# Metodi di Forward Search per la ricerca di outlier: un'applicazione ai dati Istat sui matrimoni nel 2011[1]

*Simona Toti,[2] Romina Filippini,[3] Francesco Amato,[4] Claudia Iaccarino [5]*

## Sommario

*La Forward Search (FS) è un metodo iterativo capace di individuare gruppi di dati anomali, nel caso di dati di regressione o multivariati. Il presente lavoro applica l'approccio FS ai dati Istat relativi ai matrimoni celebrati nel 2011 e alle caratteristiche degli sposi. Lo scopo è l'identificazione di eventuali Province o Comuni con caratteristiche anomale rispetto al resto dell'Italia. I modelli rispetto ai quali si è condotta l'analisi FS sono: il modello multivariato, il modello composizionale e il modello di regressione. Dal confronto dei risultati emerge la forte dipendenza della definizione di dato anomalo dal contesto in studio. L'analisi è stata effettuata con l'ausilio del software SiRiO (Sistema Ricerca Outlier), sviluppato in ISTAT per rendere fruibile il sistema d'analisi anche all'utilizzatore meno esperto.*

**Parole chiave**: outlier, Forward Search, indagine Istat sui matrimoni, SiRiO.

## Abstract

*Forward Search (FS) is an iterative method to detect the presence of groups of outliers in the case of regression or multivariate data. In the present paper, FS is applied to Istat data on marriages celebrated in 2011 and to the characteristics of partners. The aim is to identify Provinces and Municipalities with significantly different characteristics from the other Italian Provinces and Municipalities. The analysis was conducted referring to three FS models: the multivariate, the compositional and the regression model. The comparison of the results obtained highlights the importance of the reference framework to define an observation as outlier. The analysis was performed using SiRiO (Outlier Research System), a software developed in Istat to make the FS analysis available for non- expert users.*

**Keywords:** outlier, Forward Search, Istat marriage data, SiRiO.

[1] Gli autori ringraziano Alessandra Reale e Sabrina Prati per l'incoraggiamento e il supporto costante. Quanto pubblicato impegna esclusivamente gli Autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

[2] Istat, e-mail: toti@istat.it.

[3] Istat, e-mail: filippini@istat.it.

[4] Istat, e-mail: framato@istat.it.

[5] Istat, e-mail: iaccarin@istat.it.

## 1. Introduzione

L'identificazione dei dati anomali è un problema importante nell'analisi statistica dei dati: la presenza di osservazioni atipiche può infatti dare origine a risultati distorti (Becker e Gather, 1999). Con dati anomali o outlier si intendono le osservazioni che si trovano "lontane" dalla maggior parte delle altre osservazioni e che seguono un modello diverso da quello assunto per il fenomeno analizzato.

La Forward Search (FS) è un metodo statistico capace di individuare gruppi di dati anomali, anche nel caso di dati multivariati, e di ricercare strutture di eterogeneità nei dati (Atkinson *et al.*, 2004).

Nell'ambito dell'attività istituzionale dell'unità Istat di Controllo e Correzione dei dati censuari (MTC/D) è stata sviluppata una competenza metodologica e il software SiRiO per affrontare il problema dell'identificazione dei dati anomali tramite FS, in tre casi differenti: un modello regressivo (confronto fra archivi); un modello gaussiano multivariato; un modello log-normale composizionale. Mentre le prime due specifiche nascono dalla collaborazione con il gruppo dell'Università di Parma, l'unità MTC/D ha esteso il metodo al caso di dati composizionali ed ha sviluppato l'applicazione web per rendere fruibile il sistema d'analisi in modo semplice ed intuitivo.

Il presente lavoro nasce dalla collaborazione dell'unità MTC/D con l'unità Strutture familiari e ciclo di vita (DEM/C) e riporta i risultati dell'analisi dei dati relativi alle indagini Istat sui matrimoni celebrati nel 2011.

## 2. L'approccio Forward Search per la ricerca dei dati anomali

La FS è un metodo iterativo il cui principale risultato è quello di individuare, all'interno di un gruppo di osservazioni, il sottoinsieme privo di dati anomali a partire dal quale ottenere stime robuste dei parametri di interesse, ad un prefissato livello di confidenza.

Il primo passo della procedura consiste nell'individuazione di un sottoinsieme di unità con valori vicini al centro della distribuzione dei dati, su cui verranno calcolate opportune statistiche. Quindi, nei passi successivi, il sottoinsieme viene incrementato di una unità alla volta, in modo tale che il sottoinsieme sia sempre composto dalle osservazioni più vicine al centro. Su ogni nuovo sottoinsieme di osservazioni vengono ripetute le stesse analisi statistiche. La procedura garantisce che i dati anomali entreranno a far parte del sottoinsieme soltanto agli ultimi passi.

Ad ogni passo, una buona sintesi dell'andamento della procedura è dato dalla più piccola distanza dal centro della distribuzione delle unità non appartenenti al sottoinsieme di volta in volta analizzato: tale misura è detta "segnale" della procedura stessa. Riportando sull'asse delle ascisse i passi dell'algoritmo (ossia la dimensione del sottoinsieme) e sull'asse delle ordinate il segnale, il grafico della spezzata che unisce i punti mostrerà un picco in corrispondenza del passo immediatamente precedente all'inclusione del primo dato anomalo, se presente.

La conoscenza, anche approssimata, della distribuzione di probabilità della minima distanza utilizzata, sotto l'ipotesi nulla di normalità dei dati, permette la costruzione di bande ad un prefissato livello di confidenza per il segnale.

