Agenzia nazionale per le nuove tecnologie,
l'energia e lo sviluppo economico sostenibile

MINISTERO DELLO SVILUPPO ECONOMICO

Ricerca di Sistema elettrico

Sviluppo di una procedura automatica di individuazione e trattamento di outlier in database di micro-dati di grandi dimensioni

Maurizio Vichi, Carlo Cavicchia

SAPIENZA
UNIVERSITÀ DI ROMA

Report RdS/PAR2017/100

SVILUPPO DI UNA PROCEDURA AUTOMATICA DI INDIVIDUAZIONE E TRATTAMENTO DI OUTLIER IN DATABASE DI MICRO-DATI DI GRANDI DIMENSIONI

Maurizio Vichi, Carlo Cavicchia (Sapienza Università di Roma, Diparimento di Scienze Statistiche)

Settembre 2018

Report Ricerca di Sistema Elettrico

Accordo di Programma Ministero dello Sviluppo Economico - ENEA
Piano Annuale di Realizzazione 2017
Area:
Progetto:
Obiettivo:
Responsabile del Progetto: Ing. Giovanni Puglisi, ENEA

Il presente documento descrive le attività di ricerca svolte all'interno dell'Accordo di collaborazione "Procedura automatica per l'identificazione di outliers"
Responsabile scientifico ENEA: Dott. Alessandro Federici
Responsabile scientifico Dipartimento di Scienze Statistiche Prof. Maurizio Vichi

Indice

Sommario

Lo studio ha riguardato lo sviluppo di una procedura automatica di trattamento di outliers in 7 database di micro-dati di grandi dimensioni con differenti caratteristiche che riguardano diverse tipologie di interventi di ristrutturazione e manutenzione con effetti energetici. Per questa terza annualità l'anno analizzato è il 2017. In questa sintesi abbiamo la possibilità di descrivere solo i risultati delle elaborazioni realizzate in codice Matlab con diversi programmi appositamente realizzati composti da diverse migliaia di istruzioni che rappresentano la procedura automatica per identificare statisticamente gli outliers e i dati errati. La procedura realizza automaticamente anche la loro imputazione insieme a quella dei dati mancanti, dove esplicitamente richiesto da ENEA. In particolare, la procedura realizza per ogni comma le statistiche puntuali su ogni singola imputazione e le rappresentazioni grafiche (boxplot e distribuzioni di frequenze) delle distribuzioni prima e dopo l'analisi.
Le variabili oggetto dello studio sono: il risparmio energetico, il costo dell'intervento e l'ammontare della detrazione. Sono utilizzate numerose variabili ausiliarie durante l'analisi. Le informazioni presenti nei dataset hanno consentito un lavoro specifico in base alle caratteristiche dei singoli interventi considerati. I dataset in questione rispondono a 7 differenti commi (344, 345a, 345b, 345c, 346, 347 e BA). Gli outliers sono stati individuati sulla base delle distribuzioni delle variabili sopracitate. In sintesi, la procedura misura la distanza di ogni unità statistica (ovvero per ogni intervento) dalla mediana della distribuzione in termini di standard deviation robusta (median absolute deviation (MAD)). Si identifica la SOGLIAsup = MEDIANA + K*MAD e la SOGLIAinf = MEDIANA – 0.4*K*MAD. Se il dato è identificato come anomalo perché supera la SOGLIAsup (per eccesso) o la SOGLIAinf (per difetto) è stato successivamente imputato con una metodologia del tipo "donatore" di minima distanza per Provincia. La scelta di calcolare le due soglie in maniera asimmetrica nasce dalla natura delle distribuzioni delle variabili d'interesse. Le informazioni fornite da ENEA in ausilio alla tecnica di imputazione hanno consentito un'analisi più pertinente e realistica. Per ogni dataset è stata svolta un'analisi preliminare di ricodifica delle variabili categoriali in modo da ottenere un dataset di soli valori numerici.
Per ogni dataset si è svolta un'analisi di sensitività sul parametro K in modo da scegliere il valore del parametro iterativamente, valutando i valori delle soglie identificate per le variabili d'interesse. Si è verificato inoltre che il valore del parametro risultasse coerente con la realtà dei dati studiati.
Oltre alla letturatura disponibile sul trattamento di dati anomali ci siamo avvalsi anche dei risultati conseguiti negli scorsi anni dove sono stati analizzati dei dataset analoghi per diverse annualità. Si evidenzia inoltre che si è verificata l'esistenza di nuove metodologie per l'identificazione dei dati anomali, e quella utlizzata è risultata ancora valida.
L'approccio multivariato non ha evidenziato la presenza di ulteriori outlier significativi, quindi si è scelto di non modificare il metodo per confrontabilità con gli anni precedenti.

RIEPILOGO AMMONTARE RISPARMIO E COSTI ANNO 2017

	344	345a	345b	345c	346	347	BA
Risparmio pre (kWh)/anno	189.890.947	602.472.589	39.159.034.906	17.356.981	6.926.223.630	623.568.390	8.733.056
Risparmio Post (kWh)/anno	90.956.886	429.915.065	613.493.437	21.110.915	53.511.507	268.368.649	9.180.349
Costi pre (€)	326.322.805	818.220.880	1.834.856.757	294.600.993	72.237.591	967.413.768	18.383.916
Costi post (€)	351.640.050	876.822.217	1.870.107.289	211.105.235	74.396.691	740.011.961	18.099.515

Introduzione

Lo studio ha riguardato lo sviluppo di una procedura automatica di trattamento di outliers in 7 database di micro-dati di grandi dimensioni con differenti caratteristiche che riguardano diverse tipologie di interventi di ristrutturazione e manutenzione con effetti energetici. Per questa terza annualità l'anno analizzato è il 2017. Descriveremo principalmente i risultati delle elaborazioni realizzate in codice Matlab con diversi programmi appositamente realizzati composti da diverse migliaia di istruzioni che rappresentano la procedura automatica per identificare statisticamente gli outliers e i dati errati. La procedura realizza automaticamente anche la loro imputazione insieme a quella dei dati mancanti, dove esplicitamente richiesto da ENEA. Illustreremo inoltre una sintesi delle metodologie utilizzate nella identificazione degli outliers e daremo una descrizione dei programmi MATLAB che sono stati realizzati appositamente per i sei database. In particolare, la procedura automatica di trattamento degli outliers realizza per ogni comma le informazioni puntuali su ogni singolo dato mancante o outliers identificato e sulle corrispondenti imputazioni. Inoltre la procedura produce le statistiche e le rappresentazioni grafiche (boxplot e distribuzioni di frequenze) delle distribuzioni prima e dopo l'analisi di controllo e correzione dei dati.
Le variabili oggetto dello studio sono: il risparmio energetico, il costo dell'intervento e l'ammontare della detrazione. Sono utilizzate numerose variabili ausiliarie durante l'analisi. I dataset in questione rispondono a 7 differenti commi (344, 345a, 345b, 345c, 346, 347 e BA). Nella prima parte dello studio si riporta una breve sintesi delle metodologie in uso per il controllo e la correzione dei dati. Si sono esaminati gli errori casuali, dovuti alla misurazione e gli errori sistematici, che si manifestano sempre nella stessa direzione. Tra gli errori sistematici sono stati distinti i dati mancanti, le incongruenze logiche e i valori fuori campo.
Enea ha fornito le seguenti regole per identificare le incongruenze logiche ossia le contraddizioni che si possono osservare tra i dati rilevati per ciascun record (unità statistica).

Regole di correzione degli errori fornite da ENEA:

Laddove emerga la necessità di correggere/sostituire il dato dichiarato o mancante, questo va corretto secondo le regole logiche che riterrete opportune valutando PRIORITARIAMENTE l'ambito provinciale o – se questo non significativo – regionale.

Per tutti i commi tali vincoli necessitano di una successiva post correzione:

Valore massimo di detrazione minore uguale al 65% valore di investimento dichiarato;
Valore minimo di detrazione maggiore di 0;
Valore investimenti maggiore di 0;
Risparmio energetico > 0;
Valore di detrazione inferiore a valore di investimenti.

Implementate le regole che individuano le incongruenze logiche si sono individuati i valori da correggere.
Gli outliers sono stati identificati secondo una procedura statistica appropriata che realizza il controllo e la correzione dei dati che ha le specifiche illustrate nel paragrafo che segue.

1 Descrizione delle attività svolte e risultati

1.1 Gli outliers

1.1.1 Individuazione di errori e dati anomali

L'individuazione, il trattamento e la misura dell'errore non campionario e dell'eventuale errore campionario rientra nella metodologia di controllo della qualità dei dati nota come “error profile”.
In ogni indagine statistica, parallelamente, ma anche successivamente alla fase di registrazione dei dati su memoria di massa e comunque prima di effettuare le elaborazioni statistiche, viene messa in atto una vera e propria procedura di controllo e di revisione, che ha lo scopo di verificare il rispetto dei vincoli di integrità imposti dal fenomeno in esame.
In ogni fase di una indagine statistica si possono commettere errori, che se non sono individuati e corretti, possono pregiudicare in maniera significativa l'interpretazione dei risultati relativi al fenomeno osservato.
Quindi prima di effettuare le elaborazioni statistiche e la successiva interpretazione dei dati, conviene attivare una fase controllo e di correzione che, in estrema sintesi, permetta di individuare ed eventualmente rimuovere errori e incongruenze nel materiale di rilevazione.
Infatti si devono controllare il rispetto dei vincoli di integrità del fenomeno in esame verificando; per le variabili: l'ammissibilità delle modalità assunte da ciascuna unità statistica rilevata, la congruenza logica delle risposte al questionario relativo ad una unità statistica rilevata, le congruenza logica fra le risposte di diverse unità statistiche; per le unità, ossia per i tipi di record ad esse associati: la loro corretta identificazione tramite la chiave primaria del record, la corretta identificazione di una relazione tra unità (associazione 1:1, 1:n, n:m) , tramite la chiave esterna del record.
Gli errori e le incongruenze riguardano quindi le unità statistiche e le modalità delle variabili esaminate.
L'errore sistematico è quell'errore che si presenta con una certa frequenza (dal 3% al 5% secondo NCBS Statistics Sweden) nell'insieme dei dati rilevato, dovuto a cause che operano sempre nella stessa direzione. Gli errori sistematici possono manifestarsi come dati mancanti, incongruenze logiche, valori fuori campo.
Un dato mancante si verifica quando, per esempio, ad una domanda di un questionario non viene data risposta. Spesso la mancata risposta può dipendere dalla natura delicata dell'argomento trattato, o dalla mancanza di una risposta aderente alle idee dell'intervistato in una domanda strutturata.
L'incongruenza logica si determina quando si manifestano delle contraddizioni nelle informazioni rilevate su una unità statistica. Per esempio, in un questionario si verificano delle evidenti contraddizioni nelle risposte fornite dall'intervistato.
Un dato inammissibile o fuori campo è una modalità registrata tra le risposte di una variabile nella domanda di un questionario, ma che non risulta nella scala delle modalità della variabile.
Una volta individuati i dati mancanti, quelli fuori campo e le incongruenze logiche dobbiamo decidere se intervenire nella loro correzione. Nell'effettuare le eventuali correzioni dobbiamo sempre verificare se le notizie introdotte a rettifica degli errori sistematici soddisfino due principi:
quello della verosimiglianza delle correzioni, mediante il quale si vuole mantenere la coerenza fra le notizie registrate nei questionari e quelle corrette;
quello del minimo cambiamento nelle correzioni che consiste nel minimizzare i cambiamenti delle informazioni raccolte.

Le operazioni di individuazione e correzione degli errori sono state condotte in un piano di compatibilità e di correzione. Gli errori basati su incongruenze sono individuati mediante le regole

di incompatibilità sopra enunciate, mentre le correzioni avvengono attraverso due differenti criteri: il criterio deterministico e quello stocastico.
Il criterio deterministico consiste nel correggere un errore, quale un dato mancante in una domanda, immettendo un valore predeterminato, come può essere la moda o la media aritmetica delle modalità osservate o una modalità scelta casualmente da una distribuzione relativa alle risposte di quella domanda, distribuzione determinata generalmente da una precedente indagine. Si noti che i dati utilizzati per la correzione con un criterio deterministico sono esterni alla rilevazione. La correzione avviene attraverso una forzatura.
Il criterio stocastico consente di correggere gli errori attraverso una unità statistica che risulta la più simile rispetto a quelle osservate in cui non si è riscontrato alcun errore. Quest'ultima unità viene denominata donatore. I dati utilizzati per la correzione degli errori nel caso del criterio stocastico sono interni al file dei dati osservati. Si distinguono due metodi da donatore il cold-deck e l'hot-deck. Nel primo si effettua uno screenning per distinguere le unità senza errori da quelle con almeno un errore, mentre nel secondo l'insieme delle unità senza errore viene aggiornato mano a mano che le unità vengono esaminate. Nella presente procedura automatica si è adottato il criterio stocastico che utilizza i dati stessi per individuare il donatore più idoneo per l'imputazione del dato mancante o del dato anomalo.
Prima di procedere alla identificazione degli outliers è necessario soffermarsi alla loro definizione e brevemente illustrare le diverse metodologie utilizzare per individuarli.

1.1.2 Le diverse tipologie di Outliers

"Un outlier è un'osservazione molto distante dalle altre osservazioni a tal punto da supporre che sia stato prodotto tramite un meccanismo differente." [Hawkins, 1980]

Definzioni:

1. Outlier di singolo costrutto: dati troppo grandi o troppo piccoli comparati agli altri valori dello stesso costrutto. Di solito questi valori sono nelle code delle distribuzioni dei dati;

2. Outliers frutto di errori: dati distanti dalla nuvola dei punti per mancanza di accuratezza nella raccolta dei dati. Sono valori che non fanno parte della distribuzione in quanto fuori dai limiti massimi consentiti, spesso sono frutto di errori di manipolazione dei dati in una fase preliminare dell'analisi;

3. Outliers di interesse: dati che fanno parte della popolazione ma sono comunque ai margini di questa e vanno ad individuare una particolare caratteristica del fenomeno studiato;

4. Outliers di discrepanza: dati con un grande valore residuale e che possono condizionare il fit del modello e la stima dei parametri;

5. Outliers per il fit del modello: dati che con la loro presenza influenzano il fit del modello;

6. Outliers per la previsione: dati che con la loro presenza influenzano la stima dei parametri del modello.

Gli outlier possono essere distinti in:

1. outlier non rappresentativi: si tratta di valori anomali a causa di veri e propri errori in fase di compilazione del questionario. Un caso classico è costituito dall'errore nell'unità di misura utilizzata per la risposta (ad esempio, euro invece di migliaia di euro, per cui i valori

dichiarati dovrebbero essere divisi per mille). La loro non rappresentatività va intesa con riferimento alle unità della popolazione non incluse nel campione, perché non contribuiscono alla variabilità campionaria fornendo informazioni su di esse. Si tratta di veri e propri errori che occorrerebbe individuare e correggere a monte. Questo studio è particolarmente ricco di errori di questo tipo che riguardano le misurazioni delle superfici (ad esempio infissi);

2. outlier rappresentativi: si tratta di valori anomali non dovuti ad errori di misurazione, bensì ad eventi relativi all'unità di riferimento non (del tutto) valutabili sulla base delle informazioni disponibili su di essa. Si tratta comunque di osservazioni rare rappresentative di un certo numero di unità della popolazione che spesso non sono incluse nel campione e di cui generalmente non si conosce l'ammontare.

1.1.3 Tecniche per l'identificazione di outliers:

Lo studio dei dati anomali è un punto cruciale di ogni analisi statistica e quindi in letteratura sono presenti molti modelli ed approcci per la loro individuazione.
Negli anni sono state sviluppate molte tecniche, spesso nuovi modelli, però, sono utilizzabili solamente sotto ipotesi ed assunzioni molto restrittive. Si possono distinguere:

1. **Tecniche per singoli costrutti:**

- Box-Plot: un grafico che mostra una sintesi dei più piccoli ed i più grandi valori del costrutto (esclusi gli outliers), gli outliers possono essere considerati i punti che vanno oltre il grafico.

- Stem and leaf plot: un grafico che simultaneamente ordina i dati quantitativi e fornisce informazione riguardo la forma della distribuzione.

- Schematic plot analysis: un grafico simile al Box-Plot, ma utilizzato nella meta-analisi.

- Standard deviation analysis: distanza di un punto dalla media in unità di deviazione standard.

- Pencetage analysis: relativa alla normalità di un punto in una distribuzione di punteggi indicizzata per il suo percentile.

2. **Tecniche per costrutti multipli:**

- Scatter Plot: un grafico dei valori di due variabili. Una delle quali sull'asse x (variabile indipendente) ed una sull'asse y (variabile dipendente). Un potenziale outlier potrebbe essere individuato dalla sua distanza dal centroide dei dati.

- q-q Plot: un grafico che confronta due distribuzioni di probabilità misurando la distanza tra i quantili di una con i quantili dell'altra. Un andamento non lineare indica la presenza di outliers.

- p-p Plot: un grafico che assegna il grado di similarità di due insieme di dati confrontando le due funzioni di distribuzioni cumulate. Un andamento non lineare indica la presenza di outliers.

- Residui Standardizzati: un valore residuale calcolato dividendo il residuo della i-esima osservazione per la deviazione standard. Osservazioni con valori residuali più alti sono candidati per essere outliers. Tuttavia, il valore residuale standardizzato delle osservazioni non è una misura di quanto esse siano outliers.

- Distanza Euclidea: lunghezza di un segmento tra due specifici punti in una spazione uni-bi-tri-n dimensionale. Una grande distanza tra due punti potrebbe significare che uno dei due punti è un outliers.

- Distanza di Mahalanobis: simile a quella Euclidea, ma la distanza di Mahalanobis è calcolata per ogni punto dal centroide calcolato tra gli altri punti. Una grande distanza tra due punti potrebbe significare che il punto su cui è calcolata è un outliers.

3. Tecniche basate sull'influenza:

- Cook's Di: valuta l'influenza che il punto i ha su tutti i coefficienti di regressione come un intero.

- Cook's Di modificata: simile alla Cook's Di ma questa utilizza i residui standardizzati corretti piuttosto che i residui standardizzati.

- Cook's Di generalizzata: simile alla Cook's Di ma questa tecnica valuta l'influenza che il punto i ha sulla stima dei parametri.

