sfide concettuali e metodologiche per quanto riguarda la determinazione della qualità in un contesto completamente diverso".[7]

2. Una proposta metodologica per la rappresentazione sintetica della qualità degli archivi amministrativi

L'utilizzo di un archivio amministrativo per fini statistici deve prevedere una validazione statistica dell'informazione amministrativa. In particolare, quando un archivio amministrativo viene acquisito da un Istituto Nazionale di Statistica (INS), con l'obiettivo ultimo di utilizzare l'informazione amministrativa nel proprio patrimonio informativo statistico, deve essere completato con indicatori di qualità per tutte le variabili dell'archivio.

Supponendo che la descrizione delle variabili e degli annessi indicatori di qualità sia formalizzata con una tavola a 2 vie, con gli indicatori di qualità dell'input come colonne e con le variabili dell'archivio come righe, per cui ogni riga conterrà un set di indicatori riferiti ad una specifica variabile, mentre ogni colonna indicherà i punteggi di un particolare indicatore rispetto a tutte le variabili dell'archivio, si avrebbe una rappresentazione dei dati nel seguente modo:

Tavola 1 - Schema dati del processo d'integrazione di un singolo archivio amministrativo

	Archivio singolo integrato			
	Indicatore di qualità 1	Indicatore di qualità 2	…	Indicatore di qualità n
Variabile 1	…	…	…	…
Variabile 2	…	…	…	…
…	…	…	…	…
Variabile m	…	…	…	…

Il precedente schema, inoltre, può essere direttamente esteso anche ad un complesso di archivi amministrativi, da validare prima di utilizzare nella realizzazione e/o nell'aggiornamento di un registro statistico.

[4] Blue-Ets, 2010, Working Project n.4, cui partecipa l'autore della presente ricerca.
[5] Wallgren, A., Wallgren, B., 2007.
[6] Blue-Ets, 2010.
[7] Ibidem.

Tavola 2 - Schema dati del processo d'integrazione di un complesso di archivi amministrativi

	Sistema di archivi integrati			
	Indicatore di qualità 1	Indicatore di qualità 2	…	Indicatore di qualità n
Archivio 1	…	…	…	…
Archivio 2	…	…	…	…
…	…	…	…	…
Archivio m	…	…	…	…

In questo caso la tavola a 2 vie ha lungo le righe gli archivi oggetto d'integrazione e, lungo le colonne, gli indicatori di qualità dell'input che saranno diversi dai precedenti perché ora essi dovranno dare un punteggio rispetto all'archivio nel suo complesso e non rispetto ad una singola variabile.

I dati presenti nelle celle di entrambe le tavole possono essere i più svariati da un punto di vista metrico. Ad esempio, nella tavola 2 l'indicatore i-esimo potrebbe contenere misure continue sul grado di copertura dei vari archivi, oppure potrebbe descrivere un ordinamento dei vari archivi, con misure discrete del numero delle indagini di un INS che essi possono sostituire, oppure potrebbe riportare misure categoriche tipo "poco adeguato", "adeguato", "molto adeguato" confrontando come i vari archivi si siano comportati rispetto alle aspettative che hanno spinto alla loro acquisizione. Analoghe considerazioni possono essere fatte per la tavola n.1, dove ad esempio l'indicatore j-esimo potrebbe indicare il livello di dati mancanti per le variabili dell'archivio.

Le tavole sopra esposte, infine, possono essere complesse, cioè con molte righe e molte colonne. Tabelle semplici potranno essere rappresentate con un numero di strumenti, sovente univariati, eguali al numero delle celle: ad esempio, per ogni cella di tavola 1, rappresentante una combinazione di una variabile amministrativa con un indicatore di qualità, vi potrebbe essere un indice statistico, un grafico esplorativo tipo istogramma con kernel, ecc., ma tale metodo diventa meno percorribile quando vi sono tabelle con decine se non centinaia di celle. In tale situazione si è del parere che sia opportuno iniziare con una lettura della tavola complessiva avvalendosi di opportune tecniche di riduzione e che, solo per le celle male rappresentate dalle citate tecniche, sia necessario avvalersi di analisi specifiche.

2.1 Impostazione dell'analisi delle corrispondenze

L'analisi delle corrispondenze (AC) appare uno strumento idoneo per la riduzione della dimensionalità di tavole di dati come quelle prima esposte. Essa è in grado di rappresentare le variabili di riga, cioè per tavola 1 le singole variabili dell'archivio e per tavola 2 i singoli archivi, e le variabili di colonna, cioè gli indicatori di tavola 1 o tavola 2, come punti relazionali in uno spazio di ridotta dimensionalità. L'AC permette di osservare tramite l'*asymmetric* plot l'associazione incrociata dei punti variabile (tavola 1), o dei punti archivio (tavola 2), verso i rispettivi punti indicatori, nonché tramite il *symmetric* plot le somiglianze interne ai due gruppi di categorie (punti variabile per tavola 1, punti archivio per tavola 2, e punti indicatore), un fatto questo che, per quanto concerne le categorie di colonna, può far emergere la presenza di indicatori in qualche misura correlati e quindi ridondanti. Poiché l'interesse dell'analisi è qui rivolto a ricercare un modo con cui descrivere la relazione degli indicatori di qualità con le variabili, i primi hanno ricevuto il ruolo di punti di riferimento e, quindi, sono stati trattati come punti vertice con coordinate standard mentre le seconde sono state gestite come punti profilo riga con coordinate principali.

Per realizzare l'AC ci sono stati due problemi da superare, entrambi di ordine metrico ma differenti poiché originati da diversi fattori.

Il primo è stato originato dal fatto che i dati impiegati hanno metriche eterogenee e, in ogni caso, non sono valori di *count*, come tipicamente avviene con l'AC. La metrica che è stata applicata per analizzare tali generi di dati è quella del *double ranking*, come indicato in letteratura (Greenacre, 2007), e cioè le variabili sono state prima trasformate in ranghi misurati lungo le colonne, cioè lungo gli indicatori, e poi raddoppiate poiché se in una tavola vi fosse una riga con ranghi tutti eguali, ed un'altra riga anch'essa con ranghi tutti eguali ma diversi dai prece-

denti, tali righe avrebbero il loro profilo indistinguibile. La soluzione consiste nel rimpiazzare i ranghi - calcolati lungo le colonne (qui gli indicatori) - con due misure: la distanza (numero posti) dell'elemento rispetto al rango minimo e la distanza rispetto al rango massimo.[8.]

Il lavorare sui ranghi, inoltre, conferisce all'analisi una maggiore robustezza e, infatti, tale analisi viene definita in letteratura come AC non parametrica.[9]

Il secondo problema di ordine metrico sorge dall'esigenza di dare un significato univoco a una distanza di un punto variabile, o di un punto archivio, da un punto indicatore. A tal fine si è ricercata una metrica tale che l'avvicinarsi di un punto variabile, o di un punto archivio, verso un punto indicatore implichi un miglior comportamento del primo rispetto al secondo, e viceversa.[10] [11]

Entrambe le esigenze metriche, quella dettata da ragioni di riporto dei dati ad un'unica metrica e quella richiesta da motivi di uniformità interpretativa, sono state soddisfatte e conciliate proprio attraverso l'impiego del *double ranking*.[12]

2.2 I dati degli SDS esaminati

In questa applicazione è stata utilizzata l'AC sulla precedente tavola 1 popolata con 21 variabili presenti nel Quadro G degli SDS e con 7 indicatori di qualità. Il campo di osservazione degli SDS comprende tutti i soggetti aventi redditi dall'esercizio di arti e professioni inferiori ai 5 milioni annuali di euro. Per l'anno 2007 il quadro G degli SDS è stato compilato da 846.423 soggetti. Le variabili e gli indicatori trattati sono i seguenti:

[8] Ad esempio, in una tavola con N variabili (N righe) e M indicatori (M colonne), un generico vettore colonna j-esimo avrà i suoi ranghi compresi tra 1 e N, con rango 1 assegnato all'elemento con il valore più piccolo e N all'elemento con il valore più grande. Indicato con rj il vettore colonna rango desunto dal vettore j, il procedimento di double ranking sostituisce rj con due vettori: un primo, dato da rj-1, indicante la distanza (numero posti) di ogni elemento rispetto al rango minimo; un secondo dato da N-rj, rappresentante la distanza di ogni elemento rispetto al rango massimo.

[9] Greenacre 2007, p.184.

[10] Si sono definiti tutto gli indicatori in modo tale che loro valori bassi, con annessi ranghi bassi, implichino valori più soddisfacenti della variabile interessata, e valori alti con collegati ranghi alti il contrario. Come da nota 8, ad esempio, indicato con cnv il vettore colonna (indicatore) "quota % campi non valorizzati", e con rcnv il vettore colonna ottenuto da cnv sostituendo ai suoi valori i ranghi corrispettivi, l'elemento in cnv con il migliore risultato (vale a dire con la minore quota % di campi non valorizzati) avrà il valore più piccolo per cui nel vettore rango rcnv l'elemento corrispondente varrà 1, cioè per esso sarà rcnv=1. Nei due vettori sostitutivi (double ranking), per tale elemento si avrà, rispettivamente, rcnv-1=1-1=0 e rango massimo-rcnv=21-1=20. Il primo vettore sostitutivo, che avrà 0 per tale elemento, verrà indicato con il suffisso finale "m" per indicare che misura le distanze dall'elemento migliore, ed infatti l'elemento in esame è separato da zero posti rispetto all'elemento con rango minore (infatti vi coincide) che è l'elemento con la migliore prestazione. Il secondo vettore sostitutivo, che vale 20 per tale elemento, verrà distinto con il suffisso finale "p" per indicare che riporta le distanze dall'elemento peggiore, ed infatti l'elemento in discorso è lontano 20 posti dall'elemento con rango maggiore che è l'elemento con la peggiore prestazione. È evidente allora come valori bassi di rcnvm (oppure valori alti di rcnvp) indichino situazioni migliori, e viceversa.

[11] In appendice, tavola b, i dati trattati con il double ranking.

[12] Più una variabile i-esima ha un valore basso (cioè preferibile) rispetto ad un indicatore j-esimo, tanto più nel sottospazio individuato dall'AC il suo punto risulterà vicino al punto dell'indicatore con suffisso finale "p", e viceversa. Per i punti indicatore con suffisso finale "m", avendo essi coordinate eguali in modulo ma di segno opposto varrà l'esatto contrario. Essendo più facile visivamente parlando rilevare una breve distanza che una grande distanza, ne segue che è questo il motivo che ha indotto a concentrare l'esame dei punti variabile verso i punti indicatore con suffisso finale "p" trascurando, quindi, gli indicatori con suffisso finale "m".

Tavola 3 - Lista delle variabili amministrative

Identificativo	Descrizione
var1	compensi dichiarati
var2	adeguamento da studi di settore
var3	altri proventi lordi
var4	plusvalenze patrimoniali
var5	spese per prestazioni di lavoro dipendente
var6	spese per lavoro dipendente di cui per personale con contratto di somministrazione di lavoro
var7	spese per prestazioni di collaborazione coordinata e continuativa
var8	compensi a terzi per prestazioni direttamente afferenti l'attività professionale e artistica
var9	consumi
var10	altre spese
var11	minusvalenze patrimoniali
var12	ammortamenti
var13	ammortamenti - di cui per beni strumentali
var14	altre componenti negative
var15	reddito
var16	valore dei beni strumentali mobili
var17	di cui valore relativo a beni acquisiti in contratti di locazione finanziaria e non finanziaria
var18	volume di affari
var19	altre operazioni fuori campo; operazioni non soggette a dichiarazione
var20	i.v.a. sulle operazioni imponibili
var21	altra i.v.a.

Tavola 4 - Lista degli indicatori di qualità

Identificativo	Descrizione	Note	Metrica
cnv	frequenza relativa dei campi non valorizzati		cardinale
cacnv	livello di casualità dei campi non valorizzati		ordinale
o1	frequenza relativa di *outliers* con detenzione non potenziata	soglie di tolleranza più largheal di fuori delle quali si rilevano i dati estremi	cardinale
cao1	livello di casualità degli *outliers* o1		ordinale
o2	frequenza relativa di *outliers* con detenzione potenziata	soglie di tolleranza più ristrette al di fuori delle quali si rilevano i dati estremi	cardinale
cao2	livello di casualità degli *outliers* o2		ordinale
schema	trattamento statistico applicato alle variabili amministrative integrande		categorica

Riassumendo, le variabili della tavola 3[13]esplicitano le righe di tavola 1 e sono state descritte con i 7 indicatori di qualità[14] di tavola 4, che rappresentano le colonne di tavola 1, per cui quest'ultima assume la forma di una tavola a due vie con 21 righe e 7 colonne, per un totale di 147 celle, il che la può caratterizzare come tavola complessa.

Il primo indicatore, **cnv**, rappresenta una misura percentuale dei campi o celle (corrispondenti a unità d'analisi) non valorizzati. Per una corretta interpretazione, va osservato che una variabile potrebbe essere non valorizzata per un dato soggetto d'imposta non perché sia stata male rilevata ma per il semplice motivo che tale soggetto d'imposta non doveva indicarla. Allo stesso tempo l'indicatore proposto può indicare quali variabili siano maggiormente dotate di contenuto informativo e quindi siano più ricche di campi valorizzati e quali, invece, ne siano meno provviste. L'indicatore può oscillare da 0, valore per una variabile con tutti i campi valorizzati, a 100, valore per una variabile completamente priva di contenuto informativo.

Il secondo indicatore, **cacnv**, è un accertamento, per ciascuna variabile, della casualità delle osservazioni aventi campi non valorizzati. In sintesi, distinguendo per ciascuna variabile le unità ri-

[13] Rispetto ai dati presenti nel quadro G sono state escluse dall'analisi le variabili "segno del reddito" e "esenzione I.V.A." in quanto dicotomiche e, pertanto, non idonee ad essere misurate secondo alcuni degli indicatori qui trattati.