## 2.1 Il modello multivariato

Il modello gaussiano multivariato assume che ogni osservazione campionaria sia la realizzazione di una stessa distribuzione normale multivariata con vettore delle medie μ e matrice di varianza e covarianza Σ, ed è rispetto a tale modello che si definiscono i dati eventualmente anomali.

Per valutare la distanza delle osservazioni dal vettore delle medie, centro della distribuzione, ad ogni passo, si ricorre alla distanza di Mahalanobis:

$$d_i = \sqrt{(x_i - \bar{x}_m) S_m^{-1} (x_i - \bar{x}_m)} \tag{2.1}$$

dove $\bar{x}_m$ è il vettore delle medie campionarie e $S_m$ è la matrice di varianza e covarianza campionaria. Tali parametri vengono stimati con le sole osservazioni presenti nel sottoinsieme, a quella iterazione.

È da notare che, mentre la stima della media non è influenzata dal troncamento del campione, la matrice di varianza e covarianza campionaria sottostima la matrice di varianza e covarianza della popolazione. Per correggere tale distorsione la FS prevede l'uso del fattore di espansione di Tallis (1963).

Sotto l'ipotesi nulla di normalità delle variabili, è possibile costruire delle bande di confidenza per la minima distanza di Mahalanobis: fin quando la distanza delle osservazioni è stimata sulla base di un campione coerente con l'ipotesi nulla, il segnale resta entro la banda; se, negli ultimi passi della procedura, il segnale esce dalle bande, questo indica l'entrata nel sottoinsieme di stima di un'unità non omogenea, dunque di un'unità anomala.

L'utilizzo nella FS degli stimatori di massima verosimiglianza (media e matrice di varianza e covarianza campionarie), sensibili alla presenza di osservazioni anomale, rende subito evidente l'introduzione del primo outlier segnalato dalla presenza di picchi nella traiettoria del segnale.

La distribuzione della minima distanza di Mahalanobis non è nota in letteratura. Tuttavia, è possibile ottenere bande di confidenza molto accurate per tale statistica facendo riferimento alla teoria delle statistiche d'ordine sotto l'ipotesi di normalità delle osservazioni. Le bande di confidenza così ottenute sono basate sulla distribuzione F di Fisher (Riani *et al.*, 2009).

Poiché non è sempre possibile fare l'assunzione di normalità dei dati, la procedura della FS introduce un passaggio propedeutico all'analisi vera e propria, nel quale si utilizza la trasformata di Box-Cox in associazione alla FS, per riportare i dati sotto l'ipotesi di normalità.

## 2.2 Il modello composizionale

L'analisi di dati composizionali trova applicazione ogni volta che l'oggetto di interesse è un vettore v-dimensionale le cui componenti siano variabili reali positive, che possono essere viste come porzioni di un totale fissato. Se dunque l'interesse è nell'ampiezza di una componente relativamente alle altre, allora confrontare differenti composizioni si traduce nel confrontare ogni valore con tutti gli altri tramite rapporti (Aitchison, 1986). Esistono diverse funzioni che applicate ai dati composizionali v-dimensionali vincolati a una somma fissata, restituiscono vettori indipendenti in $R^{v-1}$. Tra queste la trasformata ILR (Isometric Log Ratio) garantisce una perfetta isometria tra lo spazio composizionale e quello reale

(Egozcue *et al.*, 2003). Inoltre se si assume per i vettori di dati osservati una distribuzione log-normale multivariata, la trasformata ILR restituisce vettori a v-1 componenti indipendenti nello spazio reale con distribuzione multinormale (Aitchison e Shen, 1980).

Per i dati trasformati, il modello di riferimento sarà dunque una distribuzione normale multivariata. In questo modo il problema composizionale diventa un problema multinormale ed è quindi possibile adottare la stessa procedura d'analisi descritta nel paragrafo precedente. La procedura FS per dati composizionali prevede dunque:

1. un passo iniziale di applicazione della trasformata ILR ai dati;
2. l'applicazione della procedura FS multivariata gaussiana.

## 2.3 Il modello regressivo

Un problema classico dell'analisi di dati amministrativi è quello del confronto tra fonti: avendo a disposizione per una data variabile continua, rilevata su una certa popolazione, una fonte attendibile, cioè con valori non affetti da errore, la si vuole utilizzare per valutare l'attendibilità di un'altra fonte, potenzialmente affetta da errori. Una possibile rappresentazione della relazione tra le due fonti, è data dal modello regressivo a intercetta nulla e coefficiente di regressione unitario. I punti "distanti" da tale retta esprimono una discrepanza tra le due fonti, indicativi di una potenziale situazione di errore nella fonte scelta come variabile dipendente.

Per valutare la distanza delle osservazioni dal modello di regressione lineare semplice si utilizzano i residui studentizzati[6]:

$$r_i = \frac{e_i}{\sqrt{S_m{}^2(1-h_i)}} \tag{2.2}$$

dove $e_i$ è il residuo del modello di regressione, $S_m$ la matrice di varianza e covarianza campionaria, $h_i$ i valori di leva ossia gli elementi della diagonale principale della matrice cappello:

$$H = X(X_m{}^T X_m)^{-1} X^T \ . \tag{2.3}$$

L'utilizzo nella FS della tecnica, poco robusta, dei minimi quadrati (stimatori OLS, Ordinary Least Squares) per la stima dei parametri del modello, diventa un punto di forza nella metodologia. Ad ogni passo, il più piccolo residuo delle unità non appartenenti al sottoinsieme di stima del modello è il segnale della procedura. L'introduzione di dati anomali determina forti picchi nella spezzata che monitora il segnale.