- Tecnica vicini più vicino: calcola i valori più vicini al valore d'interesse utilizzando una qualche distanza. La tecnica include i K vicini più vicini utlizzando una delle tecniche note (Type I, Type II, PAM, CLARANS, etc).

- Metodi non parametrici: consistono in metodo che fittano la curva senza porre particolari vincoli sui dati. La mancanza di un trend lineare segnala la presenza di outliers.

- Metodi parametrici: a diffenrenza dei metodi non parametrici, nei metodi parametrici si fanno forti assunzioni sulla natura dei dati, assunzioni che possono riguardare la distribuzione di probabilità degli stessi. Gli outliers vengono individuati tra i punti che non rispondo alla normale natura dei dati. I metodi più noti sono: convex peeling, ellipsoidal peeling, iterative deletion, trimming, last median square and M-estimation.

- Metodi semiparametrici: questi metodi tentano di combinare la velocità e complessità dei metodi parametrici con la flessibilità dei metodi non parametrici.

- Analisi della componente indipendente: un metodo che permette di separare le componenti indipendenti massimizzando l'indipendenza tra esse. Le componenti indipendenti isolate sono outliers.

Riassumendo, si possono utilizzare le seguenti operazioni e tecniche di analisi:

1. Correzione errori dei dati con valori più propri alla natura dei dati.

2. Rimozione outliers.

3. Studio degli outliers nel dettaglio.

4. Preso atto della presenza di outliers, non si fanno assunzioni su essi a priori.

5. Report dei risultati delle analisi con e senza outliers.

6. Trasformazione dei dati estremi con uno specifico percentile della distribuzione (per esempio la 90-esimo percentile Winsorization prevede che tutti i valori sotto al 5-to percentile vengano trasformati con il valore del 5-to percentile e tutti i valori sopra al 95-esimo percentile con il valore del 95-esimo percentile).

7. Troncare la distribuzione in modo tale di prendere in considerazione solo i valori all'interno di un certo range.

8. Applicazione di una funzione deterministica per ogni valore del dataset.
9. Cambiare manualmente il valore di un outlier in una valore meno estremo per la distribuzione.

10. Minima deviazione assoluta: si scelgono i valori dei coefficienti di regressione che limitano i residui producendo una funzione che approssima i dati.

11. Minimi quadrati trimmati: i residui vengono ordinati per ogni caso dal più alto al più basso e poi vengono trimmati i valori estremi.

12. M-estimation: un insieme di statistiche robuste che riduce l'effetto dell'influenza degli outliers rimpiazzando i residui quadrati con un'altra funzione dei residui.

13. Le statistiche bayesiane servono a stimare i parametri sfruttando le informazioni a priori che si hanno sui dati.

14. La procedura robusta Two-Stage usa la distanza di Mahalanobis per assegnare i pesi ad ogni dato, così che i casi estremi vengano sottostimati nelle variabili predittrici.

15. Simile alla Two-Stage, l'assegnazione iterativa dei pesi dei minimi quadrati usa la distanza di Mahalanobis ma essa è completata dall'uso di un algoritmo iterativo.

16. Stima della varianza e delle covarianze nella parte causale di un modello multilivello direttamente dai residui tramite il metodo GEE (Generalized estimating equations). Questo approccio si basa sulla stima degli effetti medi sulla popolazione. Nonostante GEE produca stime meno efficienti rispetto alla massima verosimiglianza, GEE necessita di assunzioni più deboli riguardo la struttura della parte casuale del modello multilivello.

17. Il metodo Bootstrapping stima i parametri di un modello ed i loro errori dal campione senza riferimento alla teoria distribuzione del campione stesso. Come risultato il ricercatore può calcolare le stime del valore atteso e della variabilità delle statistiche come se fossero prese da un'empirica distribuzione campionaria.

18. Le analisi non parametriche non necessitano di particolari distribuzioni assunte per i dati. La loro flessibilità aiuta il ricercatore a trovare risultati robusti anche in presenza di outliers.

19. Ottenere statistiche meta-analitiche che non danno maggior peso agli studi preliminari in numerosità campionaria maggiore.

20. M-estimation generalizzata: una classe di tecniche robuste che riduce l'effetto degli outliers sostituendo i residui quadrati con un'altra funzione residuale.

Ipotizzando uno scenario totalmente non supervisionato, ovvero uno scenario nel quale non è presente un training set dei dati esistono tre approcci di classificazione degli outliers:

- Rilevazione outliers globali o locali: l'anomalia di ogni punto è valutata rispetto ad un insieme di riferimento di oggetti.

- Etichettatura o assegnazione punteggio outliers: considera l'output di un algoritmo.

- Proprietà modellistiche: considera che il concetto di anomalia sia modellizzato.

- **Test statistici**

Data una certa famiglia di distribuzioni statistiche e assumendo che, dopo aver stimato i parametri, tutti i punti appartengano ad una determinata distribuzione della famiglia; gli outliers sono punti che hanno bassa probabilità di essere generati da quella stessa distribuzione (ad esempio, a distanza 3 volte la deviazione standard dalla media).
Sono disponibili un numero molto elevato di test statistici variando alcuni aspetti fondamentali nella definizione di test:

- Famiglia di distribuzione

- Numero di variabili

- Numero di distribuzioni

- Parametriche o non parametriche

- **Approccio basato sulla profondità**

L'idea di base sta nel fatto di ricercare gli outliers ai confini della nuvola di punti rappresentante i dati indipendentemente dalle distribuzioni delle variabili.
Si racchiude la nuvola di punti in un poligono convesso con vertici i punti più estremi della nuvola, i punti sui lati di tale figura sono gli outliers di profondità 1.

[Tukey, 1977]
Ipotizzando di togliere gli outliers a profondità 1 e procedere analogamente sui punti rimanenti si individuano gli outliers di profondità 2, e cosa via. Quindi i punti che risultato avere una profondità <k sono riportati come outliers.
L'approccio basato sulla profondità è particolarmente efficacie in spazi bidimensionali o tridimensionali.

- **Approccio basato sullo scostamento**

Dato un insieme di punti, gli outliers sono punti che non rispondono alla generali caratteristiche dell'insieme dato. La varianza dell'insieme è minimizzata rimuovendo tali punti.

[Arning, 1996]
Dato un fattore di omogeneità SF(I) che calcola per ogni I in DB quanto decresce il valore della varianza di DB quando I viene rimosso da DB. Se due insiemi hanno lo stesso valore di SF si considera quello di dimensione minore.
Gli outliers sono gli elementi dell'insieme "eccezione" E in DB per il quale: SF(E)>SF(I) per ogni I in DB.
La complessità computazione di questo approccio è di O(2n) per n oggetti, ed è un approccio applicabile a dati di ogni natura.

- **Approccio basato sulla distanza**

Si giudica ogni punto in base alla sua distanza dai suoi vicini. Gli outliers sono quei punti che hanno un "vicinato poco denso".

[Knorr, Ng, 1997]
Dati un raggio ε ed una percentuale π, un punto p è considerato un outlier se al massimo il π percento di tutti i punti hanno distanza da p meno di ε.

$OuliersSet(\varepsilon,\pi) = \{p|[Card(\{q \text{ in } DB|dist(p,q)<\varepsilon\})/Card(DB)]< \pi\}$

- **Approccio basato sulla densità**

Si confronta la densità intorno ad un punto con la densità intorno ai suoi vicini locali: la densità relativa di un punto comparata con quella dei suoi vicini è calcolata come un punteggio dell'outlier.
I vari approcci diferiscono in base a come la distribuzione di densità venga stimata.
La densità intorno ad un dato anomalo è considerabilmente differente rispetto a quella dei suoi vicini.

- **Approccio High-dimensional**

Lo scopo è quello di trovare outliers in proiezioni (sottospazi) dello spazio completo di partenza, si usano distanze più robuste per individuare outliers a dimensione piena.

[Kriegel et al, 2008]
ABOD – angle-based outlier degree, è un metodo di individuazione degli outliers che parte dall'assunto che l'angolo sia una misura più stabile della distanza quando si tratta di spazi a grandi dimensioni. L'oggetto o è un outlier se la maggior parte degli altri oggetti sono posizionati in simili direzioni, mentre l'oggetto o non è un outlier se la maggior parte degli altri oggetti sono posizionati in direzioni differenti.
L'assunzione posta alla base di questo metodo è che un outlier è sempre posizionato al confine della distribuzione.
Considerando per un punto p l'angolo tra xp e yp (i due vettori che collegano p rispettivamente ai punti x ed y) per due qualsiasi punti x ed y del database, si può costruire lo spettro di tutti questi

angoli: la larghezza dello spettro è un punteggio per il grado di anomalia di tutti i punti del database.

Il modello prevede di misurare la varianza dello spettro pesando le corrispondenti distanze:
ABOD(p)= VAR ((<xp,yp>/||xp||2 ||yp||2)
Quindi se il valore di ABOD risulta essere piccolo allora ci troviamo davanti ad un outlier altrimenti se il valore è grande siamo davanti ad un dato non anomalo.
Questo modello ha una complessità computazionale pari a $O(n^3)$.

La procedura automatica di identificazione degli errori e degli outliers utilizzata in questo progetto di ricerca tiene conto delle diverse tecniche che sono state precedentemente illustrate.
In particolare, sono state utilizzate tecniche per singoli costrutti:

- Box-Plot: un grafico che mostra una sintesi dei più piccoli ed i più grandi valori del costrutto (esclusi gli outliers), gli outliers possono essere considerati i punti che vanno oltre il grafico.

- Standard deviation analysis: distanza di un punto dalla media (mediana) in unità di deviazione standard (robusta).

In sintesi, quest'ultima procedura misura la distanza di ogni unità statistica (ovvero per ogni intervento) dalla mediana della distribuzione in termini di standard deviation robusta (median absolute deviation (MAD)) per quanto riguarda la soglia superiore, mentre identifica la soglia inferiore tenendo conto dell'asimmetria a destra di tutte le distribuzioni e quindi misurando la percentuale di distanza tra il minimo e la mediana stessa. Quindi si è identificata la SOGLIAsup = MEDIANA + *K**MAD e SOGLIAinf = MEDIANA – P*(MEDIANA – MINIMO), per ogni variabile e per sottopopolazione studiata.
Nell'analisi svolta si osserva che il valore SOGLIAsup per le variabili esaminate domina l'identificazione dei valori anomali anche di tipo multivariato.
La scelta di identificare gli outliers inferiori in maniera differente nasce dal fatto che le distribuzioni delle variabili prese in esame presentano tutte una forte asimmetria che bisogna tenere in considerazione in questo tipo di studio.
Una volta identificato un dato come anomalo poiché supera la SOGLIAsup per eccesso o la SOGLIAinf per difetto, viene successivamente imputato con una metodologia del tipo "donatore" di minima distanza per Provincia.
Per ogni dataset è stata svolta un'analisi preliminare di ricodifica delle variabili categoriali e binarie in modo da ottenere un dataset di soli valori numerici. Illustriamole brevemente nel prossimo paragrafo.
Al fine di valutare il valore K per la determinazione della SOGLIAsup, è stata studiata la distribuzione empirica della SOGLIAsup per vari valori di K utilizzando ricampionamenti con metodo bootstrap. Si è osservato che, per la quasi totalità dei casi, si ha il miglior bilanciamento tra la riduzione di varianza osservata e numero di dati considerati anomali per K=4.
E' stata svolta un'analisi simile anche per la scelta di P, il miglior bilanciamento tra riduzione di varianza osservata e numero di dati considerati anomali si ottiene per P=0.8 per le variabili d'interesse e P=0.5 per le variabili "superficie".
Oltre alla letturatura disponibile sul trattamento di dati anomali ci siamo avvalsi anche dei risultati conseguiti negli scorsi anni dove sono stati analizzati dei dataset analoghi per diverse annualità. Si evidenzia inoltre che si è verificata l'esistenza di nuove metodologie per l'identificazione dei dati anomali, e quella utlizzata è risultata ancora valida.

L'approccio multivariato non ha evidenziato la presenza di ulteriori outlier significativi, quindi si è scelto di non modificare il metodo per confrontabilità con gli anni precedenti.

1.2 Operazioni di pulizia e ricodifica database (Anno 2017)

Prima di effettuare le operazioni di controllo e correzione dei dati è stato necessario effettuare per ognuno dei sette database dei dati del 2017 alcune operazioni di pulizia e ricodifica dei dati che sono di seguito riportate per ciascun comma.
In tutti i dataset tutte le variabili quantitative sono espresse con due cifre decimali (separato scelto: , "virgola").

1.2.1 Comma 344

Il database è composto da 4276 records.
Le variabili prese in considerazione sono: CPID, Provincia, Regione, Destinazione d'uso generale, Fabbisogno di energia primaria per la climatizzazione invernale [kWh/anno], Indice di prestazione energetica per la climatizzazione invernale proprio dell'edificio [kWh/mq anno o kWh/mc anno], Pertinente valore limite dell'indice di prestazione energetica limite per la climatizzazione invernale [kWh/mq anno o kWh/mc anno], Volume lordo riscaldato V [mc.], Superficie utile [mq.] (autom. dall'anagrafica), Numero unità immobiliari oggetto dell'intervento, Superficie totale [m2] Pareti Verticali, Superficie totale [m2] Pareti Orizzonatali o Inclinate, Superficie totale [m2] Infissi, Totale potenza nominale al focolare del nuovo generatore termico / Potenza elettrica assorbita / Potenza termica nominale [kW], Risparmio energetico stimato in fonti primarie [Kwh], Costo dell'intervento di qualificazione energetica al netto delle spese professionali [Euro], Costo delle spese professionali [Euro], Detrazione fiscale (a cura del contribuente: 55% per le spese sostenute prima del 06.06.2013 e 65% per le spese successive), Intervento.

RICODIFICA
Il dataset è stato ripulito e le variabili categoriali sono state ricodificate con valori numerici, i dati mancanti sono stati ricodificati con il valore 0:

- Destinazione d'uso generale: Residenziale (1) e Non residenziale (2).
- Intervento: Mult (1), CI (2), Altro (3) e Vuoto (0).

1.2.2 Comma 345a

Il database è composto da 23563 records.
Le variabili prese in considerazione sono: CPID, Provincia, Regione, Destinazione d'uso generale, Fabbisogno di energia primaria per la climatizzazione invernale [kWh/anno], Indice di prestazione energetica per la climatizzazione invernale proprio dell'edificio [kWh/mq anno o kWh/mc anno], Pertinente valore limite dell'indice di prestazione energetica limite per la climatizzazione invernale [kWh/mq anno o kWh/mc anno], Volume lordo riscaldato V [mc.], Superficie utile [mq.] (autom. dall'anagrafica), Numero unità immobiliari oggetto dell'intervento, Superficie totale [m2] Pareti Verticali, Superficie totale [m2] Pareti Orizzontali o Inclinate, Superficie totale [m2] Infissi, Totale potenza nominale al focolare del nuovo generatore termico / Potenza elettrica assorbita / Potenza termica nominale [kW], Risparmio energetico stimato in fonti primarie [Kwh], Costo dell'intervento di qualificazione energetica al netto delle spese professionali [Euro], Costo delle spese professionali [Euro], Detrazione fiscale (a cura del contribuente: 55% per le spese sostenute prima del 06.06.2013 e 65% per le spese successive), Intervento.

RICODIFICA
Il dataset è stato ripulito e le variabili categoriali sono state ricodificate con valori numerici, i dati mancanti sono stati ricodificati con il valore 0:

- Destinazione d'uso generale: Residenzale (1), Non residenzale (2) e Misto (3).
- Intervento: Mult (1), PV (2), POI (3), I (4) e Vuoto (0).

1.2.3 Comma 345b

Il database è composto da 227172 records.
Le variabili prese in considerazione sono: CPID, Provincia, Regione, Destinazione d'uso generale, Superficie utile [mq.], Numero unità immobiliari oggetto dell'intervento, Superficie complessiva di telaio e vetro oggetto dell'intervento [mq.], Trasmittanza media del nuovo infisso [W/mq.K], Risparmio energetico stimato [kWh] (calcolo automatico), Costo dell'intervento di qualificazione energetica al netto delle spese professionali [Euro], Costo delle spese professionali [Euro], Detrazione fiscale (a cura del contribuente: 55% per le spese sostenute prima del 06.06.2013 e 65% per le spese successive), Tipologia intervento.

RICODIFICA
Il dataset è stato ripulito e le variabili categoriali sono state ricodificate con valori numerici, i dati mancanti sono stati ricodificati con il valore 0:

- Destinazione d'uso generale: Residenzale (1), Non residenzale (2) e Misto (3).
- Tipologia intervento: codifica basata sulla tipologia di vetro e di telaio esistenti dopo l'intervento,
  - Tipologia di telaio esistente dopo l'intervento: Legno (1), Metallo, no taglio termico (2), Metallo, taglio termico (3), PVC (4) e Misto (5);
  - Tipologia di vetro esistente dopo l'intervento: Singolo (1), Doppio (2), Triplo (3), A bassa emissione (4) e Nessuno (5).

1.2.4 Comma 345c

Il database è composto da 84953 records.
Le variabili prese in considerazione sono: CPID, Provincia, Regione, Destinazione d'uso generale, Superficie utile [mq.], Numero unità immobiliari oggetto dell'intervento, Superficie totale [m2], Risparmio energetico stimato in fonti primarie [Kwh], Costo dell'intervento di qualificazione energetica al netto delle spese professionali [Euro], Costo delle spese professionali [Euro], Detrazione fiscale (a cura del contribuente: 55% per le spese sostenute prima del 06.06.2013 e 65% per le spese successive).

RICODIFICA
Il dataset è stato ripulito e le variabili categoriali sono state ricodificate con valori numerici, i dati mancanti sono stati ricodificati con il valore 0:

- Destinazione d'uso generale: Residenzale (1) e Non residenzale (2).

1.2.5 Comma 346

Il database è composto da 8236 records.
Le variabili prese in considerazione sono: CPID, Provincia, Regione, Destinazione d'uso generale, Superficie utile [mq.], Numero unità immobiliari oggetto dell'intervento, Superficie netta totale pannelli piani (o area di apertura, da certificato allegato al collettore) [mq.], Risparmio energetico

stimato in fonti primarie [Kwh], Costo dell'intervento di qualificazione energetica al netto delle spese professionali [Euro], Costo delle spese professionali [Euro], Detrazione fiscale (a cura del contribuente: 55% per le spese sostenute prima del 06.06.2013 e 65% per le spese successive).