[14] I 7 indicatori di qualità proposti non sono evidentemente esaustivi e, pur essendo calcolati con i dati reali del Quadro G, hanno qui la funzione di riempire le celle di tavola 1 consentendo, così, un trial dell'AC. È in programma la presentazione, entro la fine del 2010, di un documento, da parte di ricercatori dell'Istituto, in cui verranno presentati più estesamente i trattamenti e i risultati ottenuti dal processo di trasformazione di tale archivio in registro statistico.

spondenti da quelle non rispondenti, si sono confrontate le distribuzioni di frequenza dei due gruppi rispetto a cinque importanti variabili di classificazione tramite la V di Cramer, valutando la significatività degli scarti tra le distribuzioni delle unità con campi valorizzati e le corrispondenti con campi non valorizzati. Assumendo, in prima approssimazione, che le 5 variabili di classificazione siano egualmente importanti, si è assegnato punteggio 0 alle variabili le cui V di Cramer erano tutte non significative; punteggio 1 alle variabili con una V di Cramer significativa;…; punteggio 5 alle variabili le cui V di Cramer erano tutte e cinque significative, per cui tale indicatore assume gli interi compresi tra 0, quale sinonimo di maggiore casualità, e 5, quale sinonimo di minore casualità.

Il terzo indicatore, **o1**, è il risultato di un'analisi *box-plot* su ogni variabile selezionando, per ogni variabile, come dati estremi della distribuzione quelli esterni all'intervallo *far upper fence – far lower fence*[15] e misurandone la percentuale rispetto al totale.

Il quarto indicatore, **cao1**, esegue sul terzo quanto visto per il secondo indicatore rispetto al primo (misura della casualità con metrica ordinale).

Il quinto indicatore, **o2**, opera come il precedente o1 ma restringendo a differenza di questo i limiti di tolleranza e quindi ampliando la zona all'interno della quale si possono trovare i dati estremi, data da upper fence – lower fence, seleziona evidentemente più *outlier*[16] dell'indicatore o1.

Il sesto indicatore, **cao2**, opera come i precedenti indicatori n.2 e n.4, cui pertanto si rinvia per i chiarimenti.

Il settimo indicatore, schema, infine, è una variabile categorica che indica la procedura statistica impiegata per l'analisi della variabile amministrativa sotto controllo. Attualmente sono state predisposte tre distinte procedure statistiche, di complessità crescente, a seconda che la variabile amministrativa integranda abbia una variabile di confronto diretta ed esplicita tratta da un'indagine Istat, oppure una variabile di confronto implicita frutto cioè di elaborazioni su dati Istat, oppure che sia sprovvista di una variabile di confronto.[17] Si è assegnato punteggio 1 alle variabili ricadenti nel primo caso, quello di più facile trattamento; punteggio 2 a quelle del caso con trattamento di difficoltà intermedia, ove è intuitivo che la presenza di una variabile di controllo non esplicita e da costruire richieda un rafforzamento del trattamento statistico; punteggio 3, infine, ai casi nei quali viene meno la possibilità di un confronto (con l'esterno) e si deve fare affidamento sulla verifica della coerenza interna delle variabili di un archivio amministrativo, fatto questo che implica una maggiore difficoltà di valutazione della qualità della variabile.

Una caratteristica comune a tutti gli indicatori, constatabile osservandone i criteri definitori, è che essi assumono valori bassi se vi sono situazioni ottimali e valori alti in caso contrario, un'uniformità questa che ha consentito l'univocità interpretativa illustrata nel paragrafo 2.1.

[15] Tale intervallo di tolleranza scorre da *far lower fence* a *far upper fence*, cioè rispettivamente da Q1-3xIQR a Q3+3xIQR, ove Q1 e Q3 sono il primo e il terzo quartile e IQR è la differenza interquartile, per cui esso ha una zona di tolleranza (ampiezza complessiva) pari a 7 IQR.

[16] L'intervallo di tolleranza di questo indicatore scorre da *lower fence* a *upper fence*, cioè rispettivamente da Q1-1,5xIQR a Q3+1,5xIQR, ove Q1 e Q3 sono il primo e il terzo quartile e IQR è la differenza interquartile; ha quindi una zona di tolleranza (ampiezza complessiva) pari a 4IQR, minore rispetto a quella dell'indicatore o1 pari a 7IQR.

[17] Può risultare un po' difficile pensare che una variabile economico-sociale di un archivio amministrativo non abbia variabili di riferimento esplicite e nemmeno ottenibili mediante elaborazioni. Ma sta di fatto che in alcuni SDS esistono variabili molto specifiche e dettagliate e può essere molto complicato trovare per esse un chiaro benchmark mentre, da un punto di vista del trattamento dei dati, può essere più agevole la strada di esaminarle e controllarle, in un'ottica di coerenza interna, mediante le altre variabili dell'archivio amministrativo.

3. L' Analisi delle Corrispondenze applicata ad un singolo archivio

L'AC non parametrica è stata eseguita sui dati del quadro G degli SDS dell'anno 2007 e cioè sulle 21 variabili di cui a tavola 3, formanti le categorie di riga, e sui 14 indicatori, ottenuti previa trasformazione e raddoppiamento dei 7 indicatori di base di cui a tavola 4, definenti le categorie di colonna.[18] I primi risultati dell'AC che si presentano sono quelli relativi alla bontà generale della procedura di riduzione della dimensionalità osservata dal lato dell'inerzia.

Tavola 5a - Misure dell'inerzia

Number of active rows	Number of active columns	Pearson chi2(260) [19]	Prob > chi2	Total inertia	Number of dimensions	Explained inertia (%)
21	14	837,4	0,0000	0,2848	3	86,37

Dalla tavola 5a si rileva che vi sono 21 righe, date alle 21 variabili di tavola 3, e 14 colonne, dovute ai 7 indicatori di tavola 4 trattati con il procedimento del *double ranking*. In realtà essendo solo 7 gli indicatori linearmente indipendenti, l'estrazione degli autovalori si è arrestata al 7°. L'inerzia complessiva è pari a 0,28, un valore lontano dal massimo teorico raggiungibile pari a 6,[20] indice questo di punti riga (variabili) abbastanza concentrati attorno all'origine del sottospazio di riduzione. [21]

Tavola 5b - Decomposizione dell'inerzia

Dimension	Singular value	Principal inertia	Chi2	Percent	Cumul percent
dim 1	0,38397	0,14743	433,45	51,76	51,76
dim 2	0,25750	0,06631	194,94	23,28	75,04
dim 3	0,17961	0,03226	94,84	11,33	86,37
dim 4	0,14796	0,02189	64,37	7,69	94,05
dim 5	0,12669	0,01605	47,19	5,63	99,69
dim 6	0,02844	0,00081	2,38	0,28	99,97
dim 7	0,00894	0,00008	0,24	0,03	100
total		0,28483	837,4	100	

Osservando nella tavola 5b il livello cumulato percentuale degli autovalori (colonna "*cumul percent*"), si rileva che i primi 3 spiegano oltre 86% dell'inerzia complessiva, misura che può essere giudicata abbastanza soddisfacente essendo vicina alla soglia limite convenzionale del 90%,[22] per cui emerge che l'AC ha realmente ottenuto una valida riduzione della dimensionalità nel sottospazio R^3.

3.1 Analisi dei punti riga

Nel complesso i punti riga hanno ottenuto una buona rappresentazione, come conseguenza della ridotta perdita d'inerzia prima misurata, sebbene, specie per tavole con parecchie celle come quella qui in esame, si può verificare che alcuni di essi non ricevano una rappresentazione qualitativa sufficiente.

[18] Cfr. appendice, tavola b.

[19] I 260 gradi di libertà del chi quadro derivano, come noto, dal prodotto di (righe-1) x (colonne-1)=(21-1) x (14-1). Il chi quadro, altamente significativo (p-value<0.0001), non svolge tuttavia un ruolo di rilievo nel contesto dell'AC, a differenza dell'inerzia.

[20] La rappresentazione perfetta di una tavola di frequenza, cioè senza alcuna perdita informativa e dunque con piena rappresentazione dell'inerzia, è pari al valore minimo tra [(numero righe della tavola - 1), (numero colonne della tavola - 1)], per cui qui con 7 colonne e 21 di righe l'inerzia potrebbe essere al massimo pari a 6.

[21] Una conseguenza in parte dovuta all'applicazione dei ranghi, maggiormente livellanti rispetto alle misure originali delle variabili.

[22] Cfr. Multivariate Statistics, p.28, Stata, 2009.

Tavola 6 - Risultati specifici per i punti riga, dati SDS

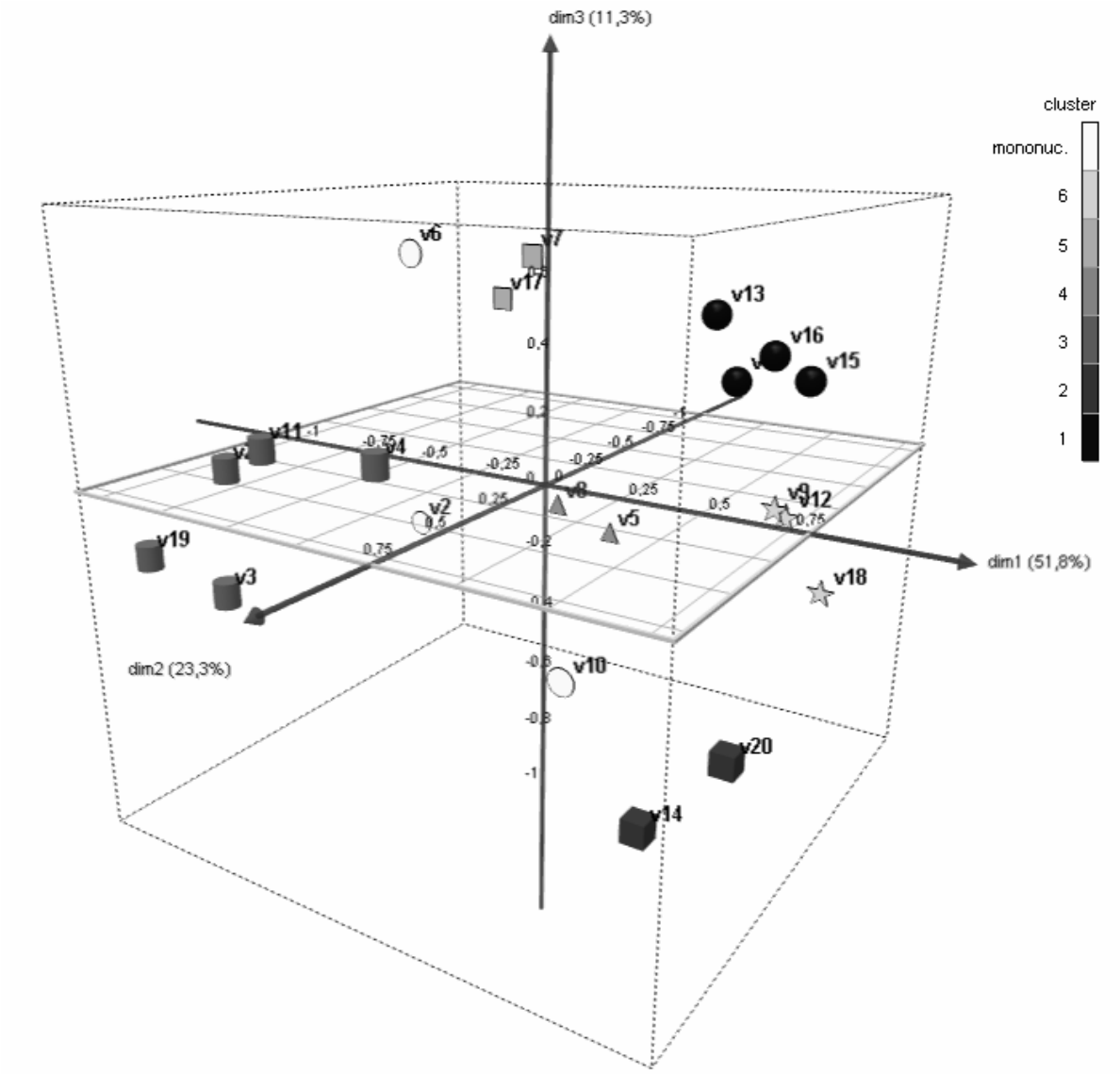
Categories	Mass	Quality	%Inert
v1	0,048	0,747	0,038
v2	0,048	0,484	0,047
v3	0,048	0,930	0,053
v4	0,048	0,878	0,016
v5	0,048	0,971	0,069
v6	0,048	0,607	0,037
v7	0,048	0,770	0,051
v8	0,048	0,910	0,074
v9	0,048	0,876	0,032
v10	0,048	0,623	0,044
v11	0,048	0,937	0,041
v12	0,048	0,912	0,034
v13	0,048	0,973	0,029
v14	0,048	0,801	0,049
v15	0,048	0,871	0,052
v16	0,048	0,951	0,040
v17	0,048	0,927	0,067
v18	0,048	0,905	0,047
v19	0,048	0,983	0,081
v20	0,048	0,966	0,043
v21	0,048	0,924	0,056

Assumendo come termine di confronto la soglia di 86% quale indicatore della qualità della rappresentazione complessiva (cfr. tavola 5 b, colonna "*cumul percent*" al 3° autovalore), si osserva infatti (colonna "*quality*") che la qualità della rappresentazione dei punti riga v2, v6 e v10 riporti livelli un po' bassi che richiederanno un supplemento d'indagine nelle prossime analisi. Si nota poi che il contributo all'inerzia dei punti riga (colonna " %*inert*") presenta una certa variabilità, con i massimi anche quadrupli dei minimi, e che le masse di riga hanno tutte lo stesso peso, dato che le variabili, anche se trasformate nel modo detto, sono state trattate come equiponderate e, pertanto, nessuna di esse ha esercitato un'influenza ingiustificata.