È possibile costruire delle bande di confidenza del minimo residuo per identificare le osservazioni omogenee rispetto al modello e quelle che, avendo associata una distanza talmente elevata da risultare esterna alla banda, risultano significativamente lontane dal modello, dunque relative a osservazioni anomale. La distribuzione del minimo residuo non è nota in letteratura ma una sua approssimazione è stata derivata sulla base della distribuzione delle statistiche d'ordine sotto l'ipotesi di normalità dei dati (Atkinson e Riani, 2006).

[6] Si parla in effetti di residui studentizzati di cancellazione per evidenziare il fatto che le quantità a numeratore ed a denominatore sono calcolate su sottoinsiemi diversi del campione (Atkinson *et al.*, 2004).

## 3. SiRiO: Sistema Ricerca Outlier

Il pacchetto Matlab FSDA, composto da funzioni per l'analisi FS di regressione e multinormale, è disponibile all'indirizzo http://www.riani.it/MATLAB.htm. Nell'ambito della collaborazione tra Istat (Unità di Controllo e Correzione dei Censimenti) e Università di Parma sono state sviluppate procedure *ad hoc* per il caso del confronto fra fonti e per il modello multinormale. Per il modello composizionale è stata sviluppata all'interno dell'Istat una procedura Python che applica la trasformata ILR ai dati (Palombi *el al.*, 2011) il cui output viene analizzato dalla procedura multinormale.

L'applicazione web SiRiO, sviluppata in Java, si interfaccia con le tre procedure in modo da renderle fruibili senza necessità di specifiche competenze Matlab e Python. Ogni utente può creare diversi progetti; ogni progetto è caratterizzato da uno o più insiemi di dati di partenza e da uno specifico tipo di analisi. Per ogni insieme di dati analizzati sono resi disponibili grafici e tabelle, scaricabili come file in formato zip. Attualmente l'applicazione è disponibile sulla rete interna Istat[7].

## 4. L'applicazione della FS ai dati Istat sui Matrimoni 2011

La presente analisi utilizza l'applicativo SiRiO per la ricerca degli outlier tramite il metodo FS, con l'obiettivo di individuare la presenza di eventuali comportamenti anomali, a livello provinciale e comunale, relativi all'evento matrimonio e alle caratteristiche degli sposi.

In particolare, è stata analizzata a livello provinciale la condizione professionale degli sposi e delle spose separatamente (fonte: Rilevazione dei Matrimoni). Lo scopo di questa prima analisi, è quello di mettere a confronto la distribuzione del numero di matrimoni per condizione professionale (Occupato, In cerca di occupazione, Inattivo), separatamente per sposo e sposa, individuando le Province che hanno un profilo che si discosta da quello dalle altre. Si è dunque utilizzato il modello multivariato classico e quello composizionale della FS per la ricerca dei dati anomali.

La seconda analisi ha riguardato il confronto tra i dati ricavati dai modelli individuali (fonte: Rilevazione dei Matrimoni) e quelli ricavati dai modelli riepilogativi (fonte: Rilevazione mensile degli eventi demografici di Stato Civile). Lo scopo è di individuare i Comuni che hanno scostamenti tra i valori riportati nelle due fonti, significativamente diversi da quelli degli altri Comuni della Provincia. In questo caso, si è utilizzato il modello di regressione della FS, assumendo come variabile dipendente il numero di modelli individuali e privi di errori i valori riportati nei modelli riepilogativi.

L'analisi dei risultati, oltre a individuare le Province con informazione statistica e distribuzione della nuzialità, anomale, permette di apprezzare le differenze dei vari modelli FS.

[7] Gli utenti Istat interessati all'utilizzo di SiRiO possono richiedere le credenziali di accesso inviando una mail a framato@istat.it.

## 5. Le fonti

Le informazioni sui matrimoni celebrati in Italia vengono rilevate da due indagini Istat: la Rilevazione dei Matrimoni (Modello Istat D.3) e la Rilevazione mensile degli eventi demografici di Stato Civile (Modello Istat D7.a).

La *Rilevazione dei Matrimoni* (D.3) è di fonte Stato Civile e fa quindi riferimento alla popolazione presente. È un'indagine individuale ed esaustiva che ha per oggetto tutti i matrimoni celebrati in Italia e che consente di analizzare il fenomeno della nuzialità con riguardo alle principali caratteristiche del matrimonio e degli sposi.

Le informazioni raccolte riguardano sia le notizie sul matrimonio, quali la data, il rito di celebrazione (religioso o civile), il comune di celebrazione e il regime patrimoniale scelto dagli sposi (comunione o separazione dei beni), sia le informazioni relative a ciascuno degli sposi, quali la data e il comune di nascita, il comune di residenza al momento del matrimonio, il luogo di residenza futura degli sposi, lo stato civile precedente, il livello di istruzione, la condizione professionale, la posizione nella professione, il ramo di attività economica e la cittadinanza.

La *Rilevazione mensile degli eventi demografici di Stato Civile* (D7.a) raccoglie, a livello comunale, le informazioni relative alle nascite, ai matrimoni e ai decessi dichiarati presso gli uffici di Stato Civile. In questo caso, quindi, i dati raccolti non sono individuali ma riepilogativi. Nello specifico, nel caso dell'evento "matrimonio", la rilevazione fornisce per ciascun mese il numero di matrimoni celebrati in ogni singolo Comune, secondo il rito religioso o civile.