RICODIFICA
Il dataset è stato ripulito e le variabili categoriali sono state ricodificate con valori numerici, i dati mancanti sono stati ricodificati con il valore 0:

- Destinazione d'uso generale: Residenzale (1) e Non residenzale (2).

1.2.6 Comma 347

Il database è composto da 87611 records.
Le variabili prese in considerazione sono: CPID, Provincia, Regione, Superficie utile [mq.], Nuovo tipo di generatore di calore, Numero unità immobiliari oggetto dell'intervento, Risparmio energetico stimato in fonti primarie (kWh) per l'impianto termico, Risparmio energetico stimato in fonti primarie (kWh) per la produzione di a.c.s, Costo dell'intervento di qualificazione energetica al netto delle spese professionali [Euro], Costo delle spese professionali [Euro], Detrazione fiscale (a cura del contribuente: 55% per le spese sostenute prima del 06.06.2013 e 65% per le spese successive), Intervento, Potenza.

RICODIFICA
Il dataset è stato ripulito e le variabili categoriali sono state ricodificate con valori numerici, i dati mancanti sono stati ricodificati con il valore 0:

- Nuovo tipo di generatore di calore: Pompa di calore (1), Caldaia a condensazione (2), Altro (3), Impianto geotermico (4), Caldaia a biomassa (5) e Null (0).
- Intervento: Pompa di calore (1), Caldaia a condensazione (2), Altro (3), Impianto geotermico (4), Caldaia a biomassa (5) e Null (0).

1.2.7 Comma BA

Il database è composto da 1913 records.
Le variabili prese in considerazione sono: CPID, Provincia, Regione, Destinazione d'uso generale, Superficie utile [mq.], Numero unità immobiliari oggetto dell'intervento, Superficie totale [m2], Risparmio energetico stimato in fonti primarie [Kwh], Costo dell'intervento di qualificazione energetica al netto delle spese professionali [Euro], Costo delle spese professionali [Euro], Detrazione fiscale (a cura del contribuente: 55% per le spese sostenute prima del 06.06.2013 e 65% per le spese successive).

RICODIFICA
Il dataset è stato ripulito e le variabili categoriali sono state ricodificate con valori numerici, i dati mancanti sono stati ricodificati con il valore 0:

- Destinazione d'uso generale: Residenzale (1) e Non residenzale (2).

1.3 Descrizione e implementazione della procedura automatica in MATLAB

La procedura automatica in MATLAB è composta, per ogni comma, da un programma base che effettua tutte le operazioni di pulizia e ricodifica delle variabili ed in alcuni casi creandone anche delle nuove come indicatori di tipologia d'intervento o variabili che rappresentano la "superficie" dell'intervento effettuato; ed altri programmi che effettuano le imputazioni sulle variabili d'interesse, ognuno di questi è richiamato dal programma base.
I programmi "imputazione" sono differenti in base alla natura della variabile studiata e alla metodologia con la quasi si è effettuato il controllo dei dati anomali. In alcuni casi i programmi sono vere e proprie funzioni che prendono in input la variabile d'interesse e l'indicatore della sottopopolazione studiata per ridare in output l'indice delle unità imputate e la nuova variabili con i valori modificati.
In alcuni casi le funzioni prendono in input due variabili, una è quella d'interesse mentre l'altra è quella utilizzata per individuare le anomalie (come nel caso di Costo su Risparmio per studiare Costo).
Nel caso delle variabili "superficie", in tutti i commi eccetto il 345b dove è stata scritta una funzione, è stato scritto un programma a parte che studia la variabile per ogni sottopopolazione.
Per alcuni commi è stato poi necessario scrivere un programma a parte per l'imputazione dei dati mancanti delle variabili categoriali.
E' importante notare che la prima operazione è sempre stata quella di individuazione delle sottopopolazioni d'interesse per lo studio ed ogni variabile è stata studiata separatamente per tutte le sottopopolazioni. Lo scopo di questa segmentazione è quello di creare sottopopolazioni omogenee per intervento nelle quali potessero evidenziarsi situazioni di anomalia o errore. In queste popolazioni si sono poi potuti identificare i "donatori" corrispondenti alla mediana di un sottoinsieme ancora più omogeneo relativo ad esempio alla provincia nella quale è stato richiesto l'intervento.
Ogni programma produce rappresentazioni grafiche delle distribuzioni studiate pre e post studio (Box Plot e Istogrammi) e uno storico delle statistiche di tutte le imputazioni svolte. Le matrici in questione presentano nella prima riga (contatore = 1) il numero di unità della sottopopolazione studiata, il valore della mediana, il valore della soglia inferiore e il valore della soglia superiore, mentre dalla seconda riga in poi presentano l'indice dell'unità nella popolazione, il valore del CPID, il valore anomalo e il nuovo valore dopo l'imputazione.
La dimensione estesa dei commi (specialmente il 345b) ha reso necessario l'utilizzo di calcolatori molto potenti e di versioni aggiornate di MATLAB per portare a termini tutte le operazioni. I programmi sono molto lunghi e le operazioni sono molto onerose ma la procedura risulta molto flessibile e adattativa alle molteplici differenziazioni di interventi.
E' importante notare che i dataset presentavano spesso situazioni di studio complesse, quindi è stata necessario una lunga fase di "taratura" della procedura tramite delle simulazioni che hanno permesso di ottenere risultati coerenti con la realtà di studio. Un'attenzione particolare è stata posta sui valori minimi assunti dalle variabili dopo le imputazioni per garantire valori sempre veritieri.

1.4 Applicazione della procedura automatica ai dati (2017)

La procedura di individuazione ed imputazione dei dati mancanti e dei dati anomali è stata ideata misurando la distanza di ogni unità statistica (ovvero per ogni intervento) dalla mediana della distribuzione in termini di standard deviation robusta (median absolute deviation (MAD)) per quanto riguarda la soglia superiore, mentre identifica la soglia inferiore tenendo conto dell'asimmetria a destra di tutte le distribuzioni e quindi misurando la percentuale di distanza tra il minimo e la mediana stessa. Quindi si è identificata la SOGLIAsup = MEDIANA + K*MAD e SOGLIAinf = MEDIANA – P*(MEDIANA – MINIMO), per ogni variabile e per sottopopolazione

studiata. Se il dato supera la SOGLIA allora è identificato come anomalo ed è stato successivamente imputato con una metodologia del tipo "donatore" di minima distanza per Provincia.

1.4.1 Comma 344

Dopo la fase di pulizia e ricodifica del database, la procedura di individuazione e correzione dei dati mancanti e dei dati anomali ha visto la creazione di quattro programmi MATLAB, uno che è alla base di tutto lo studio e da dove vengono richiamati gli altri tre che invece si occupano della fase di imputazione (ognuno di essi in base alla natura della variabile studiata).

E' importante rimarcare come ogni valutazione sulle variabili è stato preceduto da un'opportuna fase di normalizzazione per altre variabili di supporto allo studio.
Le prime operazioni svolte hanno visto la creazione di indicatori della tipologia di intervento svolto e la creazione della variabile "superficie totale". Sulla base degli indicatori creati sono state individuate alcune sottopopolazioni per rendere l'imputazione dei dati anomali più precisa ed accurata. I primi risultati ottenuti sono:

- 596 imputazioni per la sottopopolazione PARETI VERTICALI (numerosità della popolazione 3057);
- 1370 imputazioni per la sottopopolazione PARETI ORIZZONTALI O INCLINATE (numerosità 2814);
- 766 imputazioni sulla "superficie totale" per la sottopopolazione INFISSI (numerosità 3237).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

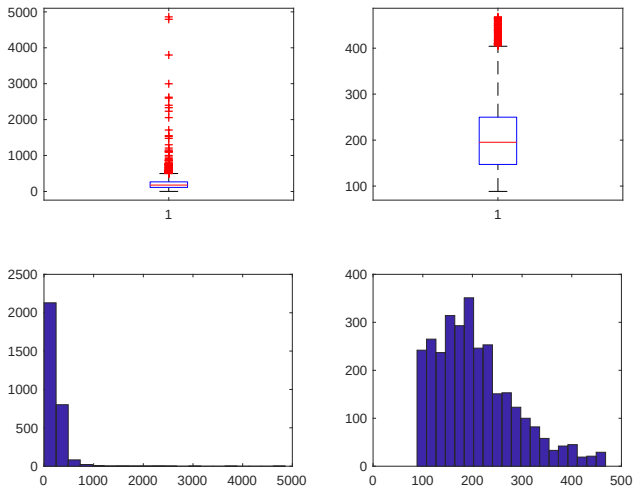

Figura 1. Superficie Parete Verticali

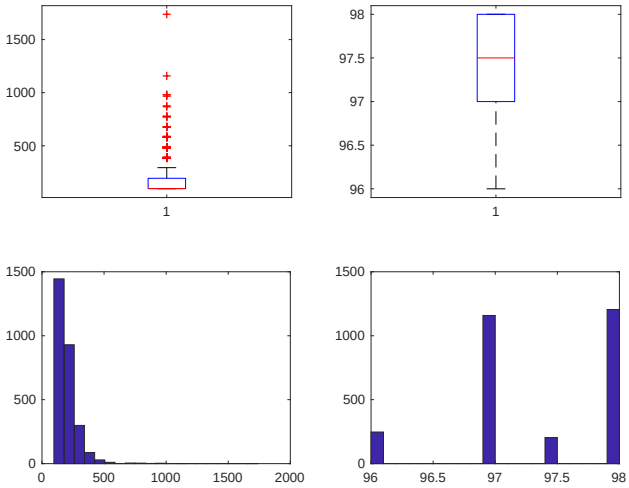

Figura 2. Superficie Pareti Orizzontali

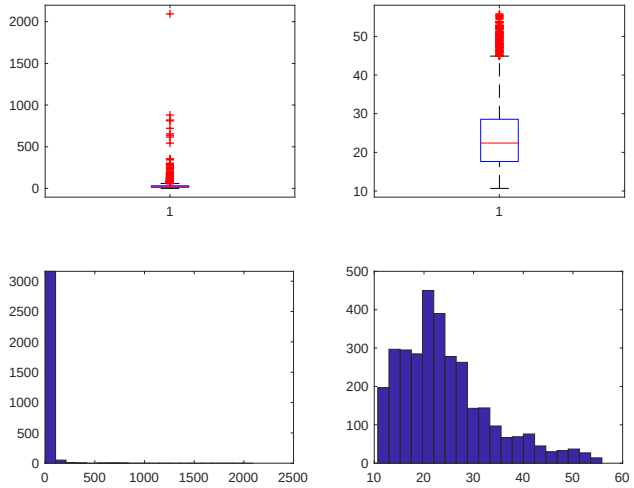

Figura 3. Superficie infissi

Dopo lo studio delle variabili riguardanti le superfici per le singole tipologie la procedura prevede la divisione in 3 sottopopolazioni, sfruttando le informazioni derivanti dagli indicatori precedentemente creati:

- INTERVENTO MULTIPLO (MULT) -3621 casi-,
- INTERVENTO SINGOLO CLIMATIZZAZIONE INVERNALE (SINGCI) -263 casi-,
- INTERVENTO SINGOLO PARETI VERTICALI O PARETI ORIZZANTI O INCLIATE O INFISSI (SINGALTRO) -168 casi- e
- INTERVENTO SENZA SPECIFICA (ZERO) -212 casi-.

Per ogni sottopopolazione sono state prese in considerazione le variabili RISPARMIO, COSTO (Costo intervento + Costo professionale), COSTO/RISPARMIO per verificare ulteriormente eventuali casi anomali sulla variabile COSTO e DETRAZIONE.

Tutte le variabili sono state studiate dopo averle normalizzate per "numero di unità immobiliare". Nel caso SINGALTRO, le variabili sono state studiate normalizzandole per la variabile superficie. Per le variabili RISPARMIO, COSTO e COSTO/RISPARMIO, i dati anomali sono stati individuati ed imputati tramite due programmi MATLAB simili a quello utilizzato per la variabile "superficie" in modo da rispettare sia la natura delle variabili stesse che lo scopo finale dell'analisi.
I risultati più immediati sono:

- RISPARMIO: 946 imputazioni (MULT), 119 imputazioni (SINGCI), 73 imputazioni (SINGALTRO) e 115 (ZERO);

- COSTO: 292 imputazioni (MULT), 75 imputazioni (SINGCI), 51 imputazioni (SINGALTRO) e 108 (ZERO);

- COSTO/RISPARMIO per imputare COSTO: 823 imputazioni (MULT), 112 imputazioni (SINGCI), 70 imputazioni (SINGALTRO) e 112 (ZERO).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