Il passo successivo è quello di analizzare il *symmetric plot*. Esso è uno strumento atto a rilevare i punti riga e i punti colonna simili, tenendo però nettamente separate tra loro queste due fasi di comparazione,[23] a differenza dell'*asymmetric plot* che è idoneo invece a far emergere schemi associativi tra i punti riga e i punti colonna. Per la metrica qui applicata, la ricerca di punti variabile simili significa verificare l'esistenza di variabili aventi ranghi proporzionali (rispetto agli indicatori), il che, dato l'evidente legame tra valori originari e ranghi, in pratica equivale a rinvenire una somiglianza delle variabili misurate sulla metrica originaria (e analogo discorso varrà, *mutatis mutandis*, per la ricerca di punti indicatore simili).

[23] Poiché i due gruppi di punti (punti riga e punti colonna) sono rappresentati su spazi con metriche differenti (Greenacre 2007, p.72).

Grafico 1 - Symmetric plot dei punti riga, dati SDS



Dal grafico 1 si osserva che i punti riga, cioè le variabili, formano 7 differenti *clusters*:

Tavola 7 - Gruppi dei punti riga

Cluster	Variabili	Forma punti
1	v1-v13-15-v16	sfere
2	v14-v20	cubi
3	v3-v4-v11-v19-v21	cilindri
4	v5-v8	triangoli
5	v7-v17	quadrati
6	v9-v12-v18	stelle
mononucleari	v2-v6-v10	cerchi

e l'interpretazione di un qualsiasi cluster è che le variabili in esso contenute hanno comportamenti (punteggi) simili rispetto agli indicatori, fatto che può essere confermato osservando la ma-

trice dei dati (appendice, tavola a) e quella delle correlazioni delle variabili nella metrica trasformata (appendice, tavola c).[24] Da notare infine che tutti i punti con minore qualità di rappresentazione sono risultati un po' separati dagli altri, richiedendo per se stessi un grappolo mononucleare.

3.2 Analisi dei punti colonna

Passando ora all'esame dei punti colonna (indicatori), si osserva che, come già per i punti riga, la qualità della loro rappresentazione in R3 è decisamente buona (colonna "quality"), con l'eccezione di due coppie di punti (rcnvm e rcnvp, rcacnvm e rcacnvm) che assumono livelli un po' più bassi e che richiederanno una maggiore attenzione nelle prossime analisi. Si nota infine che il contributo all'inerzia dei punti colonna (colonna " %inert") ha meno variabilità rispetto ai punti riga.

Tavola 8 - Risultati specifici per i punti colonna, dati SDS

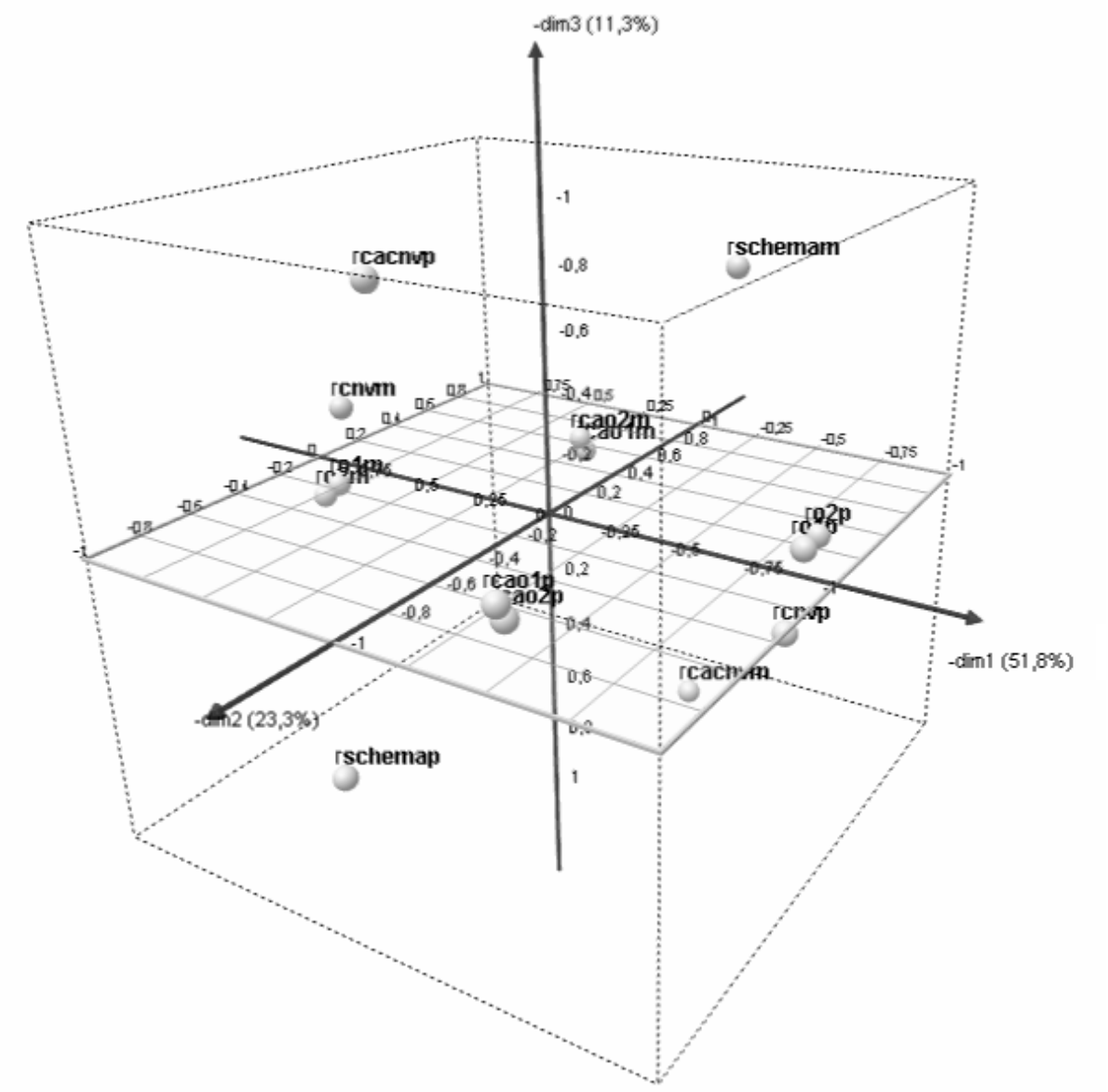
Categories	Mass	Quality	%Inert
rcnvm	0,071	0,754	0,092
rcnvp	0,071	0,754	0,092
ro1m	0,071	0,925	0,082
ro1p	0,071	0,925	0,082
ro2m	0,071	0,916	0,087
ro2p	0,071	0,916	0,087
rcacnvm	0,071	0,838	0,064
rcacnvp	0,071	0,838	0,064
rcao1m	0,071	0,906	0,051
rcao1p	0,071	0,906	0,051
rcao2m	0,071	0,856	0,051
rcao2p	0,071	0,856	0,051
rschemam	0,071	0,869	0,073
rschemap	0,071	0,869	0,073

La prima colonna della tavola riporta gli identificativi degli indicatori (tavola 4), cui è stato aggiunto, come indicato nelle note 10 e 12, il prefisso "r", per indicare che i valori originari di ogni colonna sono stati sostituiti dai rispettivi ranghi, e le finali "m" e "p", a causa del double ranking che ha raddoppiato e sostituito le variabili rango richiedendone, così, una distinzione. Le colonne riportano, come già indicato per tavola 6, la massa relativa di ogni punto riga e di ogni punto colonna, la qualità complessiva della rappresentazione e la parte di inerzia (di variabilità) nel sottospazio R3 che è spiegato dai singoli punti. La colonna della qualità è molto importante, poiché indica se vi siano punti male rappresentati in R3. Sempre assumendo come termine di confronto la soglia di 86% quale indicatore della qualità della rappresentazione complessiva (cfr. tavola 5 b, colonna "cumul percent" al 3° autovalore), si osserva che è meno bene rappresentata soprattutto la coppia di punti colonna rcnvm e rcnvp, sebbene la loro sotto rappresentazione non assuma valori tali da consigliarne l'esclusione dalle analisi. Il grafico infine è utile poiché anticipa l'asymmetric plot che sarà basato su punti profilo con coordinate identiche a quelle qui osservate e su punti vertice espansi secondo opportuni fattori di proporzionalità.[25]

[24] Tale raffronto va fatto con qualche cautela, dato che la matrice delle correlazioni è quella di ordine zero, cioè esamina correlazioni singole, mentre il *symmetric plot* assomiglia più ad una correlazione multipla, individuando simultaneamente blocchi di variabili correlate tra loro.

[25] I *punti profilo* (punti variabili) manterranno le stesse *coordinate principali*, mentre i *punti vertice* (punti indicatori) saranno espansi da fattori atti a riportarli nello stesso spazio metrico – da cui la possibilità di confronti incrociati - in *coordinate standard* (Greenacre 2007, p.60 e segg.).

Grafico 2 - *Symmetric* plot dei punti colonna, dati SDS



Premesso che nello spazio R3 i punti colonna (indicatori) con suffisso finale pari a "p" hanno coordinate eguali ma con segno invertito rispetto a quelli con suffisso finale "m",[26] per cui è sufficiente esaminare solo i primi,[27] si osservano i seguenti patterns. Primo, sembrano simili gli indicatori ro1p e ro2p (rispettivamente detenzione non potenziata e potenziata di outliers), per cui aver reso più sensibile la rilevazione dei dati estremi non ha portato a un reale incremento conoscitivo. Secondo, sono simili gli indicatori rcao1p e rcao2p (rispettivamente livello di causalità dei dati mancanti con detenzione minore e maggiore di outliers),

[26] Si può dimostrare (Greenacre 2007, pp.183-184) che questa specularità delle coordinate dipende dal fatto che i ranghi (trasformati via double ranking) hanno gli stessi valori medi (appendice, tavola b, ultima riga).

[27] Si rinvia alla precedente nota 12 per i motivi di tale scelta.

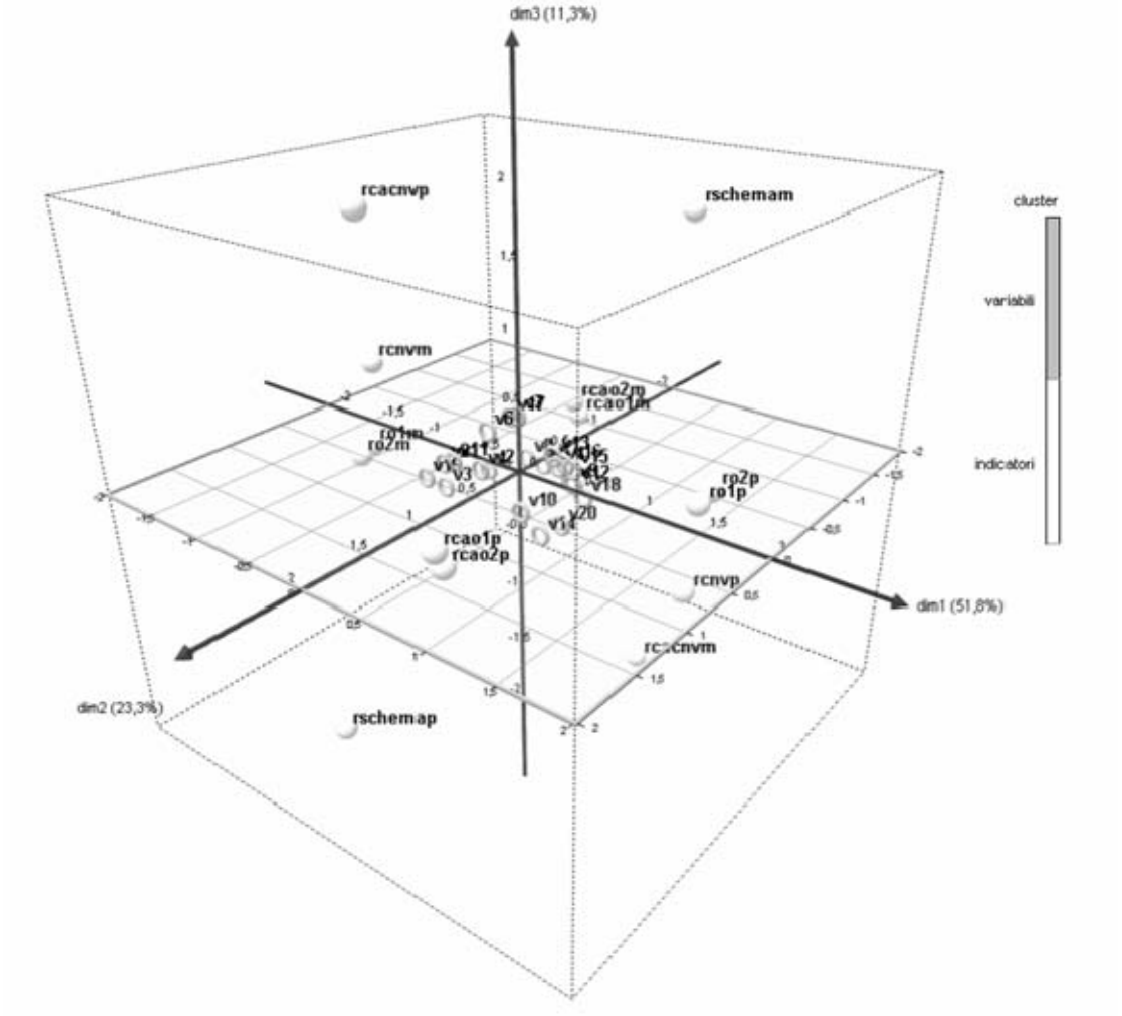
proprio per il fatto che le indicate caratteristiche di ro2 e ro1 si sono riflesse anche nella misura della causalità loro associata. Terzo, il punto indicatore rcnvp (livello dei campi non valorizzati) è vicino ai punti indicatore ro2p e ro1p, ad indicare che variabili con un numero basso di valori estremi hanno pure un limitato numero di campi non valorizzati (e viceversa), in accordo con quanto osservabile direttamente nel database (appendice, tavola a), fatto importante questo poiché sembra confermare una buona qualità dei dati. Quarto, gli indicatori di causalità sono separati dagli altri punti, evidenziando opportunamente un'indipendenza delle misure di causalità dagli altri indicatori.