## 6. Un'analisi multivariata: la condizione professionale degli sposi 2011

Dalla *Rilevazione dei Matrimoni* (D.3) è possibile ricavare alcune informazioni demo-sociali per ciascuno degli sposi. I dati rilevati a livello comunale sono stati analizzati per Provincia. In particolare, è stata considerata la variabile Condizione professionale, separatamente per sposo e sposa, allo scopo di individuare le Province con caratteristiche anomale.

### 6.1 Descrizione dei dati

La variabile Condizione professionale, inizialmente suddivisa in nove categorie, è stata ricodificata in tre modalità: Occupato, In cerca di occupazione (che aggrega: disoccupato; in cerca di prima occupazione) e Inattivo (che aggrega: ritirato dal lavoro; casalinga, solo per le spose; studente; inabile al lavoro; in servizio di leva o servizio civile, solo per gli sposi; altro). Il numero totale di matrimoni analizzati è pari a 204.830 nelle 110 Province italiane.

I dati relativi agli sposi e alle spose sono stati analizzati separatamente, sia con modello multivariato gaussiano che composizionale. Nel seguito sono illustrati solo i risultati dell'analisi sulle spose. Per quanto riguarda gli sposi, l'analisi non ha evidenziato nessuna Provincia anomala.

## 6.2 Analisi dei risultati: modello gaussiano

Le tre variabili considerate a livello provinciale sono: il numero di spose occupate, il numero di spose in cerca di occupazione, il numero di spose inattive. La matrice di dati, analizzata dalla procedura, contiene dunque in riga le Province e in colonna i valori di ciascuna delle tre variabili.

Nella procedura implementata in SiRiO il primo passo prevede l'applicazione della procedura Box-Cox per la normalizzazione dei dati, nel caso che i dati non rispettino l'ipotesi nulla di multinormalità. Poiché la trasformata di Box-Cox risente dell'eventuale presenza di dati anomali, anche per la ricerca della migliore stima del parametro di tale funzione, si procede utilizzando l'approccio FS.

In figura 1 sono riportati i grafici forniti da SiRiO, che rappresentano i valori relativi alle 110 Province prima e dopo la trasformazione. In particolare: sulla diagonale principale sono riportati gli istogrammi delle tre componenti della variabile Condizione professionale; fuori dalla diagonale i grafici di dispersione a coppie di componenti. I valori relativi alle osservazioni anomale sono indicati sul grafico da un cerchio.

Mentre i dati non trasformati presentano una distribuzione asimmetrica a valori positivi (Figura 1.a) quelli trasformati mostrano una distribuzione approssimativamente normale, per le singole componenti (Figura 1.b). L'analisi ha individuato una sola Provincia (Ogliastra, individuata dal cerchio in figura 1.b) con frequenze relative alle tre componenti della variabile Condizione professionale, che si discostano significativamente dai valori delle altre 109 Province.

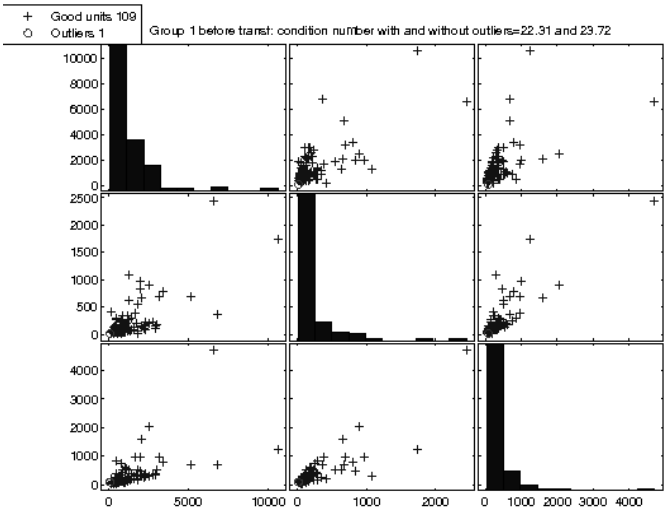
La lontananza dell'osservazione anomala dal centro della distribuzione può essere più o meno estrema a seconda della componente. Infatti, nell'analisi multivariata, ad ogni unità viene associato un valore, quello della distanza di Mahalanobis, come sintesi delle distanze delle singole componenti. Dunque può accadere che un'osservazione anomala abbia valori estremi solo su alcune componenti, mentre risulti vicina al centro della distribuzione, sulle altre.

Oltre all'output grafico, SiRiO fornisce per ogni analisi effettuata:

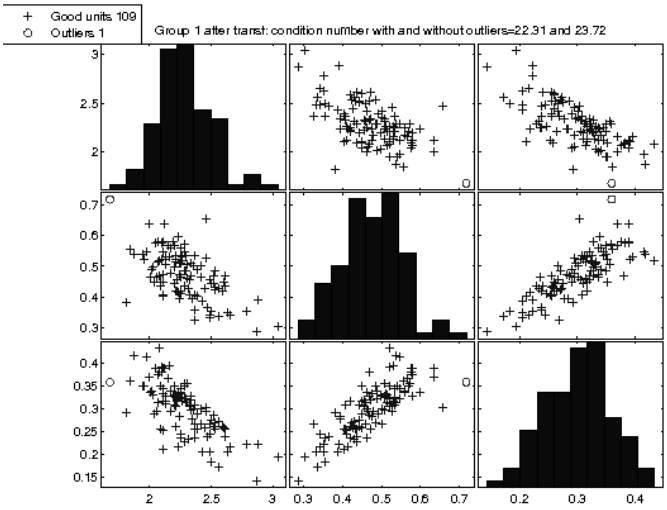
1. risultati di sintesi sul campione analizzato: numero di osservazioni del campione, numero di osservazioni anomale e valori dei parametri della trasformata Box-Cox;
2. risultati puntuali per ogni osservazione: valore iniziale, valore trasformato con Box-Cox, statistica test e relativo p-*value*.