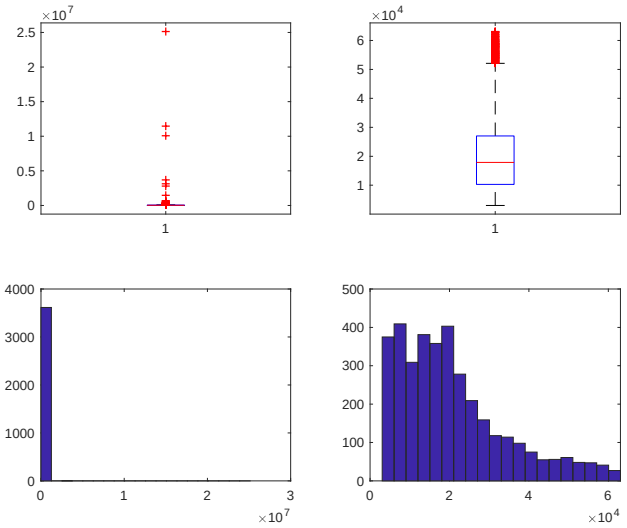

Figura 4. Risparmio MULT

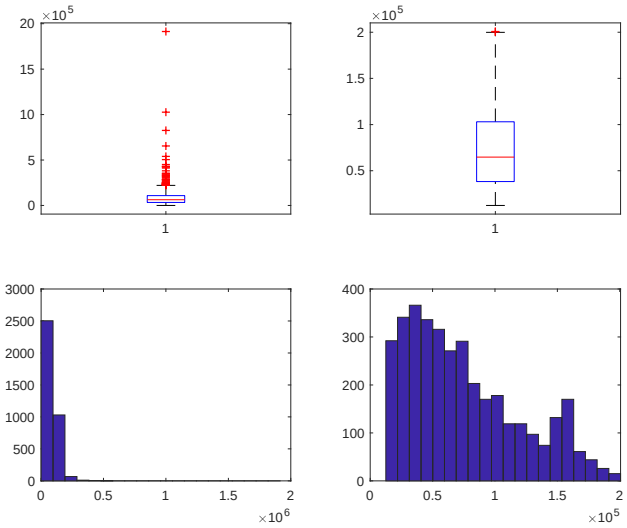

Figura 5. Costo MULT

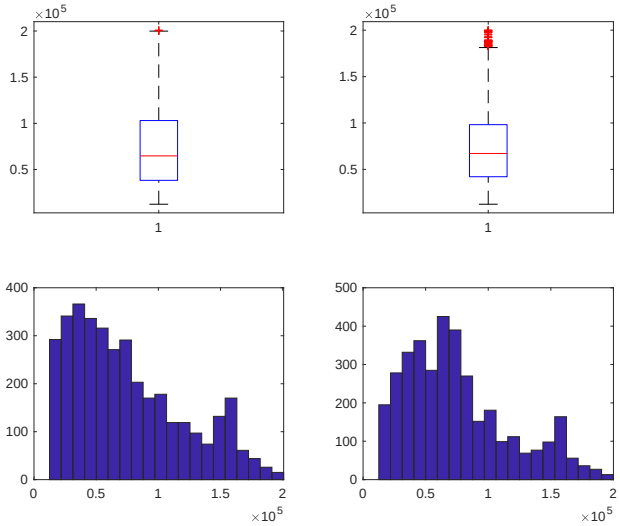

Figura 6. Costo dopo studio su Costo/Risparmio MULT

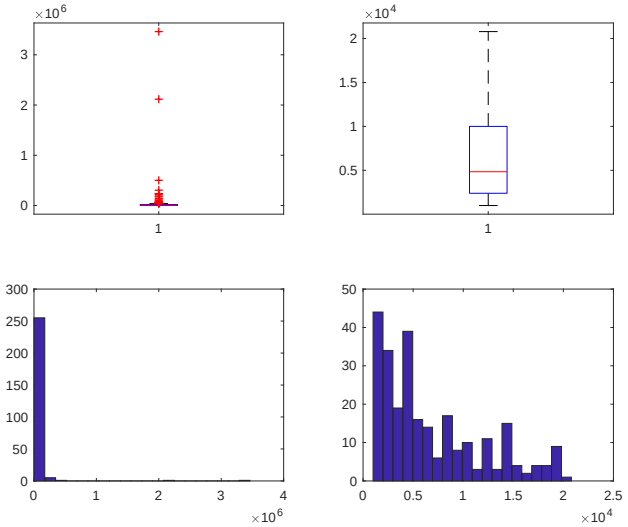

Figura 7. Risparmio SINGCI

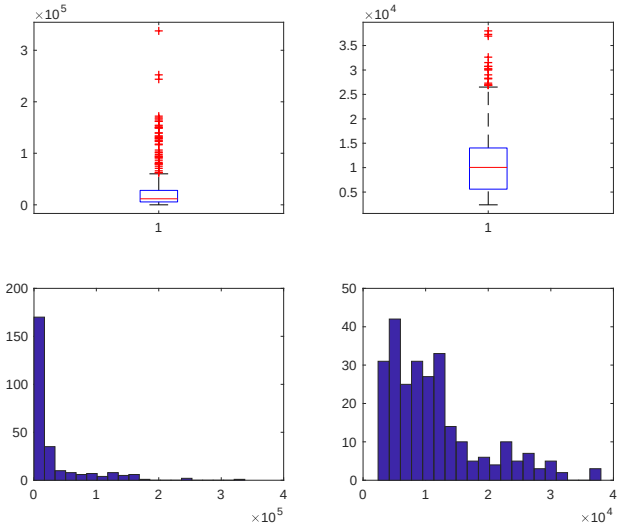

Figura 8. Costo SINGCI

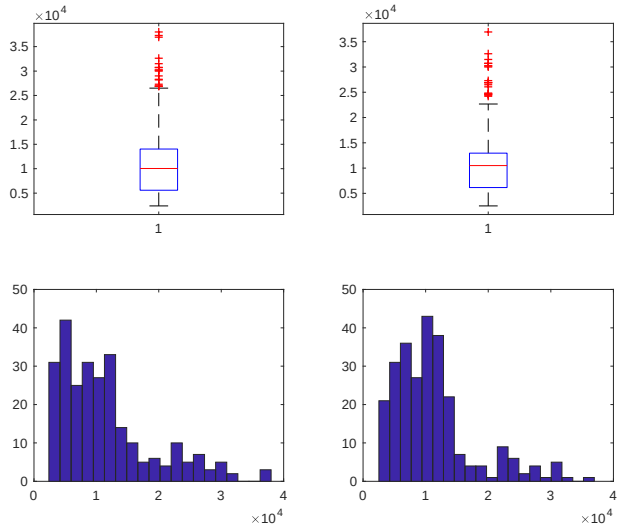

Figura 9. Costo dopo studio su Costo/Risparmio SINGCI

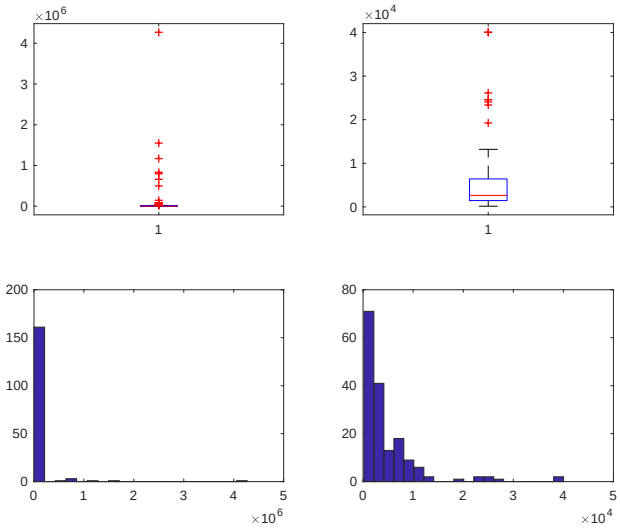

Figura 10. Risparmio SINGALTRO

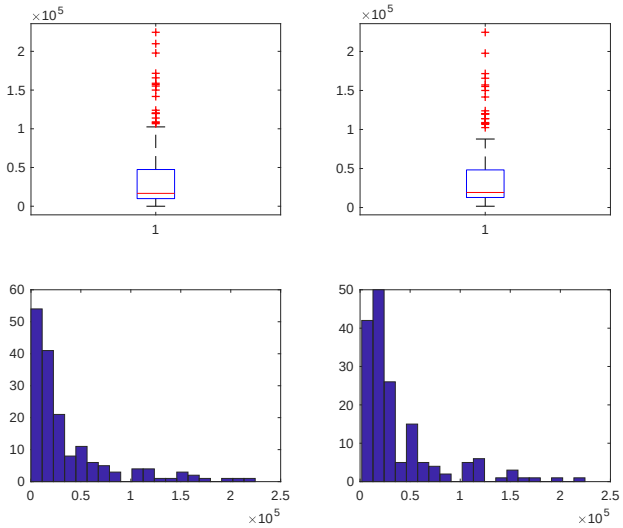

Figura 11. Costo SINGALTRO

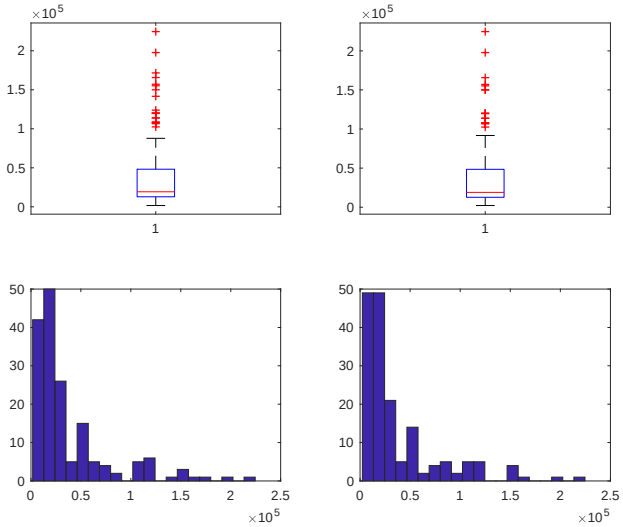

Figura 12: Costo dopo studio su Costo/Risparmio SINGALTRO

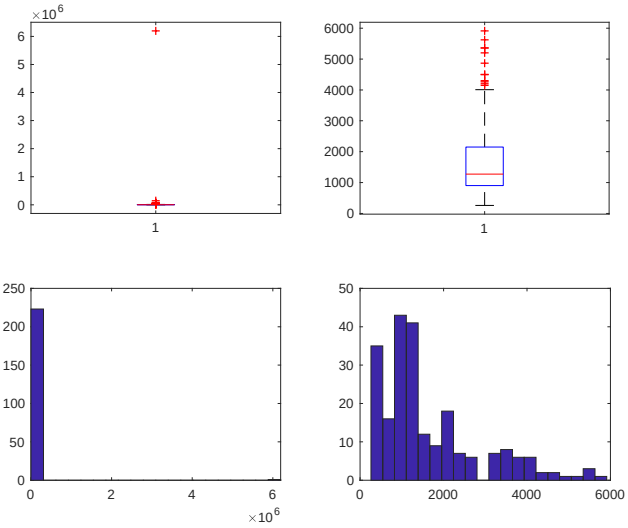

Figura 13. Risparmio ZERO

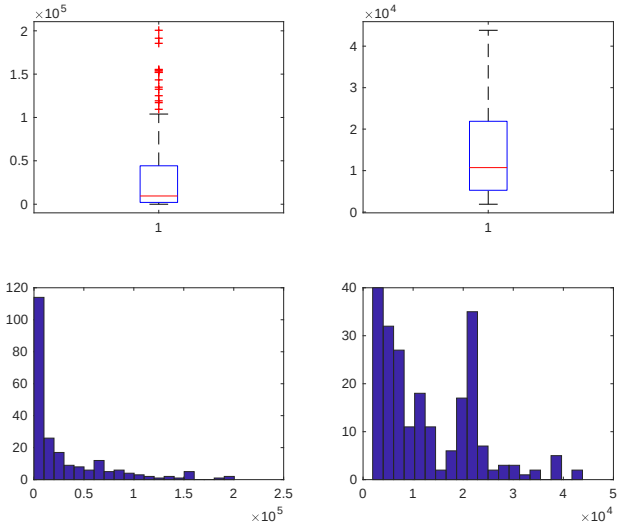

Figura 14. Costo ZERO

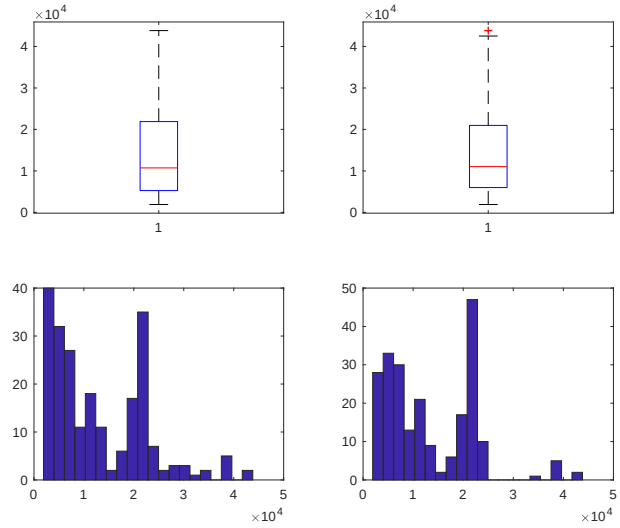

Figura 15: Costo dopo studio su Costo/Risparmio ZERO

La variabile DETRAZIONE è stata imputata secondo la regola:

detrazione= 0.65*costo

e sono stati individuate e conseguentemente imputati:

- 390 dati anomali (MULT) di cui 358 per imputazioni svolte su COSTO;
- 85 dati anomali (SINGCI) di cui 85 per imputazioni svolte su COSTO;
- 37 dati anomali (SINGALTRO) di cui 36 per imputazioni svolte su COSTO;
- 81 dati anomali (SINGZERO) di cui 80 per imputazioni svolte su COSTO.

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione risultate per la variabile Detrazione.

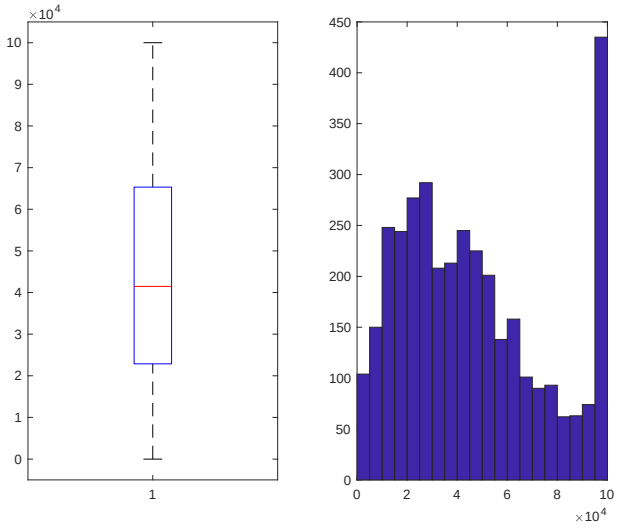

Figura 16. Detrazione MULT

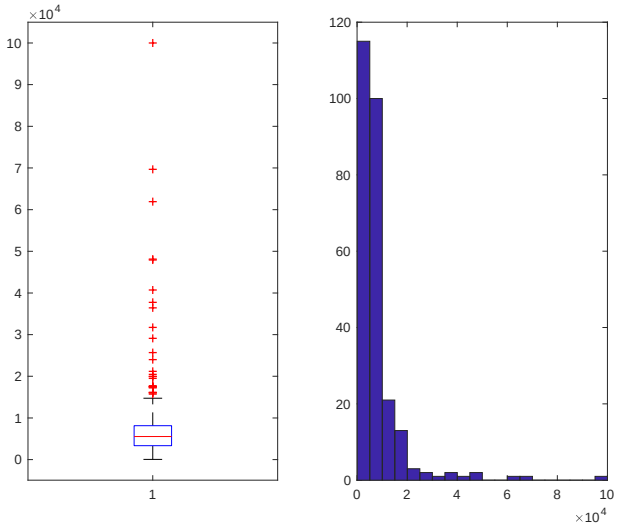

Figura 17. Detrazione SINGCI

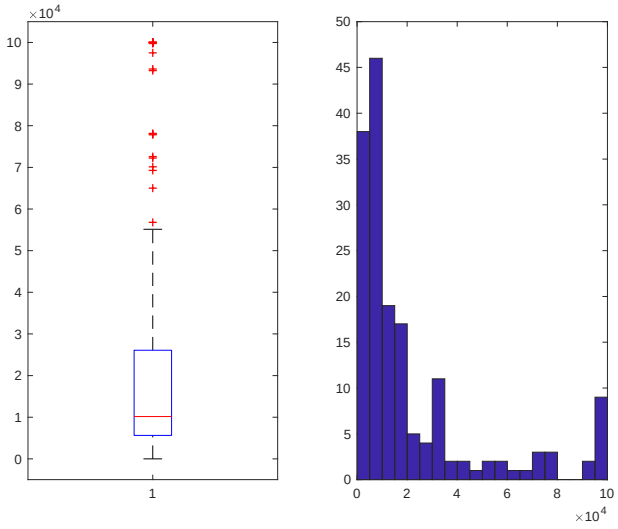

Figura 18. Detrazione SINGALTRO

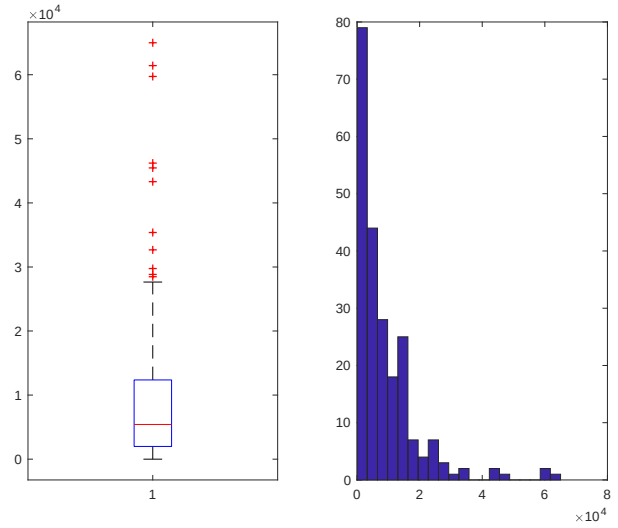

Figura 19. Detrazione ZERO

Una rapida considerazione finale arriva dal confronto della somma iniziale del Risparmio e del Costo con la somma delle medesime variabili dopo tutte le imputazioni svolte:

- Risparmio Iniziale: 189890947.746773 kWh/anno
- Risparmio Finale: 90956886.6843102 kWh/anno
- Costo Iniziale: 326322805.288999 €
- Costo Finale: 351640050.106050 €

1.4.2 Comma 345a

Dopo la fase di pulizia e ricodifica del database, la procedura di individuazione e correzione dei dati mancanti e dei dati anomali ha visto la creazione di cinque programmi MATLAB, uno che è alla base di tutto lo studio e da dove vengono richiamati gli altri tre che invece si occupano della fase di imputazione (ognuno di essi in base alla natura della variabile studiata).
E' importante rimarcare come ogni valutazione sulle variabili è stato preceduto da un'opportuna fase di normalizzazione per altre variabili di supporto allo studio.
Per questo comma è stato necessario studiare la superficie in due step separati, prima la superficie totale usando la variabile Destinazione d'uso generale per individuare le sottopopolazioni (notare che per la popolazione NON RESIDENZIALE si è studiato il Volume e non la Superficie) e poi la superficie specifica dell'intervento utilizzando il dettaglio della tipologia dell'intervento per individuare le sottopopolazioni di riferimento.

I primi risultati ottenuti sono:

- 3054 imputazioni sulla variabile "superficie totale" (numerosità 20171 casi);
- 1793 imputazioni sulla variabile "volume" (numerosità 3392 casi).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

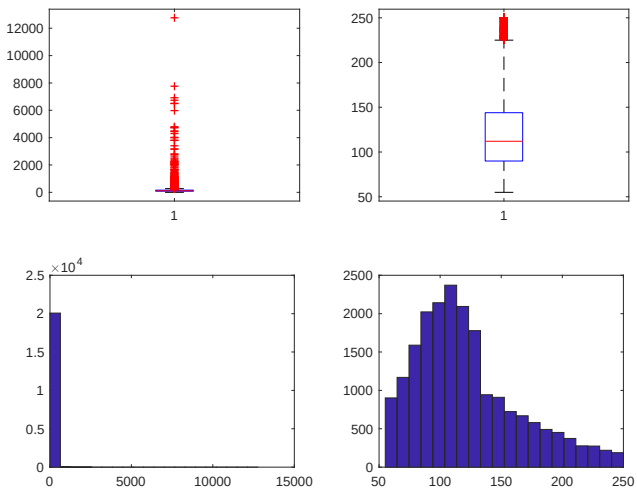

Figura 20. Superficie Residenziale

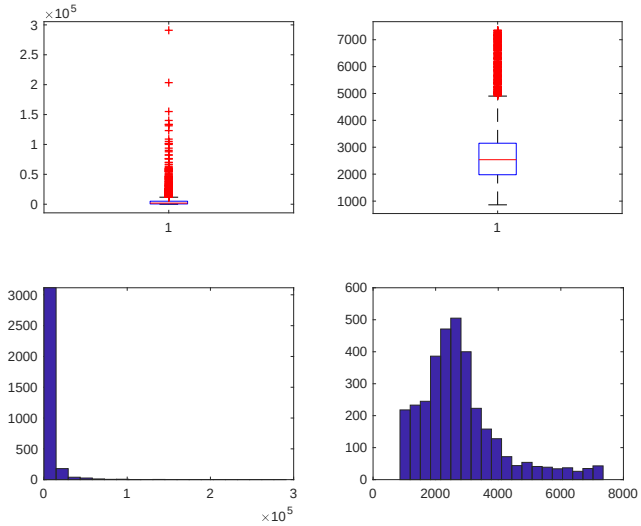

Figura 21. Volume Non Residenziale

Dopo lo studio delle variabili riguardanti le superfici per le singole tipologie la procedura prevede la divisione in 4 sottopopolazioni all'interno delle sottopopolazioni individuate precedentemente: INTERVENTO MULTIPLO (MULT), INTERVENTO SINGOLO PARETI VERTICALI (PV), INTERVENTO SINGOLO PARETI ORIZZONTALI O INCLINATE (POI) e INTERVENTO SINGOLO INFISSI (I).
Prima di procedere all'analisi delle variabili d'interesse per lo studio è stato necessario studiare in maniera dettagliata la superficie specifica dell'intervento in modo da normalizzare in maniera più precisa le variabili d'interesse. Questa operazione è stata di grande complessità ed è risultata molto onerosa anche da un punto di vista computazionale.
Si evidenziano qui solamente le numerosità delle sottopopolazioni, che verranno utilizzate poi anche per le variabili d'interesse.

- MULT (numerosità 627 casi);
- PV (numerosità 418 casi);
- POI (numerosità 1279 casi);
- I (numerosità 42 casi).