Quinto, rcacnvp (livello di causalità dei campi non valorizzati) e rschemap (schema di trattamento statistico) sono i punti indicatore più lontani, cioè i meno associati, dagli altri punti variabile.

3.3 L'analisi dei punti riga e dei punti colonna

Con l'*asymmetric plot* è possibile esaminare l'associazione tra i 21 punti variabile e i 14 punti indicatori, come indicato al paragrafo 2.2. In esso una minore distanza di un punto variabile da un punto indicatore con suffisso finale "p" (oppure una maggiore distanza di un punto variabile da un punto indicatore con suffisso finale "m") è il riflesso di ranghi (e degli annessi valori) più bassi e quindi di risultati migliori della variabile rispetto all'indicatore (e viceversa). Questo è dovuto proprio al modo con cui sono stati definite le variabili e i ranghi, ed è proprio questa caratteristica che rende univoca l'interpretazione di una distanza di un punto variabile da un punto indicatore.

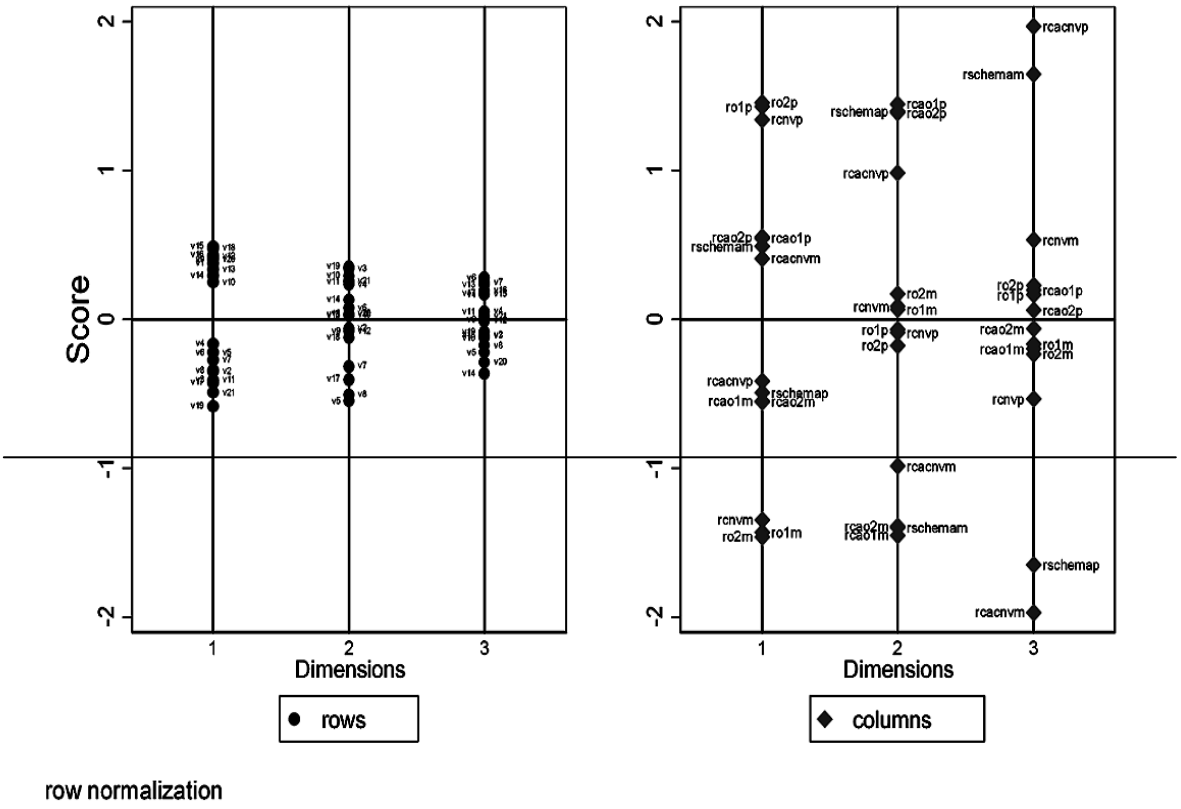
Grafico 3 - *Asymmetric* plot, dati SDS



Premesso quanto sopra, si osserva sul grafico che nello spazio R3 individuato dall'AC la posizione dei punti variabile (aventi simbolo di anello) è abbastanza concentrata attorno all'origine, mentre quella dei punti indicatore (aventi per simbolo la sfera) è dispersa sulla parte più esterna; tale disposizione è in accordo sia con i ruoli di punti profilo e punti vertice loro rispettivamente assegnati, sia (per i punti profilo) con la bassa inerzia rilevata (tavola 5a).

Tuttavia lo schiacciamento dei punti variabile, originato dall'estensione dei punti indicatore e fenomeno di per sé tipico dell'*asymmetric plot*, comporta indubbiamente una maggiore difficoltà di individuazione di schemi relazionali tra questi e i punti indicatore, specie quando, come visto, i primi sono sensibilmente addensati attorno all'origine. È allora opportuno esaminare i punti in due modi: per quelli aventi l'ordinata sul terzo asse relativamente piccola è sufficiente evidentemente studiarne la posizione in un grafico bidimensionale,[28] mentre per quelli con l'ordinata sul terzo asse non trascurabile è necessaria un'analisi con rappresentazione in 3D sotto forma di snapshot, coadiuvata da opportuni ingrandimenti e rotazioni dello scatter e da una parallela analisi della matrice delle distanze euclidee. Nella parte finale del documento questo diverso modo di procedere sarà naturalmente oggetto di alcune considerazioni.

Grafico 4 - *Asymmetric* plot, coordinate punti sui primi tre assi dimensionali, dati SDS



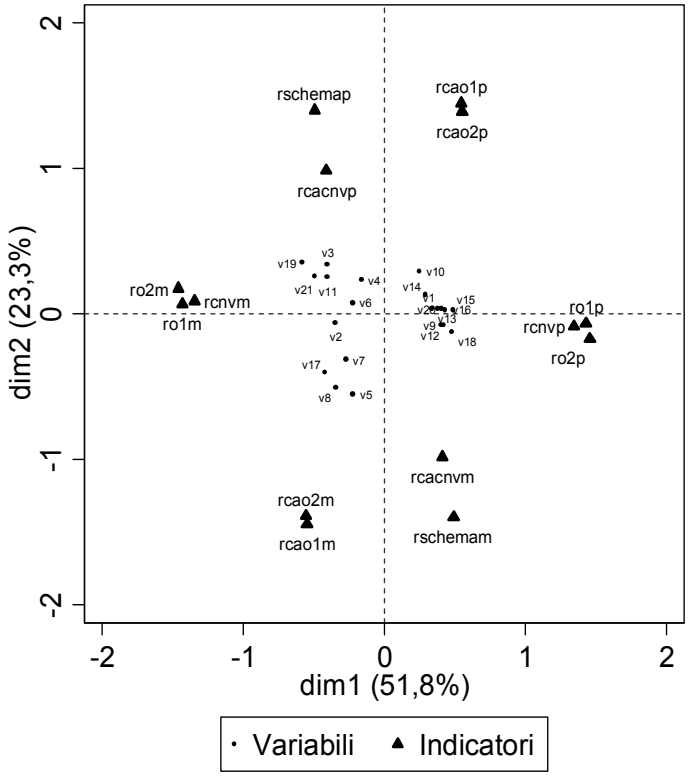
Il terzo asse, come già appariva dal grafico 3, sembra assumere una dimensione di rilievo soprattutto per i punti indicatore rcacnv e rschema, mentre per gli altri punti indicatore e per tutti i punti variabile appare meno rilevante.[29] Quindi, per 5 indicatori su 7, vale a dire ro1, ro2, rcao1,

[28] Nell'AC gli assi sono ortogonali, cioè le coordinate sui primi due assi restano invariate sia che si estragga, oppure no, un terzo asse.

[29] Cfr. appendice, tavola d, ove sono riportate le differenze percentuali tra la matrice delle distanze calcolata con i primi due assi e la stessa calcolata con i primi tre assi.

rcao2 e rcnv, appare possibile una rappresentazione semplificata in un piano bidimensionale, mentre per i due restanti, rcacnv e rschema, si dovrà necessariamente fare ricorso a rappresentazioni 3D con gli accorgimenti ad hoc prima indicati.

Grafico 5 - *Asymmetric* plot, associazione punti variabili con punti indicatori rcnv, ro1, ro2, dati SDS



Si osserva nel piano bidimensionale la presenza nitida di due gruppi di punti variabile, uno a destra della linea tratteggiata verticale - costituito dai punti v1, v9, v10, v12, v13, v14, v15, v16, v18 e v20 - che è più vicino ai punti indicatore ro1p, ro2p e rcnvp,[30] [31] e uno a sinistra, definito dai punti restanti, per il quale varrà il contrario.

Emerge quindi un primo schema relazionale, costituito da variabili che hanno valori migliori rispetto agli indicatori dei campi non valorizzati e dei dati estremi. Queste associazioni suggeriscono che i punti indicati abbiano una qualità migliore, nel senso che per essi, meno frequenti sono i campi non valorizzati, minori tendono ad essere i dati lontani dal centro della distribuzione (e viceversa).

Un secondo schema relazionale, è costituito dai punti variabile che sono vicini ai punti indicatori rcao1p e rcao2p. L'individuazione di tali punti variabile avviene più facilmente per esclusione, cioè osservando il loro insieme complementare dato dai punti variabile lontani, vale a dire v2, v5, v7, v8 e v17, per cui tutti i punti variabile non citati sono vicini ai punti indicatori rcao1p e rcao2p.

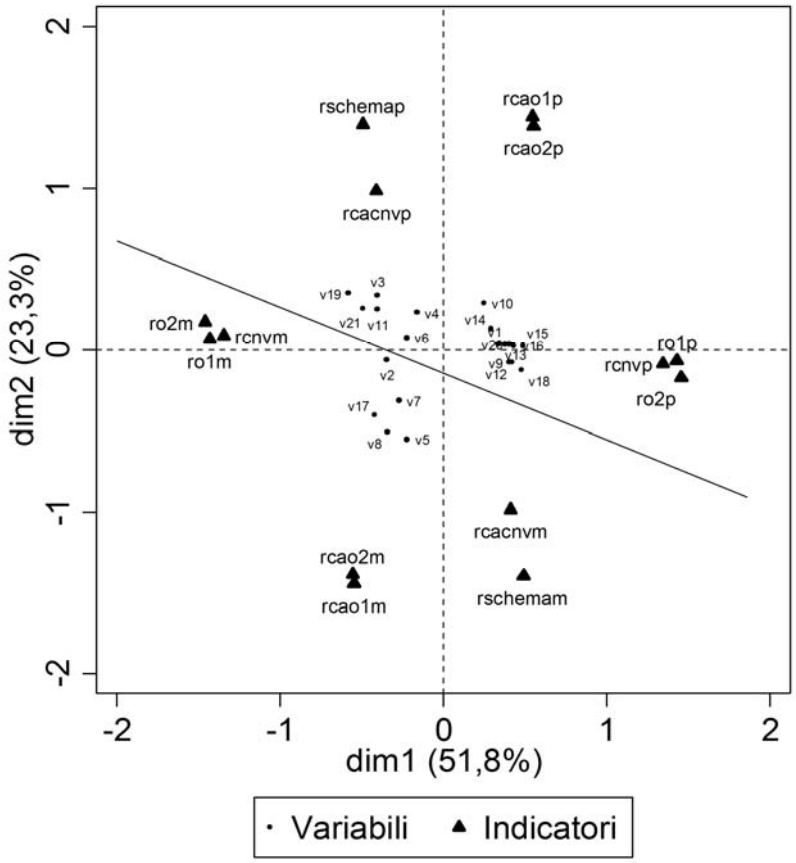
Un terzo schema relazionale osservabile è un di cui dei due precedenti. Se si traccia (grafico 6) una linea di separazione con pendenza a quasi -45° e passante poco sotto l'origine, si osserva che i

[30] E più lontano dai corrispondenti indicatori con suffisso finale "m".

[31] Un'altra caratteristica tipica dell'AC è la forma di horseshoe – detta anche arch effect - assunta dai punti profilo, come sembra di vedere anche in questo grafico sui punti profilo, aventi forma di "U" rovesciata e leggermente ruotata in senso antiorario.

punti variabile di cui al primo schema (v1, v9, v10, v12, v13, v14, v15, v16, v18 e v20), quelli per intendersi vicini a ro1p, ro2p e rcnvp, sono pure a minore distanza da rcao1p e da rcao2p, per cui tali punti hanno l'ulteriore attrattiva di avere misure migliori nel livello di casualità dei dati estremi.