**Figura 1 – Grafico fornito da SiRiO per l'analisi FS multivariata: distribuzione univariata e dispersione bivariata delle tre modalità della condizione professionale (Occupato, In cerca di occupazione, Inattivo) relativa alle spose. a) Dati grezzi b) Dati trasformati secondo la procedura FS Box-Cox**

a)




b)



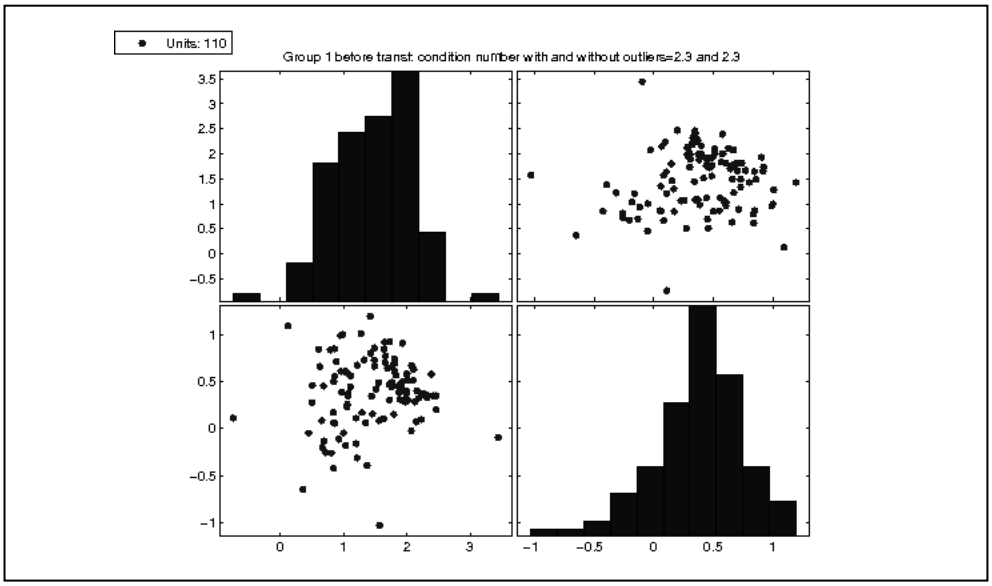

*Fonte*: Rilevazione dei Matrimoni (Modello Istat D.3). Anno 2011.

## 6.3 Analisi dei risultati: modello composizionale

Per il modello composizionale, SiRiO richiama il codice Python che opera la trasformata ILR sui dati. Le tre modalità della variabile Condizione professionale, dopo l'operazione di chiusura sulle frequenze, ossia la trasformazione delle frequenze assolute in frequenze relative, rispetto al totale dei matrimoni per Comune, vengono ridotte a due componenti ILR. A questo punto, i dati trasformati vengono analizzati dalla procedura Matlab per dati multinormali. La matrice di dati, analizzata dalla procedura, in questo caso contiene in riga le Province e in colonna i valori delle due componenti.

La figura seguente, fornita da SiRiO, riporta i valori relativi alle 110 Province dopo la trasformazione ILR.

**Figura 2 – Grafico fornito da SiRiO per l'analisi FS composizionale: distribuzione univariata e dispersione bivariata delle due componenti della variabile condizione professionale relativa alle spose dopo la trasformata ILR**




*Fonte*: Rilevazione dei Matrimoni (Modello Istat D.3). Anno 2011.

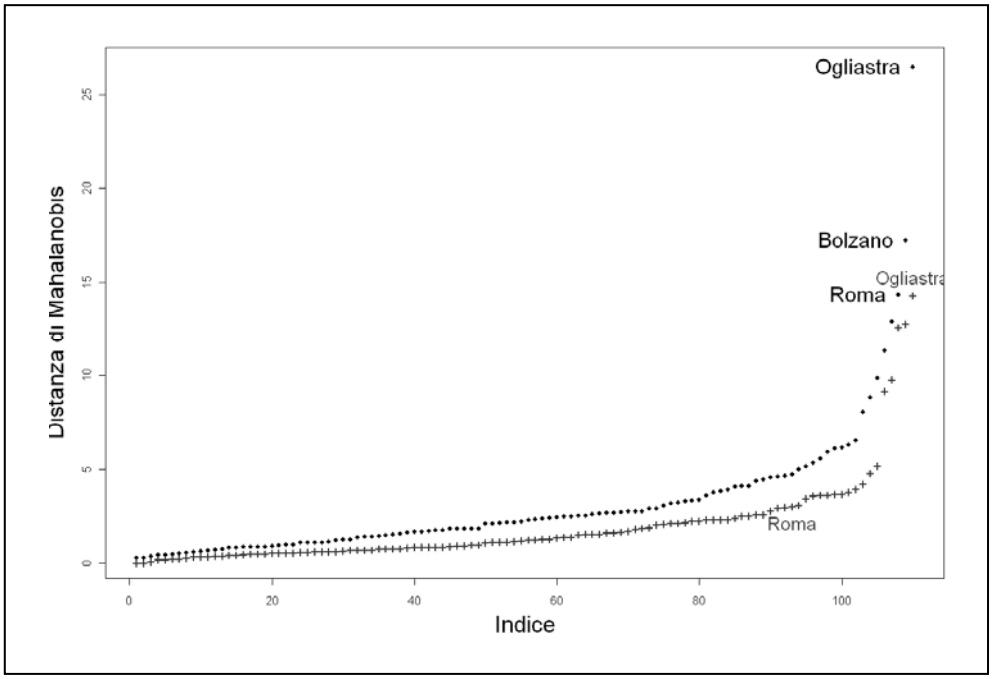
In questo caso la procedura non rileva Province anomale, in particolare la Provincia dell'Ogliastra non risulta significativamente distante dal centro della distribuzione.