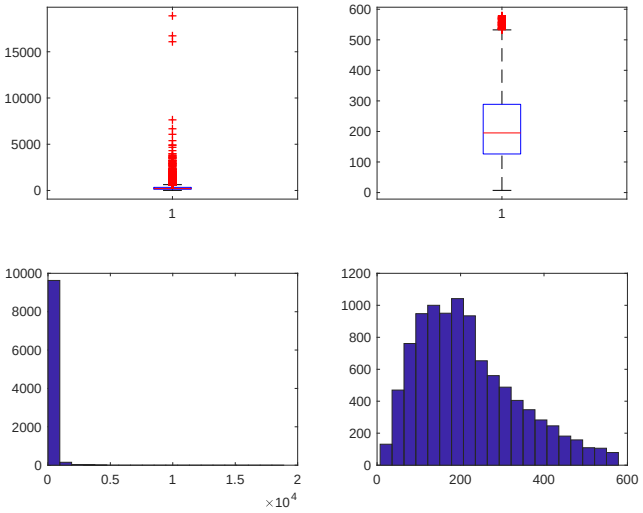

Figura 22. Superficie Specifica MULT

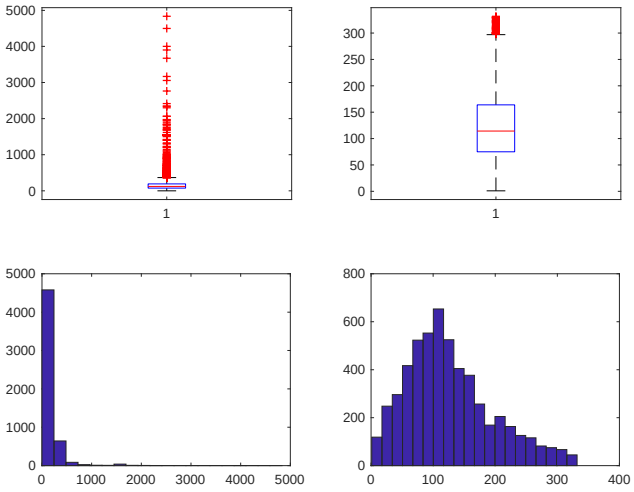

Figura 23. Superficie Specifica PV

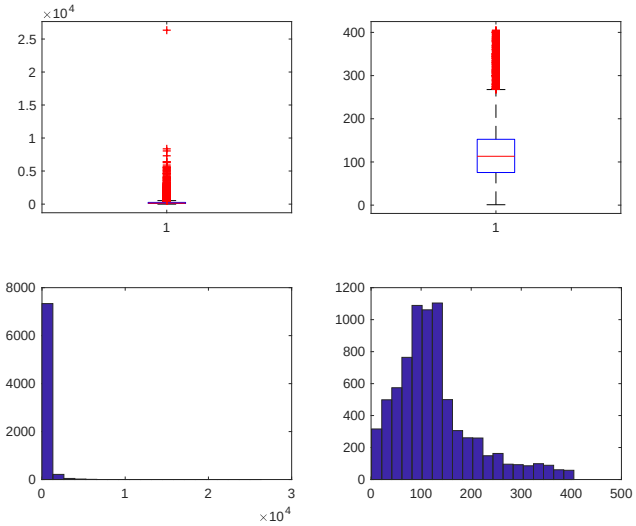

Figura 24. Superficie Specifica POI

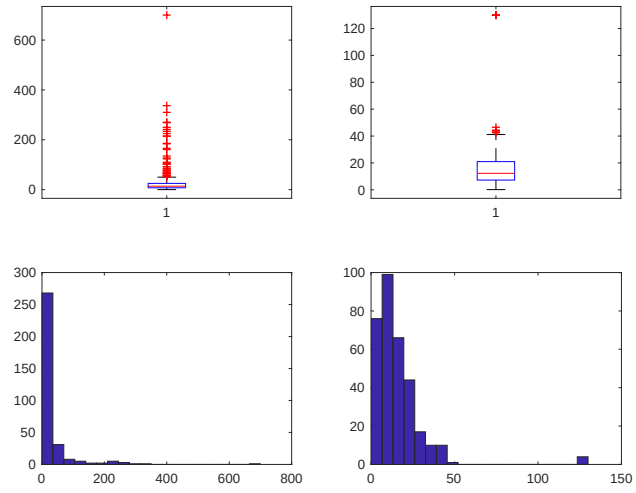

Figura 25. Superficie Specifica I

Per ogni sottopopolazione sono state prese in considerazione le variabili RISPARMIO, COSTO (Costo intervento + Costo professionale), COSTO/RISPARMIO per verificare ulteriormente eventuali casi anomali sulla variabile COSTO e DETRAZIONE.
Tutte le variabili sono state studiate dopo averle normalizzate per "numero di unità immobiliare".
Nel caso SING, le variabili sono state studiate normalizzandole per la variabile superficie.
Per le variabili RISPARMIO, COSTO e COSTO/RISPARMIO, i dati anomali sono stati individuati ed imputati tramite due programmi MATLAB simili a quelli utilizzati per la variabile "superficie" e per la variabile "superficie specifica" in modo da rispettare sia la natura delle variabili stesse che lo scopo finale dell'analisi.
I risultati più immediati sono:
Sottopopolazione RESIDENZIALE:

- RISPARMIO: 1483 imputazioni (MULT), 984 imputazioni (PV) 1497 imputazioni (POI) e 109 imputazioni (I);
- COSTO: 1483 imputazioni (MULT), 984 imputazioni (PV-RES), 1497 imputazioni (POI-RES) e 109 imputazioni (I-RES);
- COSTO/RISPARMIO per imputare COSTO: 1618 imputazioni (MULT), 957 imputazioni (PV), 1688 imputazioni (POI) e 104 imputazioni (I).

Sottopopolazione NON RESIDENZIALE:

- RISPARMIO: 780 imputazioni (MULT), 984 imputazioni (PV), 1497 imputazioni (POI) e 109 imputazioni (I);
- COSTO: 1483 imputazioni (MULT), 984 imputazioni (PV), 780 imputazioni (POI) e 22 imputazioni (I);
- COSTO/RISPARMIO per imputare COSTO: 228 imputazioni (MULT), 95 imputazioni (PV), 569 imputazioni (POI) e 19 imputazioni (I).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