Grafico 6 - Asymmetric plot, associazione punti variabili con punti indicatori rcao1 e rcao2, dati SDS



L'esame di schemi associativi per i due ultimi punti indicatore rcacnvp e rschemap fin qui non trattati è, come detto, rischioso farlo su grafici bidimensionali, ed in effetti si potrebbe facilmente dimostrare che le distanze euclidee calcolate nel sottospazio ottimale R^3 sono solo in parte coincidenti con quelle presenti sui grafici 5 e 6.[32] Tranne le situazioni nelle quali i punti profilo tendono ad avere un'alta inerzia e quindi ad essere prossimi ai punti vertice, situazioni che ragionevolmente non sono tra le più frequenti, occorre riconoscere che è difficile, ad occhio, definire le distanze tra i punti su scatter 3D tipo quello del grafico 3. Occorre in questi casi ricorrere agli snapshot prima indicati, leggendoli anche con l'ausilio della sottostante matrice delle distanze euclidee.

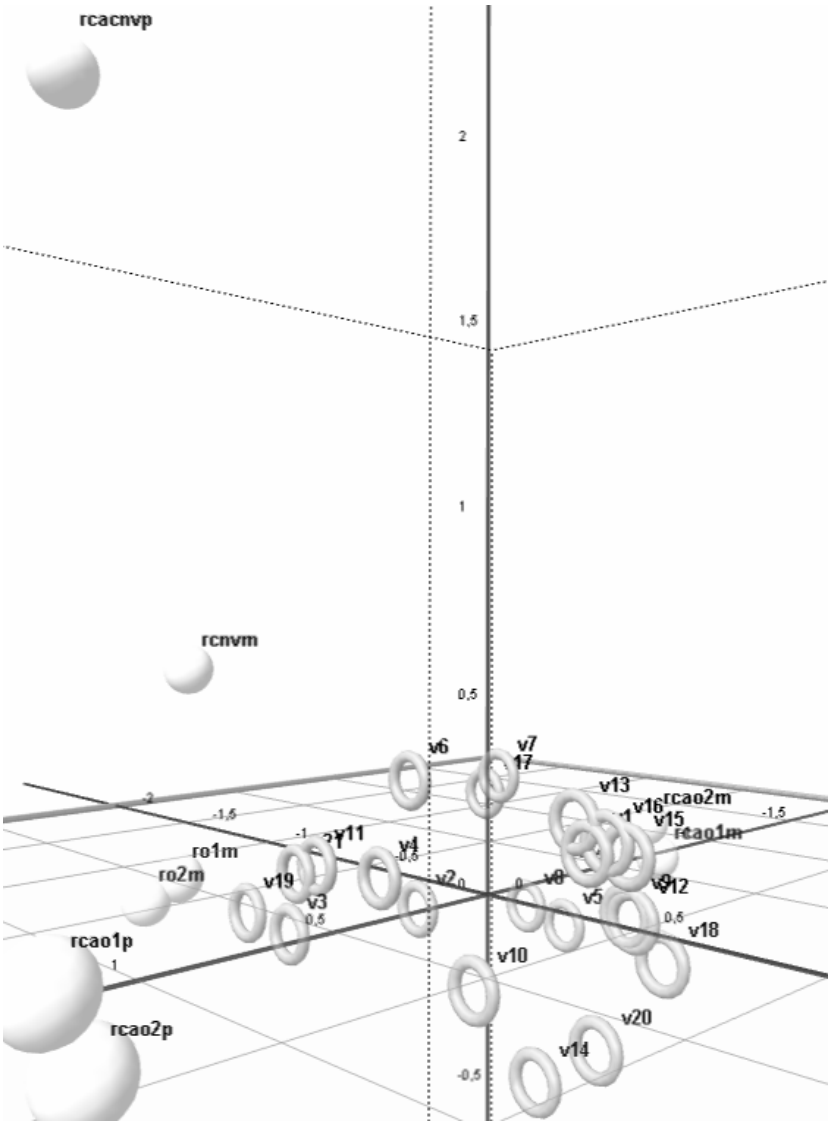
Tavola 9 - Distanze euclidee tridimensionali punti variabile rispetto a punto indicatore rcacnvp, dati SDS

Var 6	Var 11	Var 4	var 21	Var 13	Var 19	Var 7	Var 3	Var 16	Var 1	Var 15	Var 17	Var 10	Var 2	Var 9	Var 12	Var 18	Var 20	Var 14	Var 8	Var 5
1,9	2,1	2,1	2,1	2,1	2,2	2,2	2,2	2,2	2,2	2,2	2,3	2,3	2,32	2,38	2,39	2,52	2,58	2,58	2,61	2,68

[32] Ad esempio, con riferimento alla zona delimitata dal semipiano sopra la linea a circa -45° e dal semipiano a sinistra della linea tratteggiata, i punti v3 e v19 sono più vicini dei punti v4,v6,v11 e v21 rispetto all'indicatore rcacnv, mentre misurando le corrispondenti distanze in R3 tale ordine di vicinanze si inverte.

Le distanze sono riportate in ordine crescente, da sinistra verso destra, e un possibile *snapshot*, ottenuto dopo vari e non facili tentativi, è dato qui di seguito.

Grafico 7 - Asymmetric plot, associazione punti variabili con punto indicatore rcacnv, dati SDS



Il grafico, in forma di portrait al fine di preservare lo stesso aspect ratio,[33] rende visivo quanto osservabile nella tavola 9, e cioè la presenza di tre grappoli di punti variabile, un primo mononucleare dato da v6, un secondo dato da 15 punti decorrenti dal punto v11 fino al punto v12, e un terzo dato da 5 punti decorrenti dal punto v18 fino al punto v5, nettamente arretrati rispetto agli altri punti. Per tali insiemi vale quanto osservato in precedenza, cioè minore la distanza dall'indicatore rcacnvp maggiore la casualità presente nei campi non valorizzati, e viceversa.

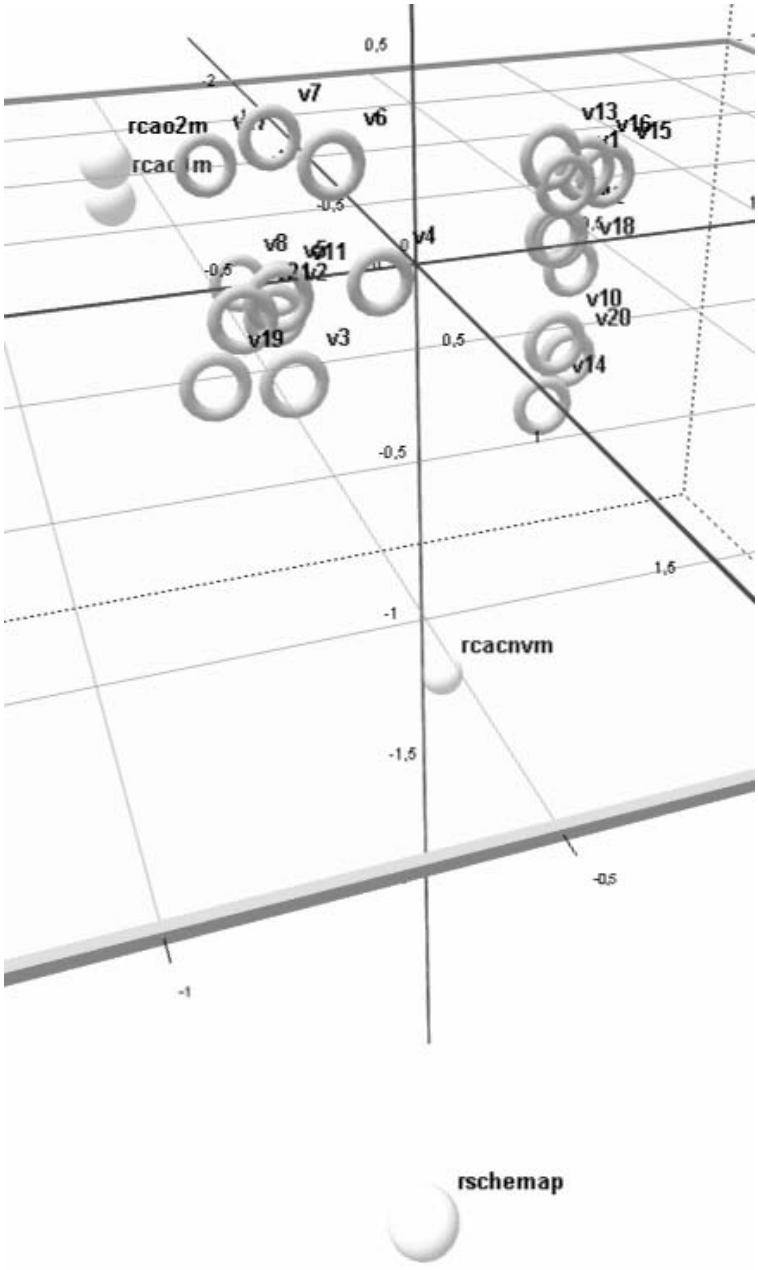
[33] Cioè la stessa unità di misura sui tre assi.

Discorso analogo si può fare per l'ultimo punto indicatore rschemap, del quale pure si riportano le distanze euclidee dai punti variabile.

Tavola 10 – Distanze euclidee tridimensionali punti variabile rispetto a punto indicatore rschemap, dati SDS

Var 3	Var 19	Var 14	Var 10	Var 21	Var 11	Var 4	Var 20	Var 2	Var 6	Var 18	Var 9	Var 12	Var 8	Var 1	Var 5	Var 13	Var 15	Var 16	Var 7	Var 17
1,87	1,88	1,97	2,02	2,02	2,05	2,09	2,12	2,13	2,35	2,37	2,39	2,39	2,41	2,43	2,43	2,47	2,48	2,48	2,56	2,56

Grafico 8 - *Asymmetric* plot, Associazione punti variabili-punto indicatore rschemap, dati SDS



Anche questo grafico, in forma di *portrait* come il precedente per gli stessi motivi, rende chiaro quanto rilevabile in tavola 10 e cioè che esistono tre gruppi di punti variabile: un primo grappolo, dato dai 2 punti v3 e v19, poi un insieme dato da 7 punti decorrenti da v14 fino a v2, infine un terzo gruppo dato dai 12 punti restanti, cioè da v6 fino a v17.

Per tali insiemi vale quanto osservato in precedenza, cioè minore la distanza dall'indicatore rschemap più agevole (migliore) la procedura di trattamento statistico applicato, e viceversa. Da notare, infine, che tale indicatore ha, a causa della terza coordinata, una posizione nel sotto-spazio come contrapposta rispetto all'indicatore rcacnvp, facendo così emergere una sorta di *trade off* tra i due.

4 - L' Analisi delle Corrispondenze applicata ad archivi integrati

Si presenta ora un'applicazione della precedente metodologia al complesso di archivi amministrativi integrati dall'Istituto di statistica dell'Olanda (CBS),[34] limitando l'analisi all'associazione incrociata dei punti archivio con i punti indicatore.

Tavola 11 - Indicatori di qualità per un complesso di archivi integrati, dati CBS

		Indicatori				
		Fornitura	Rilevanza	Sicurezza dei dati	Consegna	Procedure
Archivi	pa	+	+	+	o	+
	sfr	+	+	+	+	+/o
	cwi	+	+	+	-	+
	err	+	o	+	+	+/o
	ghe	+	+	+	+	+/o
	gse	o/-	+	+/o	o	+/o

Legenda: (+) valutazione "buona"
(+/o) valutazione intermedia tra "buona" e "ragionevole"
(o) valutazione "ragionevole"
(o/-) valutazione intermedia tra "ragionevole" e "scarsa"
(-) valutazione "scarsa"

Il primo archivio, **pa** (Policy record Administration), è gestito dall'ente governativo Institute for Employee Benefit Schemes su assegnazione del Ministero degli Affari Sociali, e riporta dati su tutti gli addetti e i pensionati aventi copertura assicurativa. Il secondo archivio, **sfr** (Student Finance Register), mantenuto da Information Management Group, contiene informazioni su tutti gli studenti che hanno ricevuto una borsa di studio. Il terzo archivio, **cwi** (Centre for Work and Income), trattato da Institute for Employee Benefit Schemes, riporta dati sulle persone in cerca di occupazione. Il quarto archivio, **err** (Exam Results Register), predisposto da Information Management Group, tratta dei risultati ottenuti agli esami da ragazzi e ragazze nella scuola secondaria. Il quinto archivio, **ghe** (Central Register of Higher Educational Enrolment), è un registro sui livelli di educazione più elevati. Il sesto e ultimo archivio, **gse**, trattato da Information Management Group, è un registro più recente che riceve informazioni scolastiche dal Ministero dell'Educazione Cultura e Scienza, dall'Ispettorato per l'Educazione, dal Secondary Education Council e dallo stesso CBS.

Gli indicatori di qualità della tavola sono riferiti agli archivi nel loro complesso e forniscono una valutazione qualitativa sui seguenti aspetti. La fornitura (F) è una misura del grado di facilità o difficoltà nelle relazioni con l'ente amministrativo detentore dell'archivio. La rilevanza (R) è una misura sull'importanza, sulle potenzialità, sulle richieste degli utilizzatori e sulla riduzione del carico statistico. Il rispetto delle norme per la sicurezza dei dati (S) rappresenta una misura delle eventuali operazioni necessarie per dotare l'archivio dei requisiti atti a soddisfare le norme legali in materia. La consegna dei dati (C) rappresenta una misura sulle operazioni richieste in materia di puntualità, eventuale necessità di selezionare una parte dei dati, fabbisogni per decodifiche e spese complessive per l'utilizzo della fonte. Le procedure (P) è una mi-

[34] Dati tratti da Daas P.J.H., Ossen S.J.L., Arends T.J., 2009, e qui riportati con alcuni adattamenti.

sura complessiva sul livello di familiarità o scarsa conoscenza del modo con cui i dati sono stati raccolti dall'ente detentore e di eventuali cambiamenti pianificati nella loro raccolta, delgrado di facilità o di difficoltà nell'avere un *feedback* con l'ente amministrativo, operazioni d'emergenza richieste qualora i dati non venissero consegnati come previsto dagli accordi.