E' possibile confrontare i 2 ranghi ottenuti dall'ordinamento delle unità in base alla distanza di Mahalanobis, fornita in output da SiRiO, dall'analisi multinormale e dalla composizionale. In particolare, in figura 3 sono indicate: con il punto le distanze di Mahalanobis ordinate, ottenute dall'analisi multinormale; con la croce le distanze di Mahalanobis ordinate, ottenute dall'analisi composizionale.

Si può notare come il valore relativo all'Ogliastra si allontani sensibilmente dal resto dei valori solo nell'analisi multivariata, basata sulle frequenze assolute delle tre modalità. I valori relativi all'Ogliastra sono infatti sensibilmente inferiori in valore assoluto dalla media delle restanti 109 province (Tavola 1). Se si considerano le frequenze relative o

percentuali, rispetto al totale dei matrimoni per Comune, la differenza fra il profilo dell'Ogliastra e il profilo medio è decisamente meno evidente. La distanza di Mahalanobis relativa all'Ogliastra, nell'analisi composizionale, è comunque l'ultima nell'ordine crescente ma non risulta significativamente lontana dal gruppo principale delle altre osservazioni.

**Figura 3 – Distanze ordinate di Mahalanobis della variabile Condizione professionale delle spose relativa alle 110 Province italiane, 2011, ottenute da SiRiO per l'analisi multivariata gaussiana (punto) e per quella composizionale (croce).**




*Fonte*: Rilevazione dei Matrimoni (Modello Istat D.3). Anno 2011.

**Tavola 1 – Valori assoluti e percentuali della variabile Condizione professionale per le Province di Ogliastra e Roma e media delle osservazioni**

| | Occupati | In cerca di occupazione | Inattivi | Totale |
|---|---|---|---|---|
| | | Valori assoluti | | |
| Ogliastra (outlier) | 74 | 8 | 86 | 168 |
| Roma | 10.602 | 1.732 | 1.244 | 13.578 |
| **Media (su 109 Province)** | **1.326** | **218** | **333** | **1.877** |
| | | Valori percentuali | | |
| Ogliastra | 44,0% | 4,8% | 51,2% | 100,0% |
| Roma | 78,1% | 12,8% | 9,2% | 100,0% |
| **Media (su 109 Province)** | **70,6%** | **11,6%** | **17,8%** | **100,0%** |

*Fonte:* Rilevazione dei Matrimoni (Modello Istat D.3). Anno 2011.

Nella tavola 1 sono riportati anche i valori relativi alla Provincia di Roma, che permettono di evidenziare le differenze nel significato dell'analisi composizionale rispetto a quella multinormale. Infatti, mentre i valori assoluti relativi a tale Provincia, si discostano sensibilmente (ma non significativamente) da quelli medi, i valori percentuali risultano in linea con il profilo percentuale medio italiano. Ciò è confermato dalla diversa posizione occupata dalla Provincia di Roma, nei due ordinamenti delle distanze di Mahalanobis ottenute dai due modelli. Nel caso multivariato, Roma è tra le Province con distanza di Mahalanobis più elevata (terzultimo valore, in ordine crescente) tanto da risultare lontana dalla maggioranza delle osservazioni (figura 3). Nel caso composizionale invece, il valore della distanza per questa Provincia, risulta vicino alla maggioranza delle distanze relative alle altre unità.

Mentre l'analisi composizionale permette di fare confronti fra le unità indipendentemente dall'ordine di grandezza dei valori assunti delle stesse, nell'analisi multivariata l'unità di misura gioca un ruolo fondamentale nel confronto fra gli elementi dell'insieme. È dunque importante avere chiaro l'obiettivo dell'analisi prima di procedere alla ricerca dei dati anomali.

## 7. Un confronto tra fonti: il numero di matrimoni per Comune dai modelli Istat D.3 e D7.a

I dati individuali raccolti mediante la Rilevazione dei Matrimoni (D.3) vengono sottoposti a una serie di controlli di tipo quantitativo (copertura della rilevazione) e qualitativo. È proprio nell'ambito dei controlli quantitativi che si utilizzano le informazioni ricavate dalla Rilevazione mensile degli eventi demografici di Stato Civile (D7.a): su base mensile e annuale il numero di matrimoni per Comune risultante dalla raccolta dei modelli individuali deve corrispondere a quello ricavato mediante i modelli riepilogativi.

Nella presente analisi vengono messe a confronto le due indagini precedentemente descritte per individuare i Comuni che abbiano differenze significative nel numero di matrimoni ottenuto mediante le due diverse fonti, nel 2011.