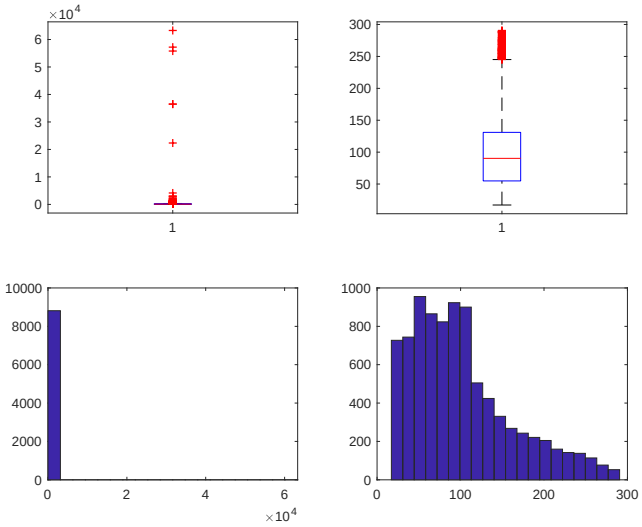

Figura 26. Risparmio RESMULT

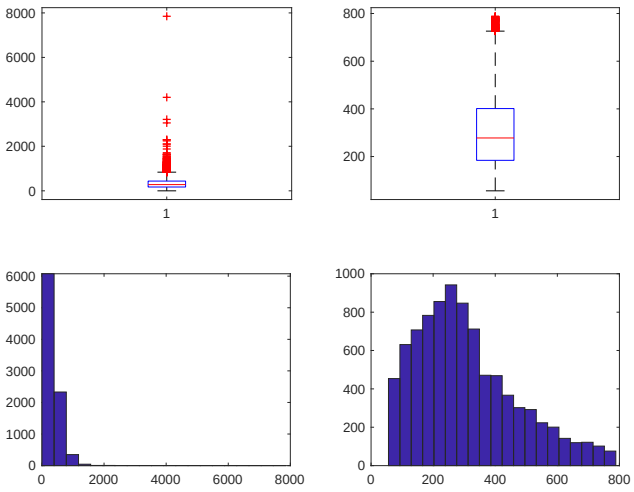

Figura 27. Costo RESMULT

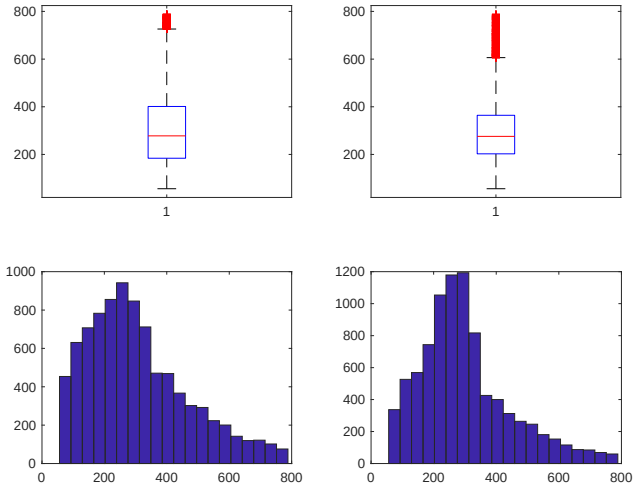

Figura 28. Costo dopo studio su Costo/Risparmio RESMULT

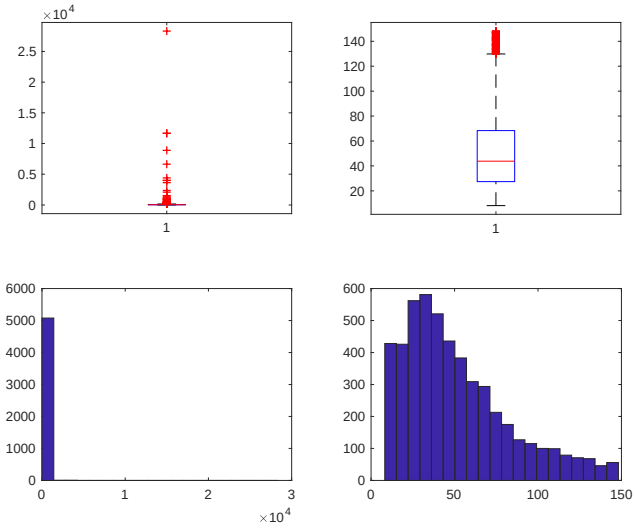

Figura 29. Risparmio RESPV

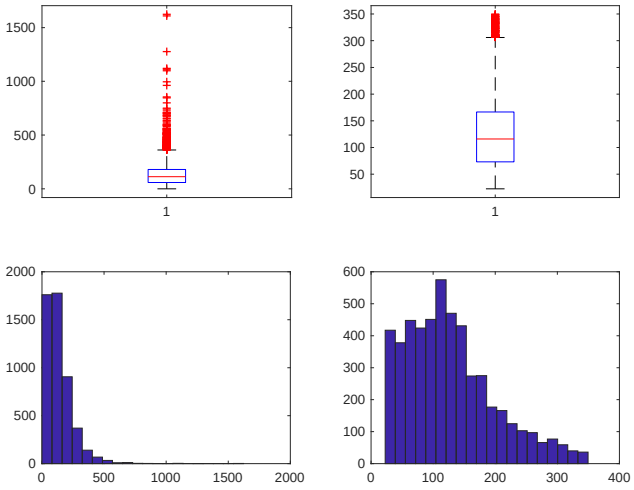

Figura 30. Costo RESPV

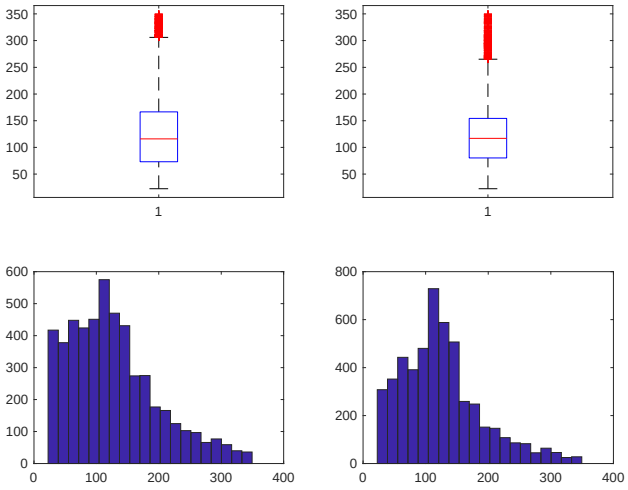

Figura 31. Costo dopo studio su Costo/Risparmio RESPV

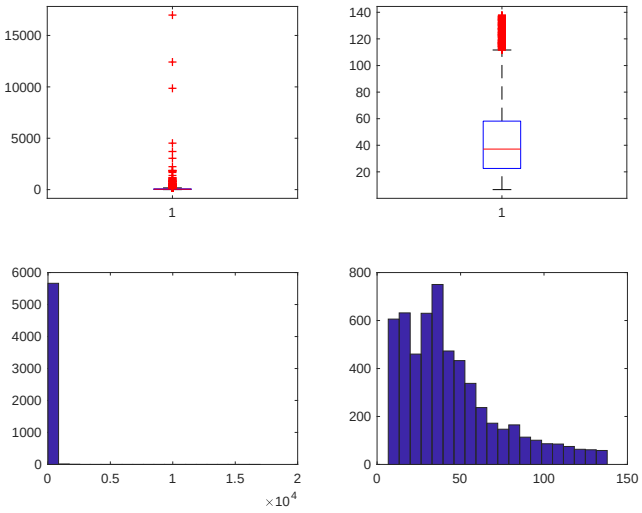

Figura 32. Risparmio RESPOI

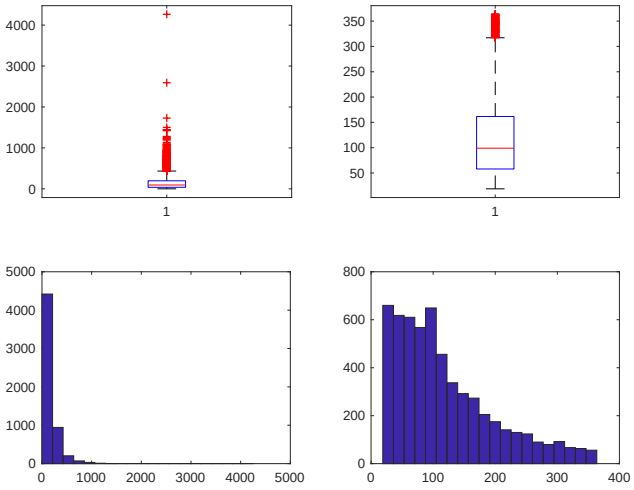

Figura 33. Costo RESPOI

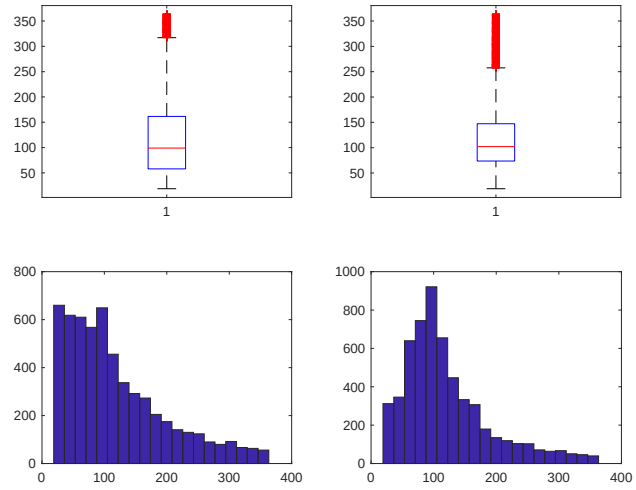

Figura 34. Costo dopo studio su Costo/Risparmio RESPOI

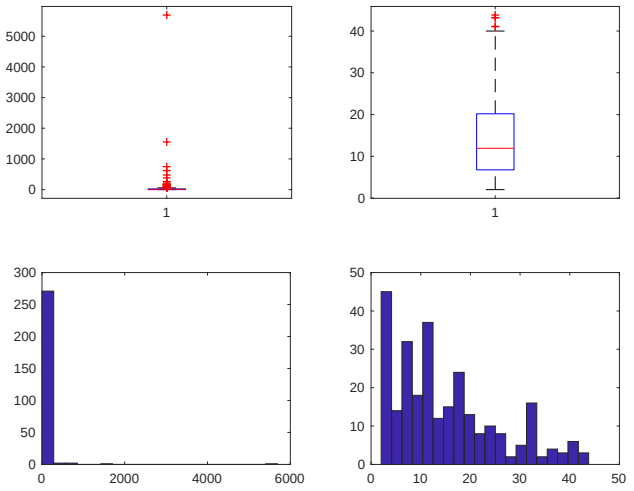

Figura 35. Risparmio RESI

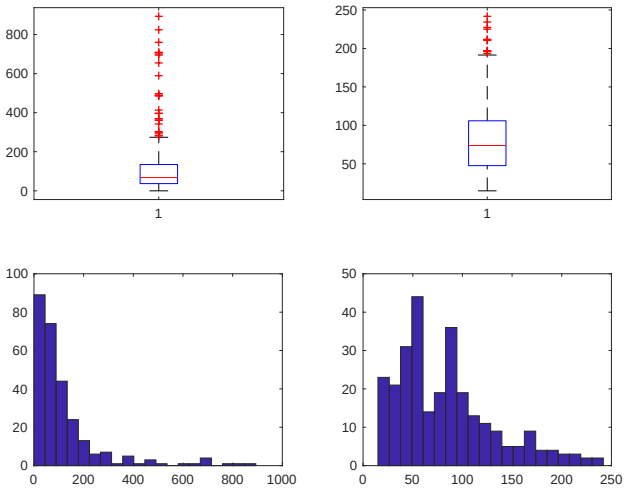

Figura 36. Costo RESI

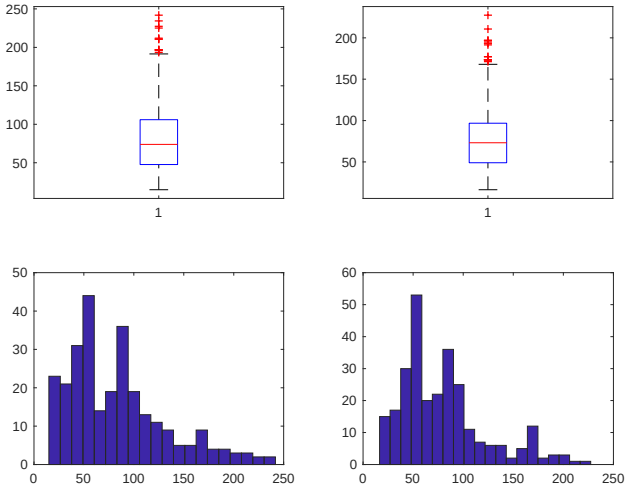

Figura 37. Costo dopo studio su Costo/Risparmio RESI

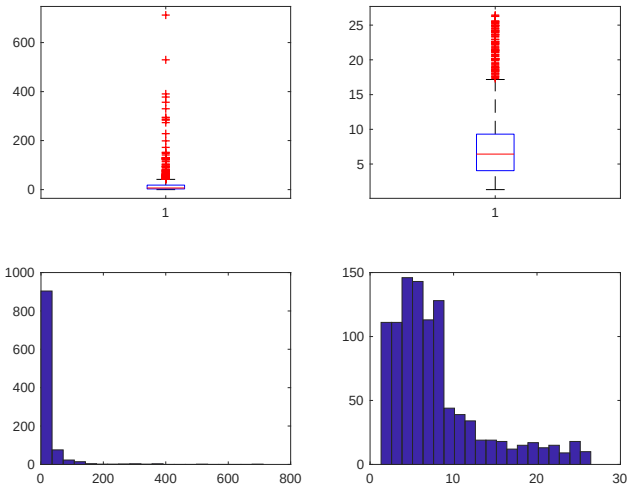

Figura 38. Risparmio NORESMULT

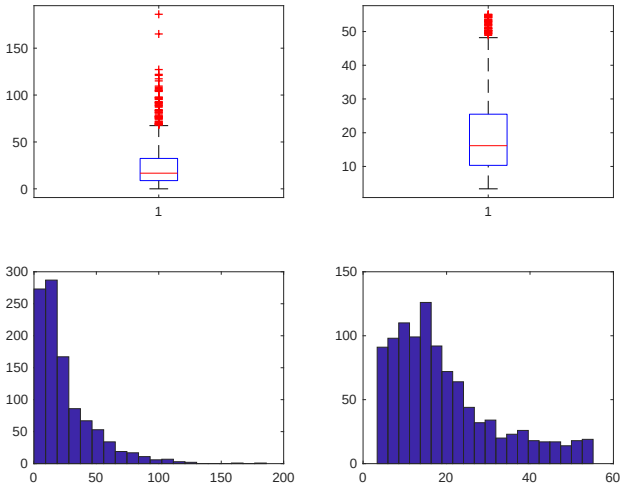

Figura 39. Costo NORESMULT

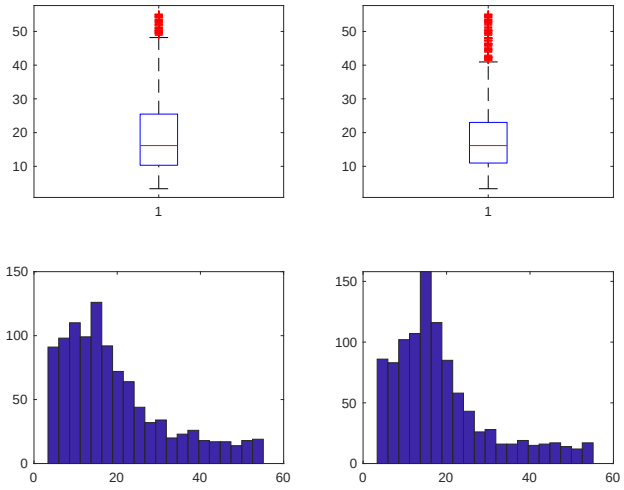

Figura 40. Costo dopo studio su Costo/Risparmio NORESMULT

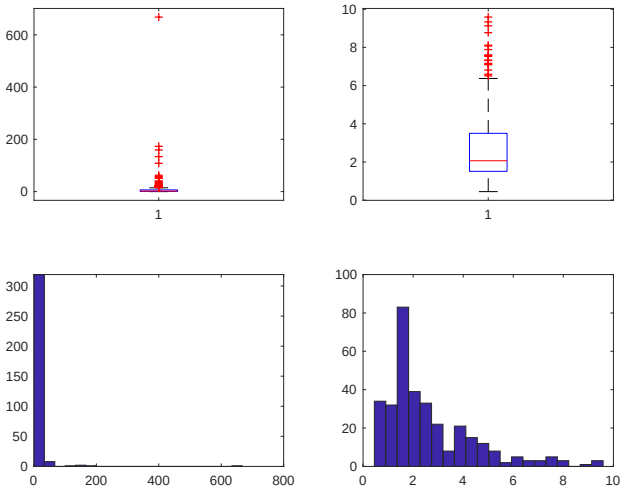

Figura 41. Risparmio NORESPV

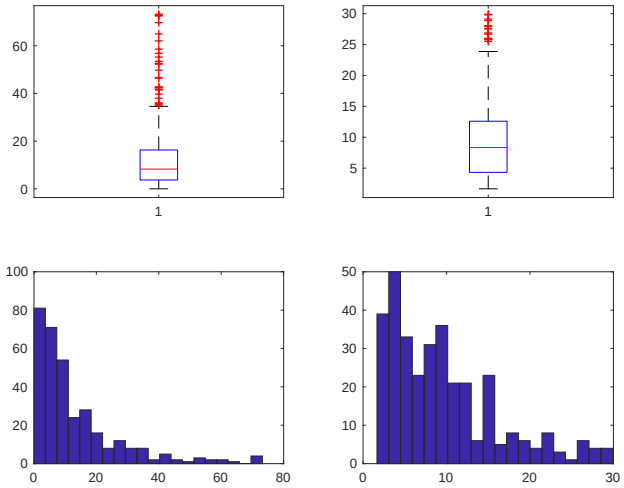

Figura 42. Costo NORESPV

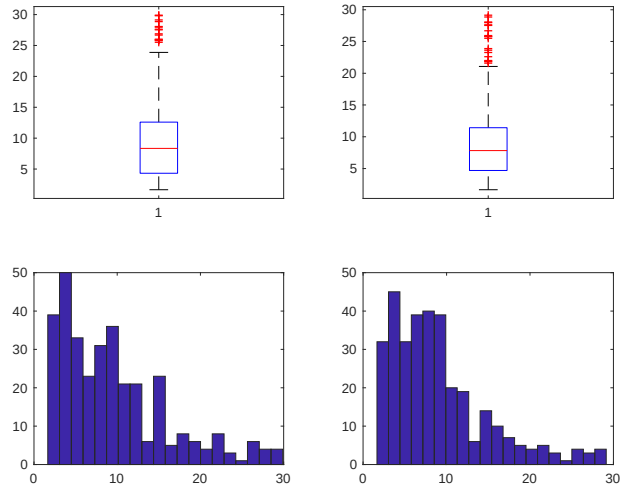

Figura 43. Costo dopo studio su Costo/Risparmio NORESPV

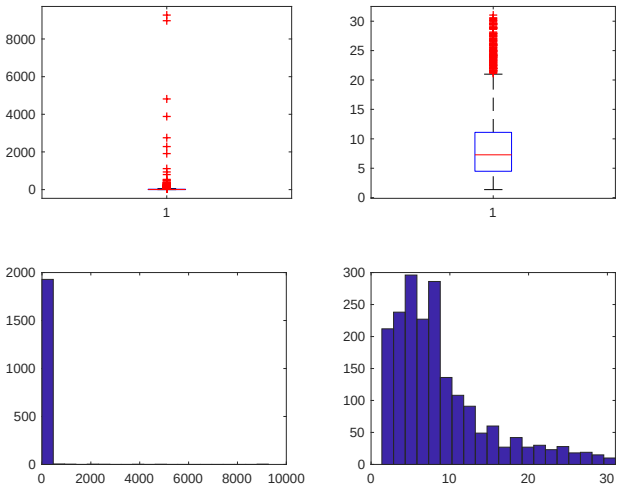

Figura 44. Risparmio NORESPOI

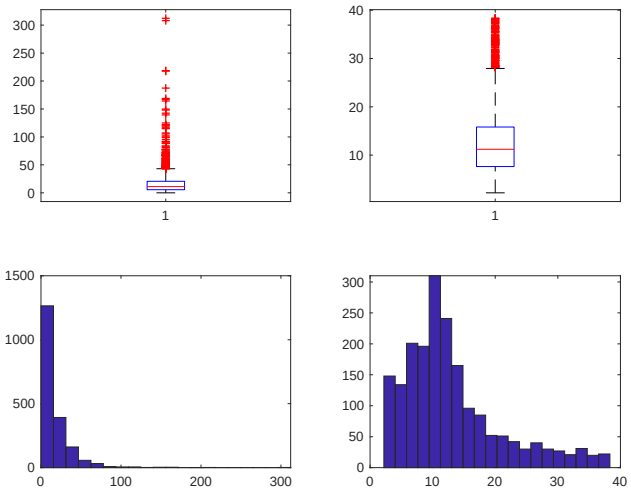

Figura 45. Costo NORESPOI

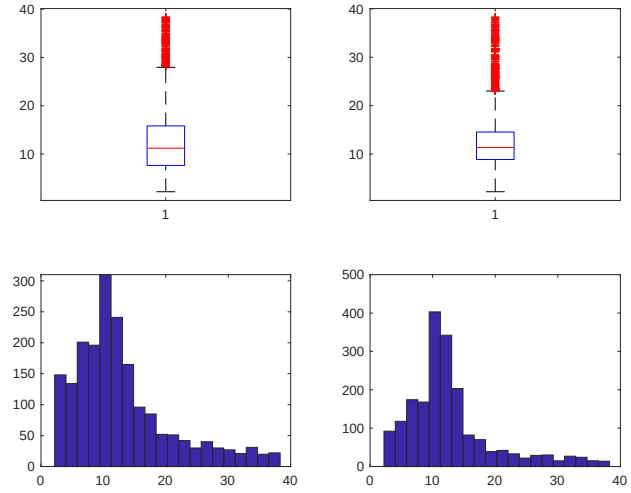

Figura 46. Costo dopo studio su Costo/Risparmio NORESPOI

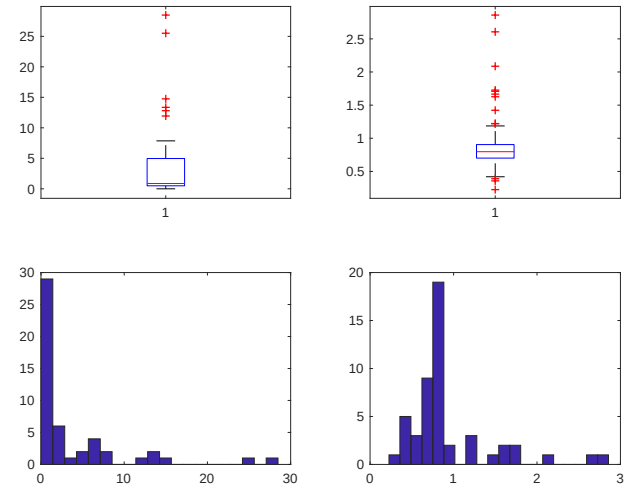

Figura 47. Risparmio NORESI

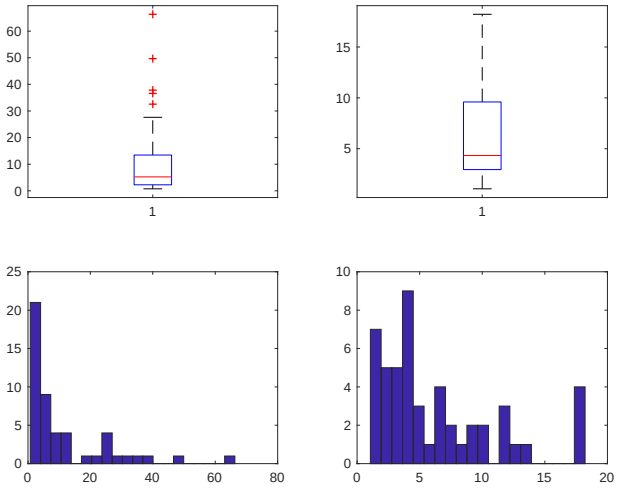

Figura 48. Costo NORESI

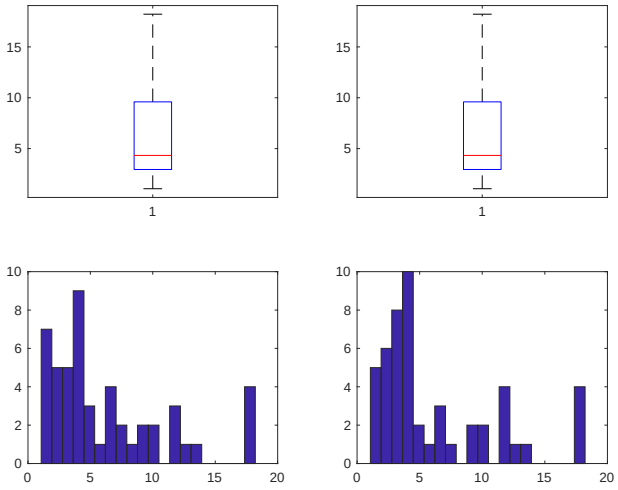

Figura 49. Costo dopo studio su Costo/Risparmio NORESI

Per questo comma è stato necessario rivalutare anche la variabile Costo sia direttamente che tramite la variabile Costo/Risparmio ai fini di correggere eventuali errori ancora presenti.
I risultati sono:

- COSTO: 972 imputazioni (MULT), 672 imputazioni (PV), 1514 imputazioni (POI), 300 imputazioni (I);
- COSTO/RISPARMIO per imputare COSTO: 1862 imputazioni (MULT), 1325 imputazioni (PV), 2378 imputazioni (POI), 126 imputazioni (I).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione della variabile Costo per le 4 sottopopolazioni cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