La tavola 9, concettualmente corrispondente alla tavola 2, ha necessitato una trasformazione preliminare per poter essere oggetto dell'AC, e cioè si sono trasformati i suoi punteggi, aventi valori categorici ma ordinabili, codificando 0 al posto del segno "+", 1 al posto del segno "+/o", 2 al posto del segno "o", 3 al posto del segno "o/-" e infine 4 al posto del segno "-".[35] Ai dati così ricodificati (appendice, tavola e) poi è stata applicata la procedura del double ranking, con le categorie di riga (punti archivio) trattate come punti profilo in coordinate principali e con le categorie di colonna (punti indicatori) gestiti, in quanto punti di riferimento, come punti vertice in coordinate standard.[36]

Tavola 12 a - Misure dell'inerzia, dati CBS

Number of active rows	Number ofactive columns	Pearson chi2(45) [37]	Prob > chi2	Total inertia	Number of dimensions	Explained inertia (%)
6	10	39,6	0,6994	0,264	3	99,13

Tavola 12 b - Decomposizione dell'inerzia, dati CBS

Dimension	Singular value	Principal inertia	chi2	percent	Cumul percent
dim 1	0,371052	0,13768	20,65	52,15	52,15
dim 2	0,307068	0,094291	14,14	35,72	87,87
dim 3	0,172447	0,029738	4,46	11,26	99,13
dim 4	0,047872	0,002292	0,34	0,87	100
total		0,264	39,6	100	

L'inerzia (tabella 10a) non è molto differente rispetto alla precedente applicazione (differenza attorno al 10%, sebbene qui il suo massimo teorico sia pari a 4 e quindi inferiore al precedente, eguale a 6, di tavola 5a), ma (tabella 10b) con 3 dimensioni (sottospazio R^3) è quasi spiegato il 100% della stessa, praticamente non vi è perdita informativa, e ciò è la conseguenza della minore dimensionalità originaria dei dati CBS che, pertanto, si adattano meglio ad essere riportati in un sottospazio. Da notare che anche con due sole dimensioni si raggiunge comunque un interessante livello esplicativo dell'inerzia, pari a 88%.

Tavola 13 - Risultati specifici per i punti riga e colonna – dati CBS

Categories	mass	quality	%inert
pa	0,167	0,956	0,116
sfr	0,167	0,999	0,081
cwi	0,167	0,985	0,222
err	0,167	1,000	0,202
ghe	0,167	0,999	0,081
gse	0,167	1,000	0,298
rFp	0,100	0,996	0,076
rRp	0,100	1,000	0,076
rSp	0,100	0,996	0,076
rCp	0,100	0,989	0,152
rPp	0,100	0,983	0,121

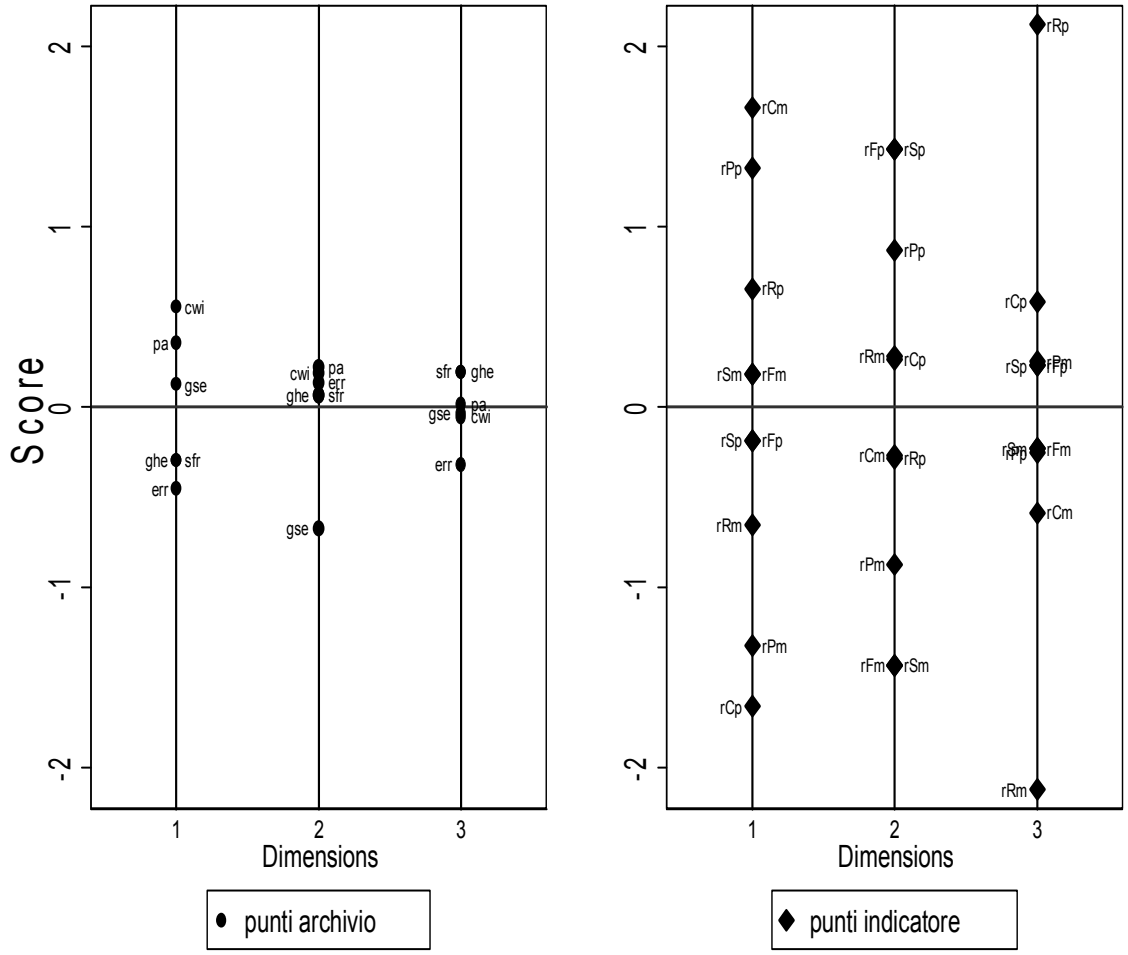
[35] In modo che, come per i dati SDS, valori minori implichino migliori prestazioni, e viceversa.

[36] Ai nomi degli indicatori, sulla falsariga precedente, si è anteposto la sigla "r" per indicare il passaggio ai ranghi e i suffissi finali "m" e "p" a causa del *double ranking*, ottenendo così la nuova lista degli indicatori impiegati nell'AC: rFm e rFp, rRm e rRp, rSm e rSp, rCm e rCp, rPm e rPp.

[37] I 45 gradi di libertà del chi quadro derivano, come noto, dal prodotto di (righe-1)x(colonne-1)=(6-1)x(10-1), ed il chi quadro questa volta non è significativo (p-value>0.05). Anche in questo caso si segnala che, nell'ambito dell'AC, non è rilevante il chi quadro ma l'inerzia.

Quanto sopra risulta naturalmente confermato dalla qualità della rappresentazione dei singoli punti, che sono tutti assestati su un livello ottimo.

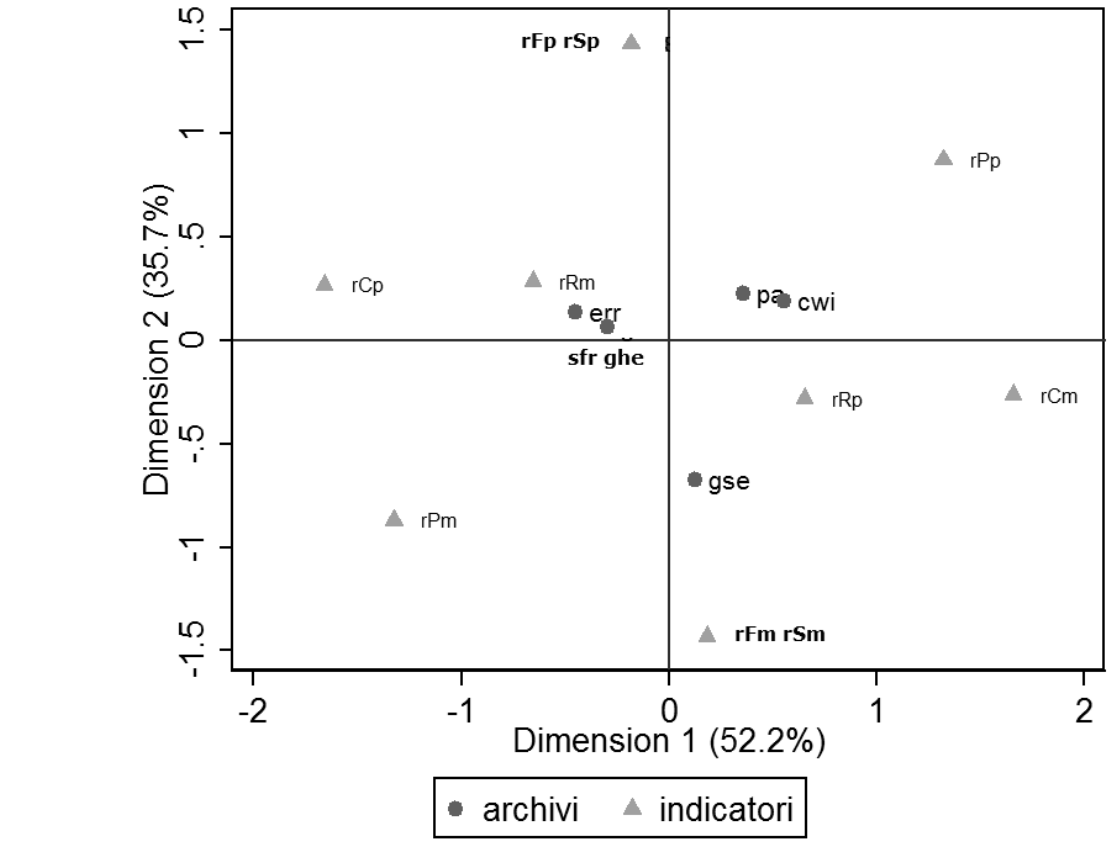
Grafico 9 - *Asymmetric* plot, coordinate punti sui primi tre assi dimensionali, dati CBS



row normalization

Il grafico mostra che il terzo asse appare rilevante solo per il punto indicatore rRp e, quindi, la gran parte delle relazioni punti archivio con punti indicatore può essere esaminata in un piano, mentre solo per il citato indicatore si esaminerà uno *scatter* 3D.

Grafico 10 - *Asymmetric* plot, associazione punti archivi con punti indicatori rF,rR, rS, rC, dati SDS



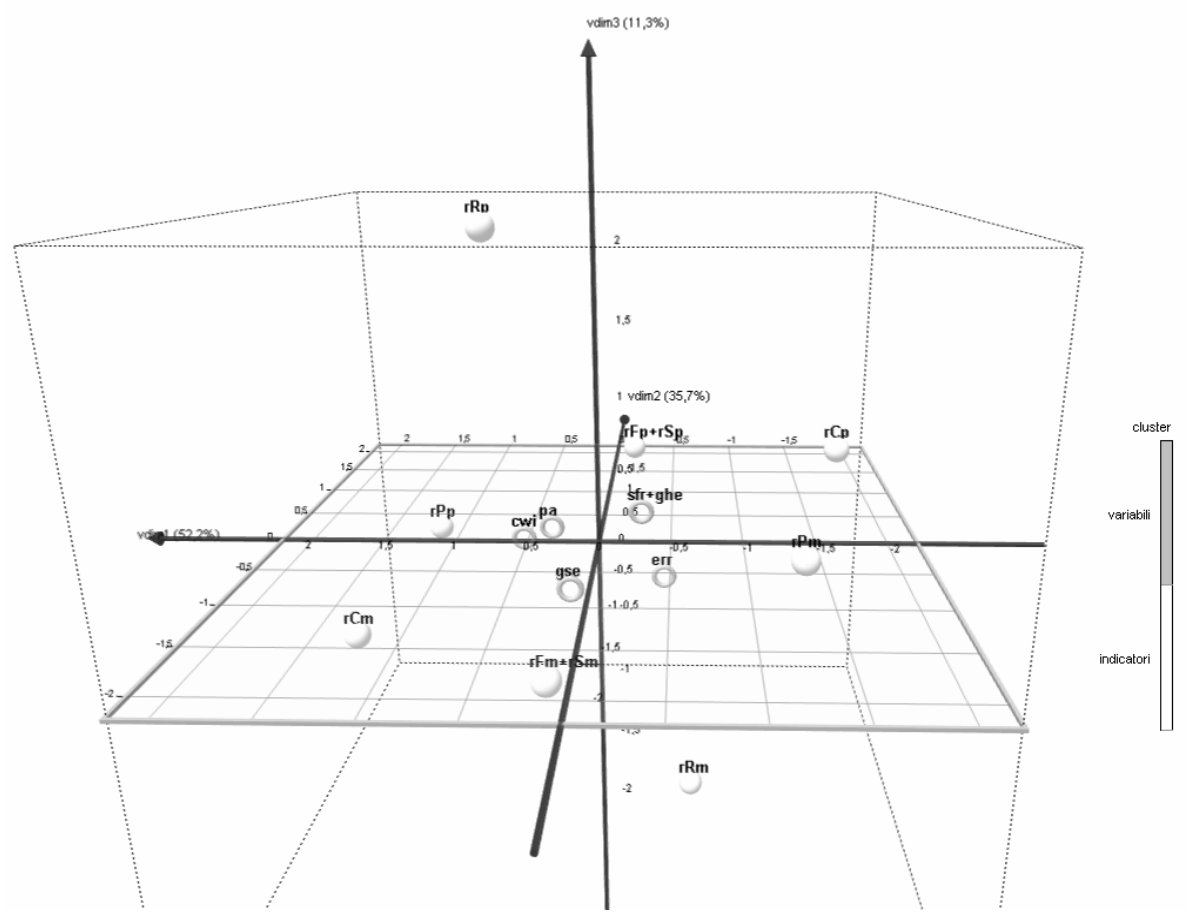
coordinates in row normalization

Premesso che due punti archivio, sfr e ghe, e due punti indicatore, rF e rS, hanno le stesse coordinate[38], si osserva che la rappresentazione bidimensionale appare abbastanza fedele ai dati di base. Ad esempio, rispetto all'indicatore rCp (consegna) i punti archivio pa, cwi e gse sono più lontani, ed infatti nel *dataset* ricodificato (appendice, tabella e) i suddetti archivi hanno prestazioni meno buone rispetto agli altri punti archivio.