Una prima analisi condotta a livello nazionale, non ha identificato alcun Comune anomalo. Di seguito sono riportati i risultati ottenuti stratificando i Comuni per Provincia.

### 7.1 Descrizione dei dati

I dati a disposizione riguardano 7.282 Comuni divisi in 110 Province. Questa dimensione iniziale è stata ridotta in base a due criteri:

1. sono stati eliminati i Comuni con un numero di matrimoni inferiore a 10 in entrambi le fonti, poiché nelle operazioni di controllo di copertura della rilevazione, non è ritenuta rilevante una differenza tra le due fonti minore di 10;
2. si sono poi eliminate le Province con meno di 10 Comuni, per rendere la dimensione dello strato d'analisi ragionevole per una ricerca di dati anomali.

La dimensione finale dei dati analizzati è quindi di 3.457 Comuni, per 103 Province, su cui sono state rilevate le due variabili: numero di matrimoni di fonte *Rilevazione mensile degli eventi demografici di Stato Civile* (variabile indipendente) e numero di matrimoni di fonte *Rilevazione dei Matrimoni* (variabile dipendente).

## 7.2 Analisi dei risultati

L'esplorazione grafica evidenzia un andamento lognormale delle distribuzioni dei dati per Provincia, caratterizzata da una forte asimmetria e dalla positività dei valori assunti. Poiché l'analisi presuppone una distribuzione normale per la variabile dipendente (D.3), è stata applicata la trasformata logaritmica.

L'analisi ha individuato 44 Province con almeno un Comune anomalo. Nella tavola seguente è riportato l'elenco delle 44 Province, il numero dei comuni anomali ed il numero totale dei Comuni nello strato. Per questi Comuni si ipotizza una potenziale situazione di errore, che richiede opportuni approfondimenti.

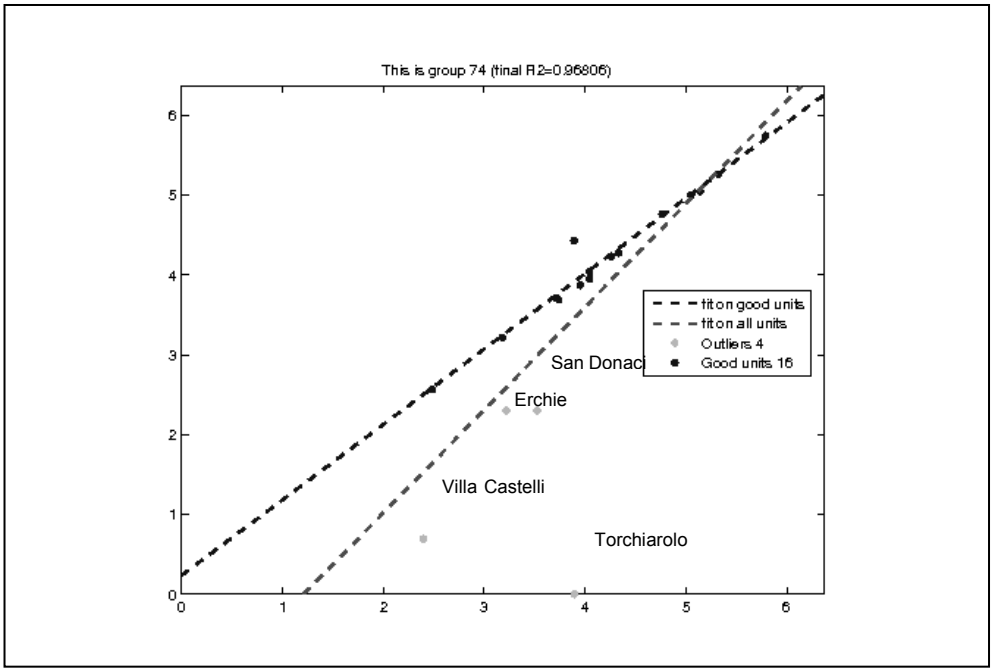
**Tavola 2 – Province con almeno un Comune anomalo**

| Provincia | n. anomali/ tot.strato | *(%)* | Provincia | n. anomali/ tot.strato | *(%)* | Provincia | n. anomali/ tot.strato | *(%)* |
|---|---|---|---|---|---|---|---|---|
| Brindisi | 4/20 | *20.0%* | Salerno | 6/141 | *4.3%* | Foggia | 1/58 | *1.7%* |
| Bari | 6/39 | *15.4%* | Gorizia | 1/24 | *4.2%* | Lodi | 1/58 | *1.7%* |
| Medio Campidano | 3/24 | *12.5%* | Monza e Brianza | 2/54 | *3.7%* | Savona | 1/59 | *1.7%* |
| Barletta-Andria-Trani | 1/10 | *10.0%* | Catania | 2/56 | *3.6%* | Campobasso | 1/62 | *1.6%* |
| Caltanissetta | 2/21 | *9.5%* | Taranto | 1/28 | *3.6%* | Sondrio | 1/68 | *1.5%* |
| Cagliari | 6/66 | *9.1%* | Potenza | 3/89 | *3.4%* | Mantova | 1/69 | *1.4%* |
| Rimini | 2/26 | *7.7%* | Lecce | 3/95 | *3.2%* | Pavia | 2/151 | *1.3%* |
| Teramo | 3/41 | *7.3%* | Milano | 4/129 | *3.1%* | Brescia | 2/186 | *1.1%* |
| Nuoro | 3/45 | *6.7%* | Frosinone | 2/71 | *2.8%* | Roma | 1/98 | *1.0%* |
| Latina | 2/32 | *6.3%* | Siena | 1/36 | *2.8%* | Bergamo | 2/222 | *0.9%* |
| L'Aquila | 4/68 | *5.9%* | Chieti | 2/76 | *2.6%* | Avellino | 1/112 | *0.9%* |
| Rieti | 2/41 | *4.9%* | Palermo | 2/77 | *2.6%* | Udine | 1/121 | *0.8%* |
| Trapani | 1/22 | *4.5%* | Pisa | 1/39 | *2.6%* | Varese | 1/131 | *0.8%* |
| Parma | 2/46 | *4.3%* | Viterbo | 1/51 | *2.0%* | Como | 1/143 | *0.7%* |
| Piacenza | 2/46 | *4.3%* | Pesaro e Urbino | 1/56 | *1.8%* | | | |