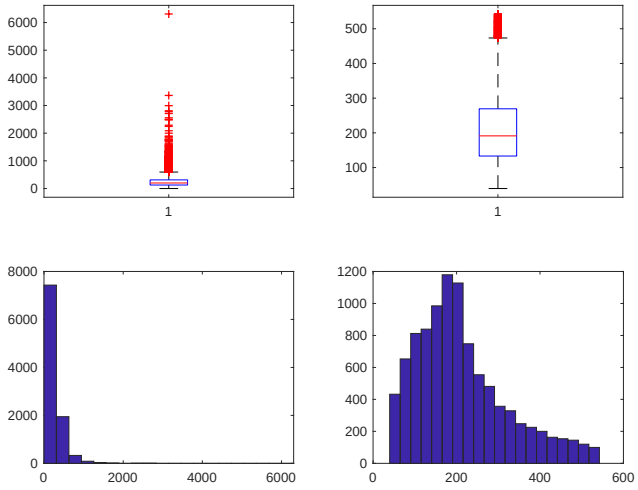

Figura 50. Costo MULT

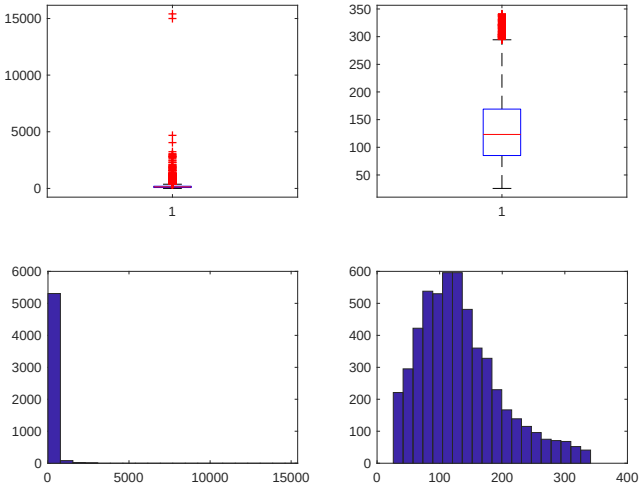

Figura 51. Costo PV

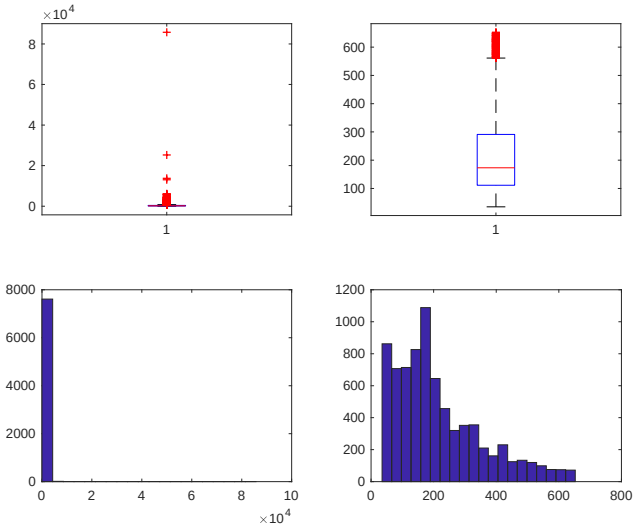

Figura 52. Costo POI

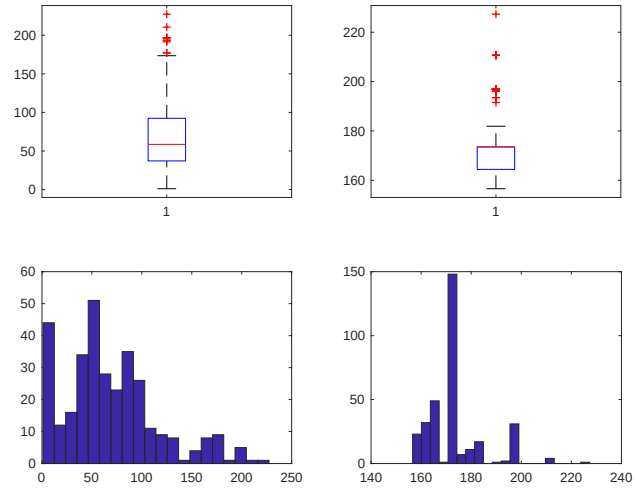

Figura 53. Costo I

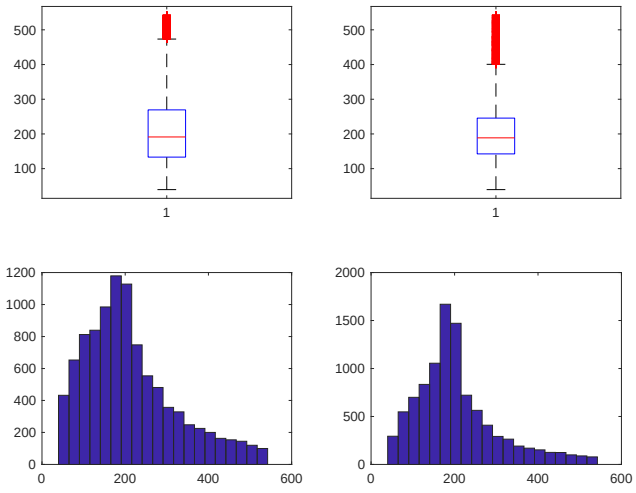

Figura 54. Costo dopo studio su Costo/Risparmio MULT

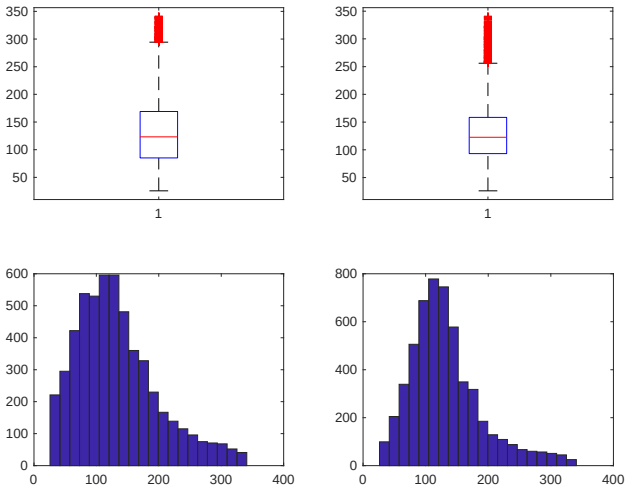

Figura 55. Costo dopo studio su Costo/Risparmio PV

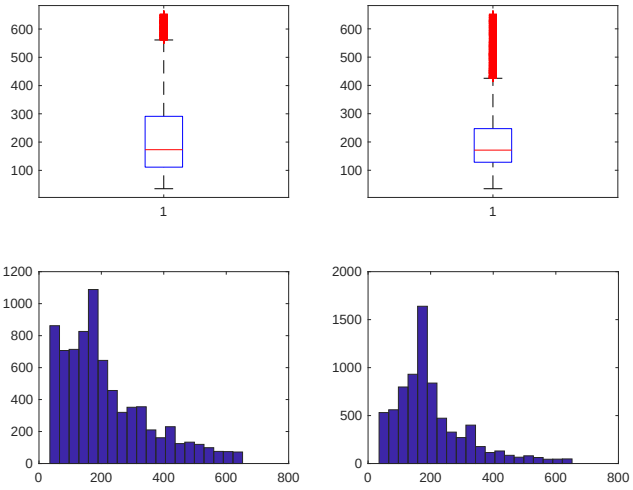

Figura 56. Costo dopo studio su Costo/Risparmio POI

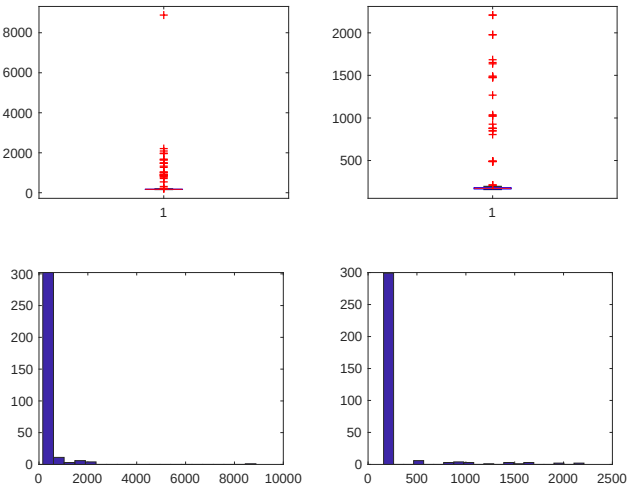

Figura 57. Costo dopo studio su Costo/Risparmio I

La variabile DETRAZIONE è stata imputata secondo la regola:

detrazione= 0.65*costo

e sono stati individuate e conseguentemente imputati:

Sottopopolazione RESIDENZIALE:

- 1138 dati anomali (MULT) di cui 1078 per imputazioni svolte su COSTO;
- 605 dati anomali (PV) di cui 573 per imputazioni svolte su COSTO;
- 960 dati anomali (POI) di cui 915 per imputazioni svolte su COSTO;
- 68 dati anomali (I) di cui 64 per imputazioni svolte su COSTO.

Sottopopolazione NON RESIDENZIALE:

- 211 dati anomali (MULT) di cui 206 per imputazioni svolte su COSTO;
- 75 dati anomali (PV) di cui 72 per imputazioni svolte su COSTO;
- 387 dati anomali (POI) di cui 376 per imputazioni svolte su COSTO;
- 16 dati anomali (I) di cui 16 per imputazioni svolte su COSTO.

Si noti che prima di arrivare a queste distribuzioni della variabile Detrazione, si è passati per una fase di studio intermedia che ha visto protagonista il calcolo delle detrazioni per le sottopopolazioni più specifiche e i risultati qui riportati sono quelli finali. Il processo di controllo e "taratura" per questo comma è stato molto articolato in quanto le variabili risultavano essere complesse da studiare per via di asimmetrie molto forti.

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

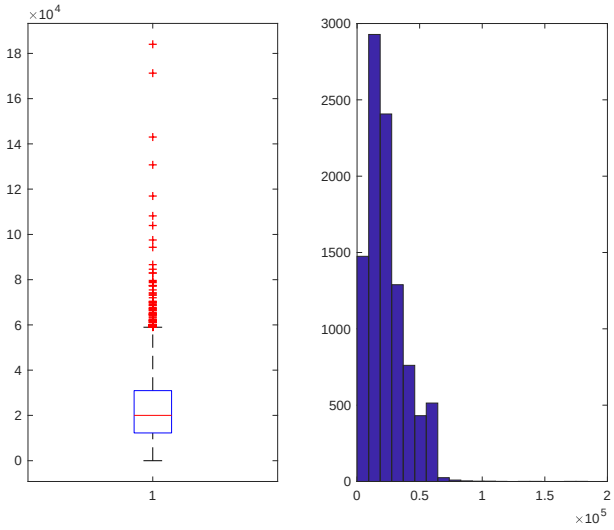

Figura 58. Detrazione MULT

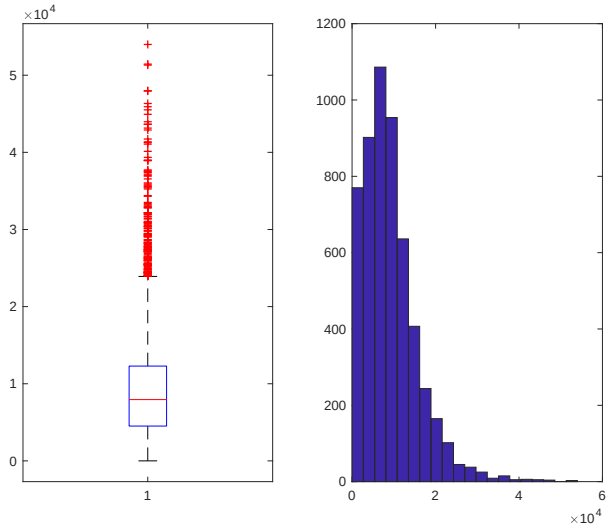

Figura 59. Detrazione PV

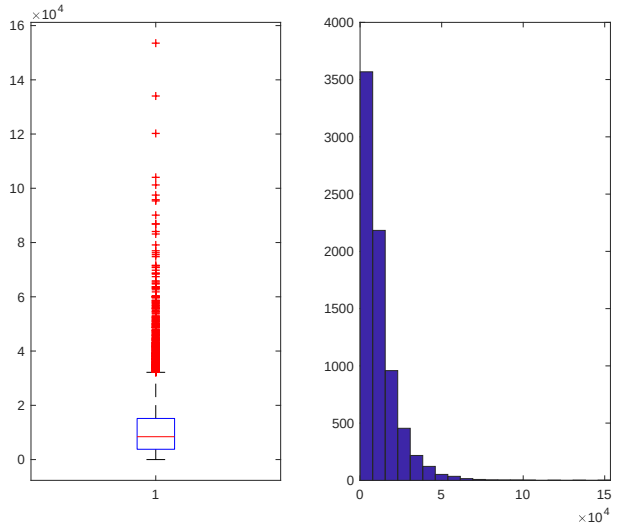

Figura 60. Detrazione POI

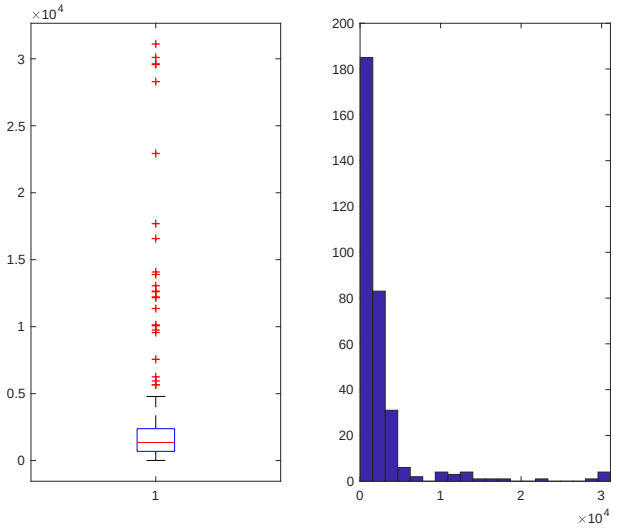

Figura 61. Detrazione I

Una rapida considerazione finale arriva dal confronto della somma iniziale del Risparmio e del Costo con la somma delle medesime variabili dopo tutte le imputazioni svolte:

- Risparmio Iniziale: 602472589.660000 kWh/anno
- Risparmio Finale: 429915065.411376 kWh/anno
- Costo Iniziale: 818220880.665000 €
- Costo Finale: 876822217.193749 €

1.4.3 Comma 345b

Dopo la fase di pulizia e ricodifica del database, la procedura di individuazione e correzione dei dati mancanti e dei dati anomali ha visto la creazione di quattro programmi MATLAB, uno che è alla base di tutto lo studio e da dove vengono richiamati gli altri quattro che invece si occupano della fase di imputazione (ognuno di essi in base alla natura della variabile studiata).

Anche in questo comma la prima operazione svolta è stata la creazione di indicatori volti ad individuare sottopopolazioni e la creazione di variabili utili allo studio.
La prima fase di imputazione è stata svolta sulla variabile superficie, il primo risultato importante è:

- 26 imputazioni per la sottopopolazione "Telaio legno esistente dopo l'intervento e vetro singolo esistente dopo l'intervento" - di seguito denominato "1;1" - (numerosità 203 casi),
- 913 imputazioni per la sottopopolazione "Telaio legno esistente dopo l'intervento e vetro doppio esistente dopo l'intervento" - di seguito denominato "1;2" - (numerosità 6690 casi),
- 317 imputazioni per la sottopopolazione "Telaio legno esistente dopo l'intervento e vetro triplo esistente dopo l'intervento" - di seguito denominato "1;3" - (numerosità 2006 casi),
- 2602 imputazioni per la sottopopolazione "Telaio legno esistente dopo l'intervento e vetro a bassa emissione esistente dopo l'intervento" - di seguito denominato "1;4" - (numerosità 18624 casi);
- 146 imputazioni per la sottopopolazione "Telaio legno esistente dopo l'intervento e vetro non esistente dopo l'intervento" - 15 - (numerosità 1270 casi),
- 3 imputazione per la sottopopolazione "Telaio Metallo, no taglio termico esistente dopo l'intervento e vetro singolo esistente dopo l'intervento" - denominato "2;1" - (numerosità 25 casi),
- 30 imputazioni per la sottopopolazione "Telaio Metallo, no taglio termico esistente dopo l'intervento e vetro doppio esistente dopo l'intervento" - denominato "2;2" - (numerosità 219 casi),
- 7 imputazione per la sottopopolazione "Telaio Metallo, no taglio termico esistente dopo l'intervento e vetro triplo esistente dopo l'intervento" - denominato "2;3" - (numerosità 27 casi),
- 58 imputazioni per la sottopopolazione "Telaio Metallo, no taglio termico esistente dopo l'intervento e vetro a bassa emissione esistente dopo l'intervento" - denominato "2;4" - (numerosità 369 casi);
- 153 imputazioni per la sottopopolazione "Telaio Metallo, no taglio termico esistente dopo l'intervento e vetro non esistente dopo l'intervento" - denominato "2;5" - (numerosità 1219 casi),
- 31 imputazioni per la sottopopolazione "Telaio Metallo, taglio termico esistente dopo l'intervento e vetro singolo esistente dopo l'intervento" - denominato "3;1" - (numerosità 191 casi),
- 1239 imputazioni per la sottopopolazione "Telaio Metallo, taglio termico esistente dopo l'intervento e vetro doppio esistente dopo l'intervento" - denominato "3;2" - (numerosità 8158 casi),
- 183 imputazioni per la sottopopolazione "Telaio Metallo, taglio termico esistente dopo l'intervento e vetro triplo esistente dopo l'intervento" - denominato "3;3" - (numerosità 1281 casi),
- 3503 imputazioni per la sottopopolazione "Telaio Metallo, taglio termico esistente dopo l'intervento e vetro a bassa emissione esistente dopo l'intervento" - denominato "3;4" - (numerosità 26409 casi);

- 287 imputazioni per la sottopopolazione “Telaio Metallo, taglio termico esistente dopo l’intervento e vetro non esistente dopo l’intervento” - denominato “3;5” - (numerosità 2314 casi),
- 45 imputazioni per la sottopopolazione “Telaio PVC esistente dopo l’intervento e vetro singolo esistente dopo l’intervento” - denominato “4;1” - (numerosità 369 casi),
- 3358 imputazioni per la sottopopolazione “Telaio PVC esistente dopo l’intervento e vetro doppio esistente dopo l’intervento” - denominato “4;2” - (numerosità 29418 casi),
- 731 imputazioni per la sottopopolazione “Telaio PVC esistente dopo l’intervento e vetro triplo esistente dopo l’intervento” - denominato “4;3” - (numerosità 6324 casi),
- 9743 imputazioni per la sottopopolazione “Telaio PVC esistente dopo l’intervento e vetro a bassa emissione esistente dopo l’intervento” - 4;4” - (numerosità 90128 casi);
- 60 imputazioni per la sottopopolazione “Telaio PVC esistente dopo l’intervento e vetro non esistente dopo l’intervento” - denominato “4;5” - (numerosità 556 casi),
- 10 imputazioni per la sottopopolazione “Telaio Misto esistente dopo l’intervento e vetro singolo esistente dopo l’intervento” - denominato “5;1” - (numerosità 75 casi),
- 313 imputazioni per la sottopopolazione “Telaio Misto esistente dopo l’intervento e vetro doppio esistente dopo l’intervento” - denominato “5;2” - (numerosità 2730 casi),
- 190 imputazioni per la sottopopolazione “Telaio Misto esistente dopo l’intervento e vetro triplo esistente dopo l’intervento” - denominato “5;3” - (numerosità 1770 casi),
- 1315 imputazioni per la sottopopolazione “Telaio Misto esistente dopo l’intervento e vetro a bassa emissione esistente dopo l’intervento” - denominato “5;4” - (numerosità 10377 casi);
- 760 mputazioni per la sottopopolazione “Telaio Misto esistente dopo l’intervento e vetro non esistente dopo l’intervento” - denominato “5;5” - (numerosità 6846 casi).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione della variabile “superficie” per le 4 sottopopolazioni cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