Per il punto indicatore rRp, poiché esso, come detto, presenta un'ordinata sul terzo asse assolutamente non trascurabile, verrà ora esaminato con uno *snapshot* accompagnato da una apposita rotazione dello *scatter* 3D.

[38] L'uguaglianza dei punti archivio sfr e ghe è presente nei dati originari, mentre quella dei punti indicatore F ed S dipende da come sono gestiti i *ties*: qui si è usato il valore medio dei ranghi appaiati, poiché tale regola, determinando eguali valori medi tra gli indicatori con suffisso finale "m" e quelli con suffisso finale "p", provoca coordinate contrapposte tra i due insiemi di indicatori e questo rende possibile esaminarne solo uno dei due, mentre se si fosse impiegata una regola di gestione dei *ties* che non avesse determinato eguali valori medi tra gli indicatori con suffisso finale "m" e quelli con suffisso finale "p" si sarebbe reso necessario un supplemento d'analisi, e cioè sarebbe stato necessario per ogni punto archivio (per ogni punto profilo cioè) verificarne le diverse distanze sia da un indicatore con suffisso finale "m" sia dallo stesso con suffisso finale "p".

Grafico 11 - *Asymmetric* plot, Associazione punti variabili-punto indicatore rRp, dati CBS



A causa del ridotto numero di punti, e nonostante la compressione dei punti archivio (*punti profilo*) tipica dell'*asymmetric plot*, si vede bene che, relativamente al punto indicatore rRp, i punti archivio cwi, pa, sfr+ghe e gse hanno distanze minori rispetto al punto err, per cui è quest'ultimo che riporta le peggiori prestazioni rispetto all'indicatore in esame, come si può accertare dai dati originari (appendice, tabella e). Anche in questo caso, infine, è stata necessaria l'analisi dalla matrice delle distanze euclidee.

Osservazioni conclusive

L'obiettivo della ricerca consisteva nel verificare se l'analisi delle corrispondenze, impiegata con la modalità del *double ranking* per la duplice esigenza di unificazione delle metriche eterogenee dei dati e di univocità interpretativa delle posizioni dei *punti profilo e vertice*, fosse uno strumento idoneo a sintetizzare le relazioni tra le variabili di un archivio amministrativo e gli indicatori di qualità annessi, nonché i rapporti intercorrenti tra un complesso di archivi amministrativi integrati e gli indicatori di qualità ad essi preposti.

Esistono anche altre metodiche con cui produrre *report* compatti, come gli indicatori compositi, oppure e con una diversa prospettiva l'analisi fattoriale e il multidimensional scaling, ma in questa ricerca si è voluto sperimentare l'analisi delle corrispondenze perché essa ha la peculiarità di trasformare entrambe le categorie di riga e di colonna di una tavola in altrettanti *punti relazionali* aprendo, così, la strada ad esami sia interni sia incrociati sui medesimi.

Per le verifiche indicate si sono scelti due *dataset* tra loro differenti non solo, evidentemente, per i contenuti - il primo, gli SDS, riferito ad un archivio singolo, il secondo, le fonti integrate CBS, ad un complesso di archivi integrati - ma anche per la dimensionalità, con i dati del primo aventi un maggior numero di celle rispetto ai dati del secondo e, pertanto, potenzialmente meno atti a ricevere un'operazione di riduzione della dimensionalità.

La valutazione che si può trarre dai risultati emersi è che il metodo sperimentato è, in generale, capace di sintetizzare entrambe le fattispecie di archivi esaminati, anche quando essi sono costituiti da tavole complesse come quella degli SDS, quasi 150 celle. La tecnica d'indagine, apparsa assai efficace nel caso di riduzione della dimensionalità a spazi bidimensionali, nel caso di contrazione a spazi tridimensionali, specie con un numero non piccolo di punti da riportare, ha richiesto invece un certo sforzo, sia per la ricerca della migliore combinazione tra rotazioni e ingrandimenti con cui elaborare lo *scatter*, sia per il calcolo della matrice delle distanze euclidee (per brevità qui omessa), senza la quale, va detto con chiarezza, sarebbe stato rischioso trarre delle conclusioni. Tuttavia, anche in tali circostanze critiche, la sintesi grafica fornita delle relazioni tra i punti è apparsa un utile strumento, in quanto ha contribuito a chiarire e puntualizzare la struttura dei dati.

Appendice

Tabella a - *Dataset* SDS

identificativo		cnv frequenza % relativa dei campi non valorizzati	o1 detenzione nonpotenziata di *outliers* in %	o2 detenzione potenziata di *outliers* in %	cacnv livello di casualità dei campi non valorizzati	cao1 livello di casualità degli *outliers* o1	cao2 livello di casualità degli *outliers* o2	schema trattamento statistico applicato alle variabili integrande
v1	compensi dichiarati	1,21	0	0,13	0	0	0	3
v2	adeguamento da studi di settore	88,61	4,07	8,55	0	1	2	1
v3	altri proventi lordi	87,18	7,09	11,95	0	0	0	1
v4	Plusvalenze patrimoniali	95,59	3,55	8,32	0	0	0	2
v5	spese per prestazioni di lavoro dipendente	85,71	4,87	9,48	3	2	2	3
v6	spese per lavoro dipendente di cui per personale con contratto di somministrazione di lavoro	99,91	4,22	7,89	0	0	0	3
v7	spese per prestazioni di collaborazione coordinata e continuativa	98,47	3,27	7,66	0	1	2	3
v8	compensi corrisposti a terzi per prestazioni direttamente afferenti l'attività professionale e artistica	69,07	6,87	11,68	2	2	2	3
v9	consumi	24,19	0	0	1	0	0	3
v10	altre spese	8,41	0	2,77	0	0	0	1
v11	Minusvalenze patrimoniali	98,14	5,95	10,62	0	0	0	2
v12	ammortamenti	23,98	0	0	1	0	0	3
v13	ammortamenti - di cui per beni strumentali	44,68	0	0	0	0	0	3
v14	altre componenti negative	51,76	0	0	2	0	0	1
v15	reddito	0,24	0	0	0	0	0	3
v16	valore dei beni strumentali mobili	19,30	0	0	0	0	0	3
v17	di cui valore relativo a beni acquisiti in contratti di locazione finanziaria e non finanziaria	95,16	5,10	10,04	0	2	3	3
V18	volume di affari	19,61	0	0	2	0	0	3
V19	altre operazioni fuori campo; operazioni non soggette a dichiarazione	97,79	10,66	14,61	0	0	0	1
v20	i.v.a. sulle operazioni imponibili	23,09	0	0	2	0	0	2
v21	altra i.v.a.	96,12	9,17	13,29	0	0	0	2
Mean		58,49	3,09	5,57	0,62	0,38	0,52	2,33

Tabella b - *Dataset* degli SDS trasformato con la metrica del *double ranking*

Id.	Contenuto	rcnv	rcnvm	rcnvp	ro1	ro1m	ro1p	ro2	ro2m	ro2p	rcacnv	rcacnvm
v1	compensi dichiarati	2	1	19	5,5	4,5	15,5	9	8	12	7,5	6,5
v2	adeguamento da studi di settore	14	13	7	13	12	8	14	13	7	7,5	6,5
v3	altri proventi lordi	13	12	8	19	18	2	19	18	2	7,5	6,5
v4	plusvalenze patrimoniali	16	15	5	12	11	9	13	12	8	7,5	6,5
v5	spese per prestazioni di lavoro dipendente	12	11	9	15	14	6	15	14	6	21	20
v6	spese per lavoro dipendente di cui per personale con contratto di somministrazione di lavoro	21	20	0	14	13	7	12	11	9	7,5	6,5
v7	spese per prestazioni di collaborazione coordinata e continuativa	20	19	1	11	10	10	11	10	10	7,5	6,5
v8	compensi corrisposti a terzi per prestazioni direttamente afferenti l'attività professionale e artistica	11	10	10	18	17	3	18	17	3	18,5	17,5
v9	consumi	8	7	13	5,5	4,5	15,5	4,5	3,5	16,5	15,5	14,5
v10	altre spese	3	2	18	5,5	4,5	15,5	10	9	11	7,5	6,5
v11	minusvalenze patrimoniali	19	18	2	17	16	4	17	16	4	7,5	6,5
v12	ammortamenti	7	6	14	5,5	4,5	15,5	4,5	3,5	16,5	15,5	14,5
v13	ammortamenti - di cui per beni strumentali	9	8	12	5,5	4,5	15,5	4,5	3,5	16,5	7,5	6,5
v14	altre componenti negative	10	9	11	5,5	4,5	15,5	4,5	3,5	16,5	18,5	17,5
v15	reddito	1	0	20	5,5	4,5	15,5	4,5	3,5	16,5	7,5	6,5
v16	valore dei beni strumentali mobili	4	3	17	5,5	4,5	15,5	4,5	3,5	16,5	7,5	6,5
v17	di cui valore relativo a beni acquisiti in contratti di locazione finanziaria e non finanziaria	15	14	6	16	15	5	16	15	5	7,5	6,5
v18	volume di affari	5	4	16	5,5	4,5	15,5	4,5	3,5	16,5	18,5	17,5
v19	altre operazioni fuori campo; operazioni non soggette a dichiarazione	18	17	3	21	20	0	21	20	0	7,5	6,5
v20	i.v.a. sulle operazioni imponibili	6	5	15	5,5	4,5	15,5	4,5	3,5	16,5	18,5	17,5
v21	altra i.v.a.	17	16	4	20	19	1	20	19	1	7,5	6,5
	Mean	11	10	10	11	10	10	11	10	10	11	10

Legenda: è basata sui dati della precedente tavola a trattati come indicato nelle note 8, 10 e 12, qui per comodità riassunte:

- indicato con j un generico vettore colonna (indicatore) tratto dalla precedente tavola a, e con rj il vettore colonna formato con i ranghi del vettore j, il procedimento di *double ranking* sostituisce rj con altri due vettori: un primo, dato da rjm=rj-1, indicante la distanza (numero posti) di ogni elemento rispetto al rango minimo (rango migliore, da cui il suffisso finale "m") che è pari a 1; un secondo dato da rjp=N-rj, rappresentante la distanza di ogni elemento rispetto al rango massimo (rango peggiore, da cui il suffisso finale "p") indicato con N;
- nella tavola in esame il rango minimo varrà evidentemente 1 e il rango massimo varrà 21 essendo 21 le osservazioni (righe) dai dati sottoposti all'AC;
- in caso di *ties* (ranghi appaiati) si è applicato il valore medio.

Tabella b segue - *Dataset* degli SDS trasformato con la metrica del *double ranking*

Id.	Contenuto	rcacnvp	rcao1	rcao1m	rcao1p	rcao2	rcao2m	rcao2p	rschema	rschemam	rschemap
v1	compensi dichiarati	13,5	8,5	7,5	12,5	8,5	7,5	12,5	15,5	14,5	5,5
v2	adeguamento da studi di settore	13,5	17,5	16,5	3,5	18,5	17,5	2,5	3	2	18
v3	altri proventi lordi	13,5	8,5	7,5	12,5	8,5	7,5	12,5	3	2	18
v4	plusvalenze patrimoniali	13,5	8,5	7,5	12,5	8,5	7,5	12,5	7,5	6,5	13,5
v5	spese per prestazioni di lavoro dipendente	0	20	19	1	18,5	17,5	2,5	15,5	14,5	5,5
v6	spese per lavoro dipendente di cui per personale con contratto di somministrazione di lavoro	13,5	8,5	7,5	12,5	8,5	7,5	12,5	15,5	14,5	5,5
v7	spese per prestazioni di collaborazione coordinata e continuativa	13,5	17,5	16,5	3,5	18,5	17,5	2,5	15,5	14,5	5,5
v8	compensi corrisposti a terzi per prestazioni direttamente afferenti l'attività professionale e artistica	2,5	20	19	1	18,5	17,5	2,5	15,5	14,5	5,5
v9	consumi	5,5	8,5	7,5	12,5	8,5	7,5	12,5	15,5	14,5	5,5
v10	altre spese	13,5	8,5	7,5	12,5	8,5	7,5	12,5	3	2	18
v11	minusvalenze patrimoniali	13,5	8,5	7,5	12,5	8,5	7,5	12,5	7,5	6,5	13,5
v12	ammortamenti	5,5	8,5	7,5	12,5	8,5	7,5	12,5	15,5	14,5	5,5
v13	ammortamenti - di cui per beni strumentali	13,5	8,5	7,5	12,5	8,5	7,5	12,5	15,5	14,5	5,5
v14	altre componenti negative	2,5	8,5	7,5	12,5	8,5	7,5	12,5	3	2	18
v15	reddito	13,5	8,5	7,5	12,5	8,5	7,5	12,5	15,5	14,5	5,5
v16	valore dei beni strumentali mobili	13,5	8,5	7,5	12,5	8,5	7,5	12,5	15,5	14,5	5,5
v17	di cui valore relativo a beni acquisiti in contratti di locazione finanziaria e non finanziaria	13,5	20	19	1	21	20	0	15,5	14,5	5,5
v18	volume di affari	2,5	8,5	7,5	12,5	8,5	7,5	12,5	15,5	14,5	5,5
v19	altre operazioni fuori campo; operazioni non soggette a dichiarazione	13,5	8,5	7,5	12,5	8,5	7,5	12,5	3	2	18
v20	i.v.a. sulle operazioni imponibili	2,5	8,5	7,5	12,5	8,5	7,5	12,5	7,5	6,5	13,5
v21	altra i.v.a.	13,5	8,5	7,5	12,5	8,5	7,5	12,5	7,5	6,5	13,5
	Mean	10	11	10	10	11	10	10	11	10	10

Legenda: è basata sui dati della precedente tavola a trattati come indicato nelle note 8, 10 e 12, qui per comodità riassunte:

- indicato con j un generico vettore colonna (indicatore) tratto dalla precedente tavola a, e con rj il vettore colonna formato con ranghi del vettore j, il procedimento di *double ranking* sostituisce rj con altri due vettori: un primo, dato da rjm=rj-1, indicante la distanza (numero posti) di ogni elemento rispetto al rango minimo (rango migliore, da cui il suffisso finale "m") che è pari a 1; un secondo dato da rjp=N-rj, rappresentante la distanza di ogni elemento rispetto al rango massimo (rango peggiore, da cui il suffisso finale "p") indicato con N;
- nella tavola in esame il rango minimo varrà evidentemente 1 e il rango massimo varrà 21 essendo 21 le osservazioni (righe) dai dati sottoposti all'AC;
- in caso di *ties* (ranghi appaiati) si è applicato il valore medio.