*Fonte*: Rilevazione dei Matrimoni (Modello Istat D.3) e Rilevazione mensile degli eventi demografici di Stato Civile (Modello Istat D7.a). Anno 2011.

Come esempio dell'output grafico fornito da SiRiO, consideriamo il caso della Provincia di Brindisi. In figura 4 è riportato il numero di matrimoni per Comune relativamente alle due fonti: in ascissa è riportata l'informazione aggregata (fonte D7.a) e in ordinata la somma ricavata dalle informazioni individuali (fonte D.3). Dei 20 Comuni analizzati, l'analisi FS ha individuato 4 Comuni anomali, identificati sul grafico dai punti chiari. Le linee tratteggiate rappresentano le rette di regressione stimate con e senza i valori relativi ai Comuni anomali. Lo spostamento dall'una all'altra retta, evidenzia l'effetto delle osservazioni anomale sulle stime dei parametri di regressione.

**Figura 4 – Grafico fornito da SiRiO sull'analisi FS di regressione: esempio della Provincia di Brinidsi**




*Fonte*: Rilevazione dei Matrimoni (Modello Istat D.3) e la Rilevazione mensile degli eventi demografici di Stato Civile (Modello Istat D7.a). Anno 2011.

## 8. Conclusione

Attraverso l'analisi robusta dei dati anomali è possibile individuare le osservazioni che hanno un comportamento diverso dal gruppo cui appartengono. Non esiste l'unità anomala in senso assoluto, ma un'unità è anomala rispetto ad un gruppo di riferimento e alle caratteristiche in studio.

Nel presente lavoro, si è utilizzata la tecnica FS per la ricerca di dati anomali univiariati e multivariati, applicata ai dati Istat sui matrimoni 2011.

L'analisi multivariata, condotta considerando i valori assoluti (modello gaussiano) e i valori relativi (modello composizionale) ha restituito risultati diversi, a sottolineare il ruolo centrale delle caratteristiche dei dati, oggetto di studio.

L'analisi di regressione, condotta per strato, restituisce 92 Comuni anomali. La stessa analisi, condotta senza stratificazione, non restituisce alcun Comune anomalo, a sottolineare l'importanza del gruppo di riferimento.

## Riferimenti bibliografici

Amato F., Filippini R., Francescangeli P., Scalfati F. e Toti S. 2013. "SiRiO: una web application per la ricerca di dati anomali multivariati". Poster presentato all'Undicesima Conferenze Nazionale di Statistica, Roma 20-21 febbraio.

Aitchison J. 1986. *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. London (UK): Chapman & Hall Ltd.

Aitchison J. e Shen S.M. 1980. "Logistic-Normal Distributions: Some Properties and Uses". *Biometrika*, 67, 2: 261–272

Atkinson A.C., Riani M. e Cerioli A. 2004. *Exploring Multivariate Data With the Forward Search*. New York: Springer Verlag.

Atkinson A.C, Riani M. 2006. "Distribution Theory and Simulations for Tests of Outliers in Regression". *Journal of Computational and Graphical Statistics*, 15: 460-476.

Becker C. e Gather U. 1999. "The Masking Breakdown Point of Multivariate Outlier Identification Rules". *Journal of the American Statistical Association*, 94: 947-955.

Egozcue J.J., Pawlowsky-Glahn V., Mateu-Figueras G. and Barcelò-Vidal C. 2003. "Isometric Logratio Transformations for Compositional Data Analysis". *Mathematical Geology*, 35, 3: 279–300.

Palombi F., Toti S., Filippini R. e Tomeo V. 2011. "A Forward Search Algorithm For Compositional Data". Key Invited Working Paper: Conference of European Statisticians (UNECE), Ljubljana 9-11 maggio.

Reale A., Torti F. e Riani M. 2012. "Robust methods for correction and control of Italian Agriculture Census data". Relazione presentata alla XLVI Riunione Scientifica della SIS, Roma, 20-22 giugno.

Riani M., Atkinson A.C. e Cerioli A. 2009. "Finding an unknown number of multivariate outliers". *Journal of the Royal Statistical Society, Series B Statistical Methodology*, 71: 447–466.

Tallis G.M. 1963. "Elliptical and radial truncation in normal samples". *Annals of Mathematical Statistics*, 34: 940-944.

Toti S., Palombi F. e Filippini R. 2011. "Outlier detection via Compositional Forward Search: application to the preliminary data of the 2011 Italian Agricultural Census". Relazione presentata al convegno: Convegno intermedio SIS, Bologna 8-10 giugno.