Tabella c - correlazioni per riga delle variabili trasformate– dati SDS

(obs=7)

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16	v17	v18	v19	v20	v21
v1	1.0000																				
v2	-0.5787	1.0000																			
v3	-0.4745	0.4187	1.0000																		
v4	-0.4391	0.2461	0.7133	1.0000																	
v5	-0.4403	0.2792	-0.1826	-0.5376	1.0000																
v6	-0.4029	-0.1131	0.2203	0.6280	-0.1414	1.0000															
v7	-0.4530	0.4982	-0.1884	0.0545	0.4392	0.4827	1.0000														
v8	-0.4463	0.3533	0.0324	-0.4307	0.9587	-0.0876	0.4230	1.0000													
v9	0.5943	-0.7761	-0.8197	-0.5530	-0.1366	-0.2493	-0.3405	-0.3480	1.0000												
v10	0.5416	0.0171	0.1030	0.0473	-0.6326	-0.6440	-0.6620	-0.6221	0.1367	1.0000											
v11	-0.6486	0.3212	0.8430	0.8824	-0.1606	0.6901	0.1749	0.0032	-0.7539	-0.2816	1.0000										
v12	0.6392	-0.7746	-0.8105	-0.5905	-0.1381	-0.3106	-0.3840	-0.3392	0.9967	0.1828	-0.7861	1.0000									
v13	0.7615	-0.6161	-0.6793	-0.2161	-0.5642	0.0146	-0.1277	-0.6744	0.7471	0.2776	-0.5449	0.7436	1.0000								
v14	0.0138	-0.1526	-0.2238	-0.0787	-0.2094	-0.4628	-0.5081	-0.4141	0.5148	0.4877	-0.3481	0.5078	0.1563	1.0000							
v15	0.9548	-0.5765	-0.5807	-0.4929	-0.4535	-0.4248	-0.4308	-0.5040	0.6987	0.5274	-0.7372	0.7382	0.8458	0.1546	1.0000						
v16	0.9260	-0.6091	-0.6324	-0.4213	-0.5047	-0.2988	-0.3486	-0.5761	0.7384	0.4656	-0.7012	0.7655	0.9246	0.1605	0.9853	1.0000					
v17	-0.4182	0.6258	0.0674	-0.1169	0.6501	0.2162	0.8298	0.7453	-0.5931	-0.5986	0.1865	-0.5984	-0.4504	-0.7237	-0.4664	-0.4775	1.0000				
v18	0.5832	-0.7394	-0.7555	-0.6781	0.0000	-0.4372	-0.4643	-0.1965	0.9633	0.1836	-0.8088	0.9750	0.5874	0.5512	0.6707	0.6673	-0.5852	1.0000			
v19	-0.6415	0.4354	0.9623	0.7828	-0.0806	0.4316	0.0159	0.1125	-0.8457	-0.1294	0.9463	-0.8565	-0.7170	-0.2908	-0.7425	-0.7601	0.1847	-0.8199	1.0000		
v20	0.3474	-0.4202	-0.4622	-0.4115	-0.1537	-0.6162	-0.6272	-0.3593	0.7616	0.5021	-0.6459	0.7739	0.3599	0.9081	0.4720	0.4524	-0.7589	0.8282	-0.5732	1.0000	
v21	-0.5797	0.3174	0.9165	0.7491	-0.0536	0.5502	0.0694	0.1579	-0.8156	-0.2585	0.9572	-0.8281	-0.6569	-0.4531	-0.7033	-0.7128	0.2458	-0.8053	0.9752	-0.6811	1.0000

Legenda: la matrice riporta le correlazioni di ordine zero calcolate lungo le righe della tabella b, cioè lungo le variabili v1, v2,...,v21, e con le colonne rcnvm, rcnvp, ro1m, ro1p,ro2m, ro2p, rcacnvm, rcacnvp, rcao1m, rcao1p, rcao2m, rcao2p, schemam, rschemap.

Tabella d - Variazioni percentuali coordinate punti asymmetric plot calcolate sui primi due e sui primi tre assi - dati SDS

	rcnvm	rcnvp	ro1m	ro1p	ro2m	ro2p	rcacnvm	rcacnvp	rcao1m	rcao1p	rcao2m	rcao2p	rschemam	rschemap	media
v1	2,3	23,0	1,7	0,0	2,3	0,2	130,8	77,3	2,1	0,0	0,9	0,3	43,4	50,4	20,6
v2	18,5	3,1	0,2	1,2	0,6	1,7	85,0	121,7	0,2	1,5	0,1	0,5	49,4	45,0	20,3
v3	20,0	2,7	0,1	1,1	0,6	1,5	55,6	237,4	0,1	2,2	0,0	0,7	34,4	75,9	20,8
v4	7,8	7,0	1,5	0,2	2,4	0,6	80,4	161,4	1,1	0,5	0,2	0,0	34,9	72,4	20,9
v5	15,9	1,8	0,1	2,5	0,0	3,4	147,9	73,3	0,0	1,9	1,6	0,9	95,9	23,5	20,4
v6	2,5	12,5	6,6	0,2	8,2	0,0	107,1	107,9	4,6	0,1	2,5	1,0	30,2	74,5	22,0
v7	3,1	10,9	5,6	0,1	6,7	0,0	151,8	65,4	7,0	0,0	3,8	0,5	45,3	48,5	21,1
v8	17,1	2,1	0,0	1,8	0,1	2,4	124,1	75,0	0,0	1,5	0,8	0,6	79,7	26,2	19,6
v9	4,5	14,9	0,4	1,3	0,8	2,3	138,6	77,8	0,7	0,8	0,1	0,1	59,6	38,3	20,2
v10	8,1	6,0	0,0	2,8	0,2	3,7	74,5	140,4	0,1	3,6	0,1	1,3	44,1	51,9	20,4
v11	12,1	5,2	2,2	0,2	3,5	0,4	69,0	179,4	1,1	0,4	0,2	0,0	31,2	78,4	20,5
v12	4,5	15,1	0,4	1,5	0,7	2,6	138,1	77,1	0,6	0,9	0,1	0,1	60,2	37,5	20,1
v13	1,6	25,3	2,5	0,2	3,3	0,0	136,9	75,0	3,0	0,0	1,5	0,8	40,0	54,6	21,2
v14	14,0	1,3	0,6	10,0	0,3	11,5	74,3	133,2	0,4	8,4	1,5	5,3	64,1	32,0	22,6
v15	1,9	29,2	1,5	0,0	2,1	0,2	133,3	69,2	2,1	0,0	0,9	0,3	43,9	47,1	20,2
v16	1,8	27,5	1,9	0,0	2,5	0,1	135,1	71,3	2,4	0,0	1,1	0,5	42,5	49,8	20,6
v17	5,8	7,5	4,5	0,0	5,6	0,0	133,1	63,5	6,0	0,0	2,7	0,1	48,1	42,1	19,6
v18	5,9	11,4	0,0	4,0	0,2	5,7	137,3	77,0	0,1	1,8	0,0	0,6	70,0	31,5	20,2
v19	26,0	2,5	0,4	0,7	1,3	1,1	50,8	228,8	0,2	1,6	0,0	0,5	30,7	79,4	20,0
v20	10,5	3,3	0,2	9,2	0,0	11,0	92,2	106,4	0,1	5,7	0,9	3,3	67,5	30,3	21,5
v21	16,1	4,3	1,9	0,3	3,3	0,6	63,4	184,6	0,8	0,6	0,1	0,0	30,9	77,2	20,1
media	8,3	8,7	1,4	1,5	1,9	2,0	101,0	102,2	1,4	1,4	0,8	0,8	47,6	47,9	20,6

Legenda: le celle evidenziate in grigio indicano le combinazioni di variabili e indicatori che hanno una distanza euclidea poco variata, sia che venga calcolata sui primi due assi sia sui primi tre assi. Gli elementi (i,j) della matrice sono stati calcolati con 100 x [d3(vari, indj)- d2(vari, indj)] / d2(vari, indj), ove d3(vari, indj) indica la distanza euclidea tra la variabile i-esima e 'indicatore j-esimo nello spazio tridimensionale (grafico 3) e d2(vari, indj) l'analoga nel piano bidimensionale (grafico 5)

Tavola e - Dataset CBS ricodificato

id	F	R	S	C	P
pa	0	0	0	2	0
sfr	0	0	0	0	1
cwi	0	0	0	4	0
err	0	2	0	0	1
ghe	0	0	0	0	1
gse	3	0	1	2	1
Mean	0,500	0,333	0,167	1,333	0,667

Tavola f - Dataset di CBS trasformato con la metrica del double ranking

identificativo	rF	rFm	rFp	rR	rRm	rRp	rS	rSm	rSp	rC	rCm	rCp	rP	rPm	rPp
pa	3	2	3	3	2	3	3	2	3	4,5	3,5	1,5	1,5	0,5	4,5
sfr	3	2	3	3	2	3	3	2	3	2	1	4	4,5	3,5	1,5
cwi	3	2	3	3	2	3	3	2	3	6	5	0	1,5	0,5	4,5
err	3	2	3	6	5	0	3	2	3	2	1	4	4,5	3,5	1,5
ghe	3	2	3	3	2	3	3	2	3	2	1	4	4,5	3,5	1,5
gse	6	5	0	3	2	3	6	5	0	4,5	3,5	1,5	4,5	3,5	1,5
Sum	21.00	15	15	21.00	15	15	21.00	15	15	21.00	15	15	21.00	15	15

Riferimenti bibliografici

Blue-Ets. 2010. *Subject Area, Working Project n.4*. http://www.blue-ets.istat.it/index.php?id=40.

Daas P.J.H., Ossen S.J.L., Arends T.J.. 2009. *Framework of Quality Assurance for Administrative Data Sources*, http://www.pietdaas.nl/beta/pubs/pubs/ISI-2009_paper.pdf

Greenacre M. 2007. *Correspondence Analysis in Practise.* Chapman & Hall.

Stata. 2009. *Multivariate Statistics*. StataCorp.

UE. 2010. *Trattato di Lisbona*. http://europa.eu/lisbon_treaty/take/index_it.htm.

Wallgren, A., Wallgren, B.. 2007. *Register-based Statistics: Administrative Data for Statistical Purposes*. Series in Survey Methodology, John Wiley & Sons, Ltd, Chichester, England.

Informazioni per gli autori

La collana è aperta ad autori dell'Istat e del Sistema statistico nazionale, e ad altri studiosi che abbiano partecipato ad attività promosse dal Sistan (convegni, seminari, gruppi di lavoro, ecc.). Da gennaio 2011 essa sostituirà Documenti Istat e Contributi Istat.

Coloro che desiderano pubblicare sulla nuova collana dovranno sottoporre il proprio contributo alla redazione degli Istat Working Papers inviandolo per posta elettronica all'indirizzo iwp@istat.it. Il saggio deve essere redatto seguendo gli standard editoriali previsti, corredato di un sommario in italiano e in inglese; deve, altresì, essere accompagnato da una dichiarazione di paternità dell'opera. Per la stesura del testo occorre seguire le indicazioni presenti nel foglio di stile, con le citazioni e i riferimenti bibliografici redatti secondo il protocollo internazionale 'Autore-Data' del *Chicago Manual of Style*.

Per gli autori Istat, la sottomissione dei lavori deve essere accompagnata da una mail del proprio dirigente di Servizio/Struttura, che ne assicura la presa visione. Per gli autori degli altri enti del Sistan la trasmissione avviene attraverso il responsabile dell'ufficio di statistica, che ne prende visione. Per tutti gli altri autori, esterni all'Istat e al Sistan, non è necessaria alcuna presa visione. Tutti i lavori saranno sottoposti al Comitato di redazione, che valuterà la significatività del lavoro per il progresso dell'attività statistica istituzionale. La pubblicazione sarà disponibile su formato digitale e sarà consultabile on line.

Gli articoli pubblicati impegnano esclusivamente gli autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat. Si autorizza la riproduzione a fini non commerciali e con citazione della fonte.