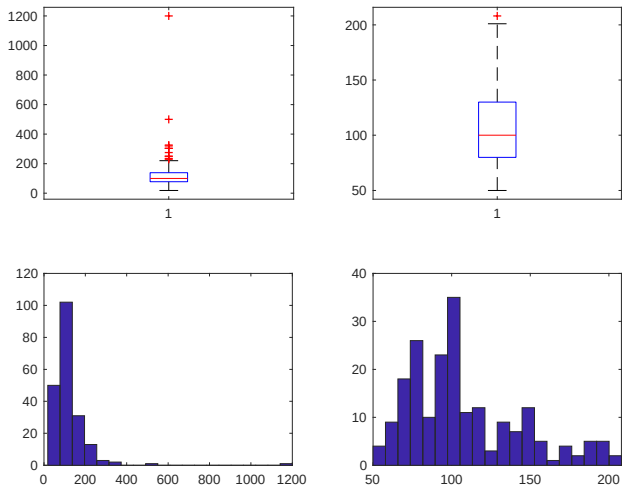

Figura 62. Superficie LEGNO-SINGOLO

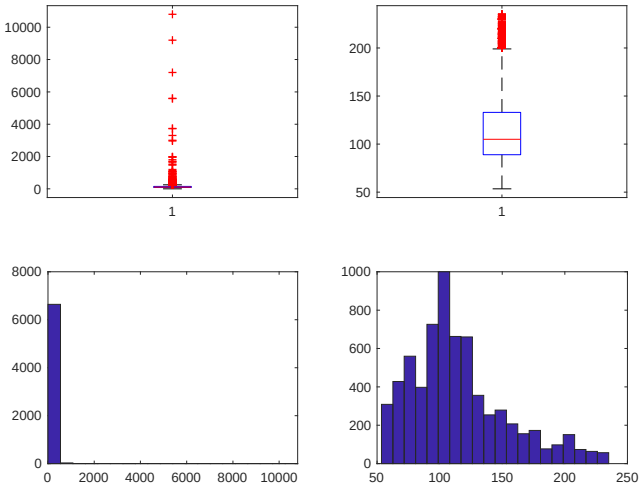

Figura 63. Superficie LEGNO-DOPPIO

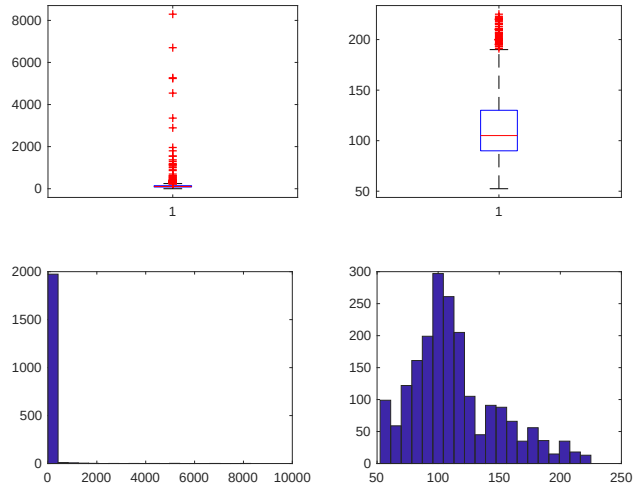

Figura 64. Superficie LEGNO-TRIPLO

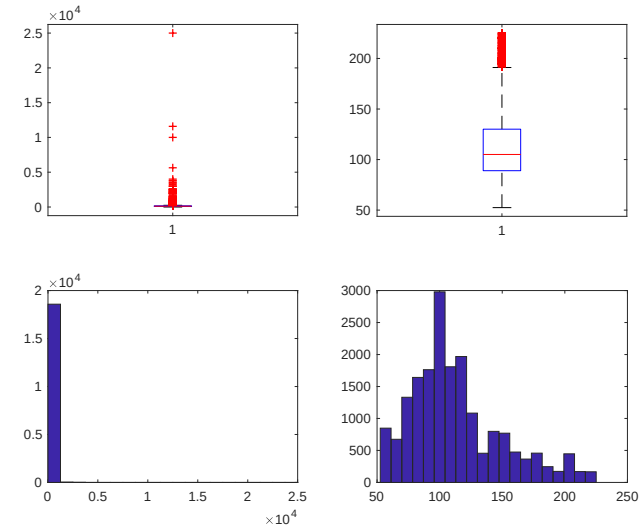

Figura 65. Superficie LEGNO-VETRO A BASSA EMISSIONE

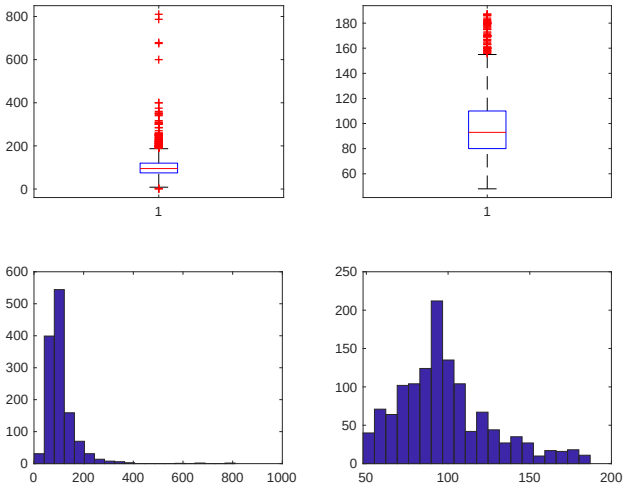

Figure 66. Superficie LEGNO-VETRO NON ESISTENTE

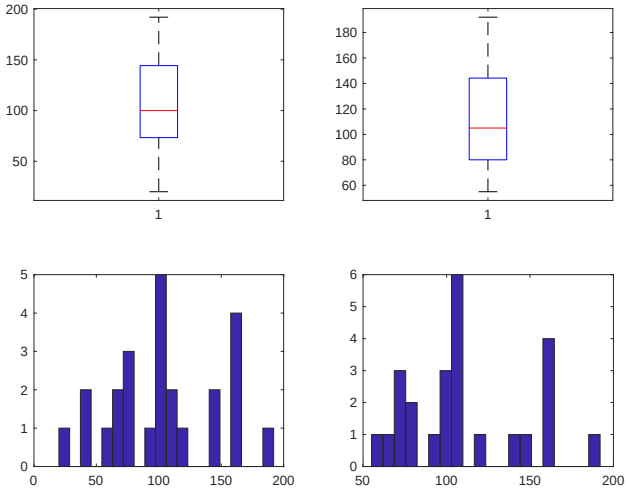

Figura 67. Superficie METALLO NO TERMICO-SINGOLO

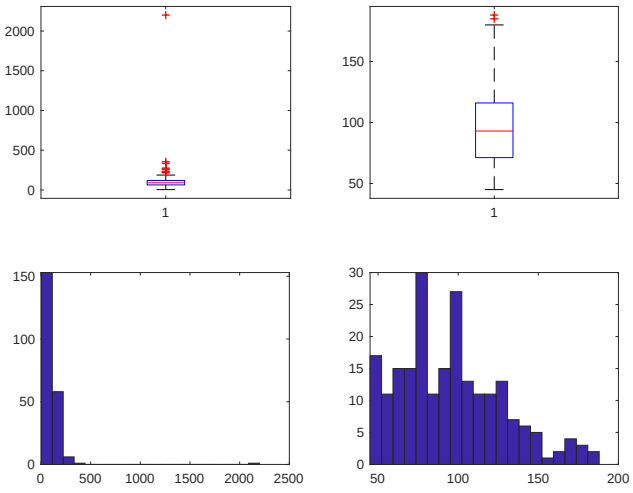

Figura 68. Superficie METALLO NO TERMICO-DOPPIO

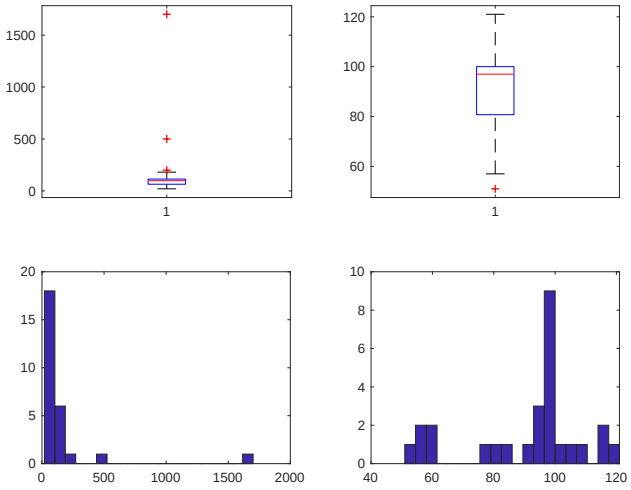

Figura 75. Superficie METALLO NO TERMICO-TRIPLO

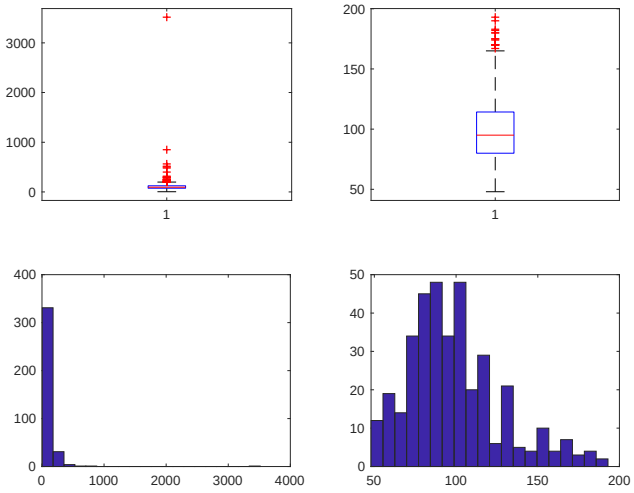

Figura 69. Superficie METALLO NO TERMICO-VETRO A BASSA EMISSIONE

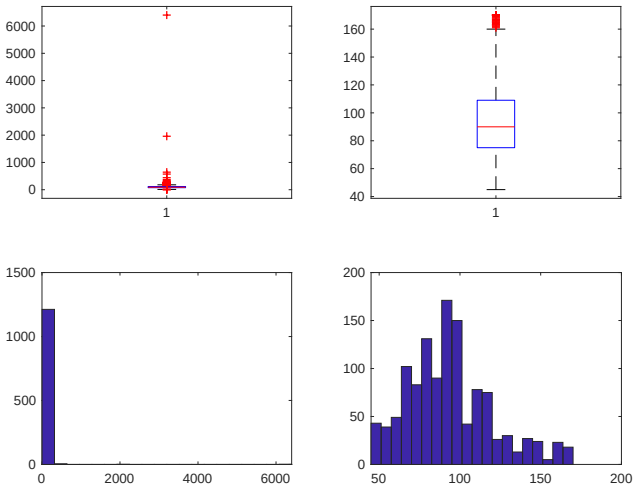

Figura 70. Superficie METALLO NO TERMICO-VETRO NON ESISTENTE

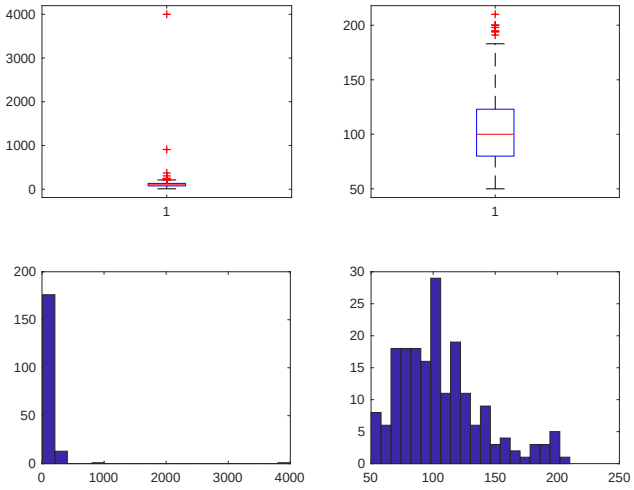

Figura 71. Superficie METALLO TERMICO-SINGOLO

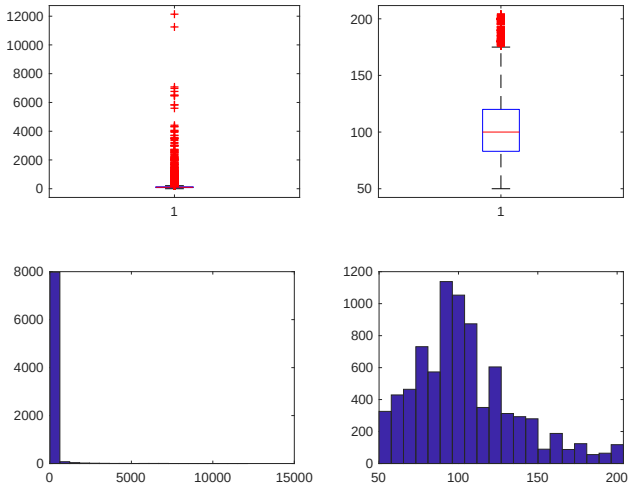

Figura 72. Superficie METALLO TERMICO-DOPPIO

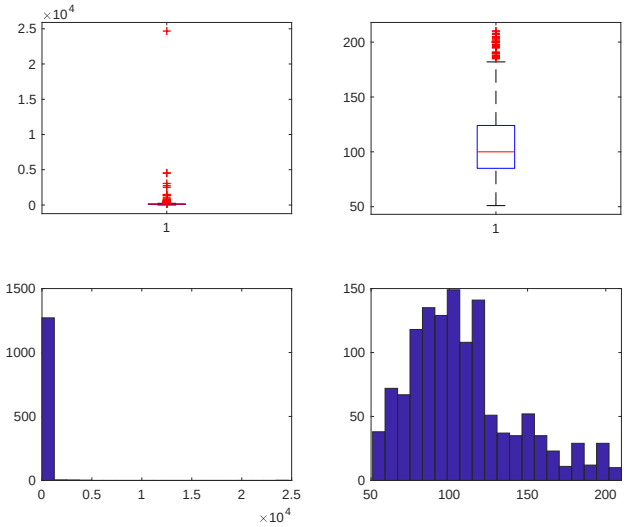

Figura 73. Superficie METALLO TERMICO-TRIPLO

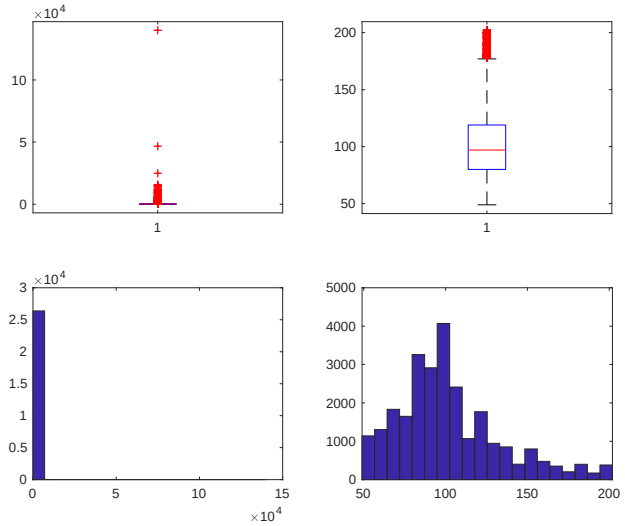

Figura 74. Superficie METALLO TERMICO-VETRO A BASSA EMISSIONE

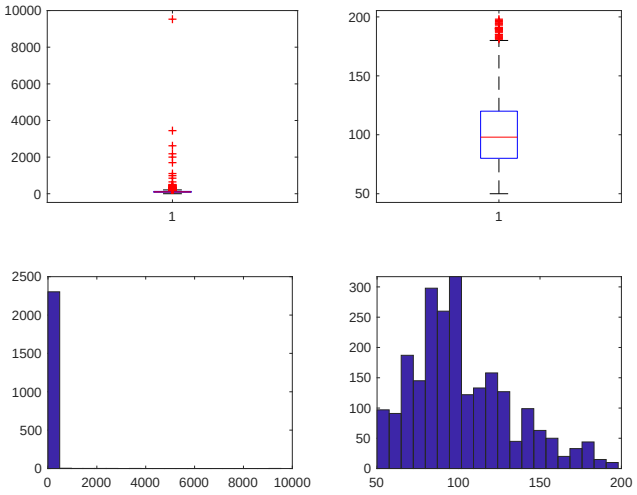

Figura 75. Superficie METALLO TERMICO-VETRO NON ESISTENTE

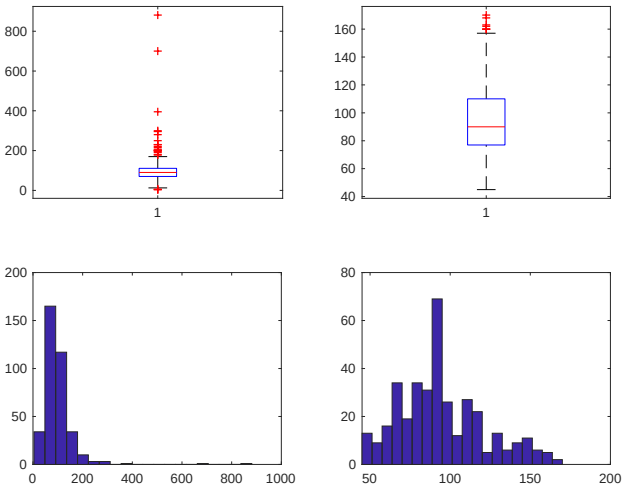

Figura 76. Superficie PVC-SINGOLO

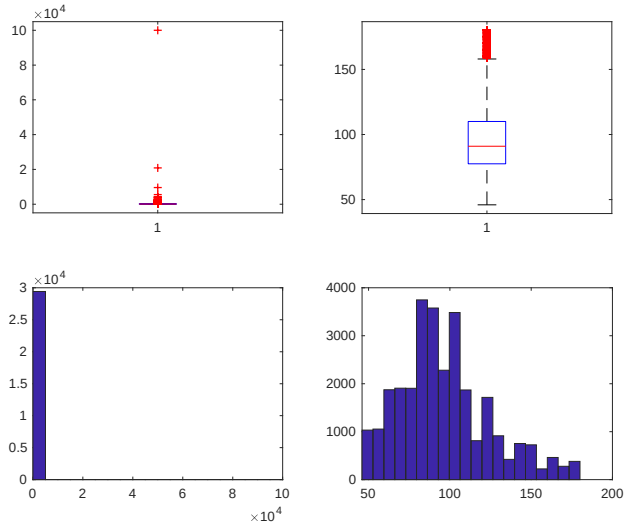

Figura 77. Superficie PVC-DOPPIO

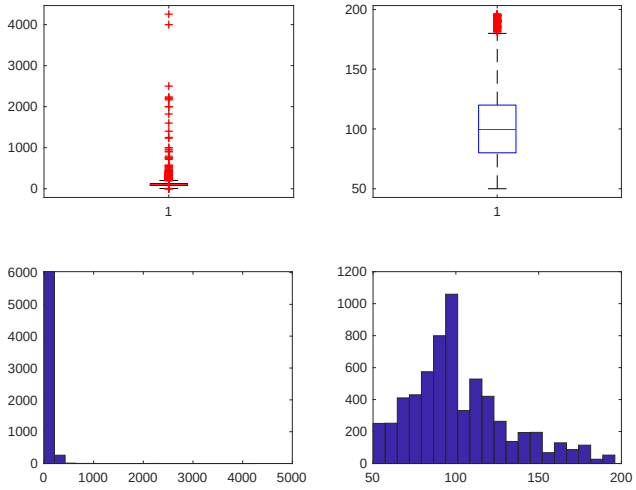

Figura 78. Superficie PVC-TRIPLO

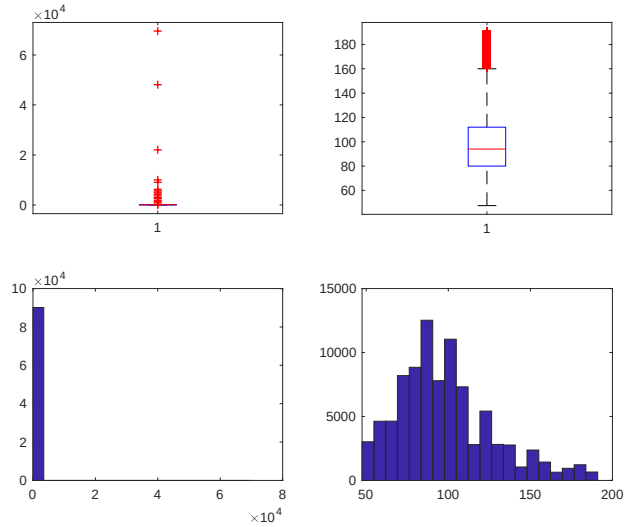

Figura 79. Superficie PVC-VETRO A BASSA EMISSIONE

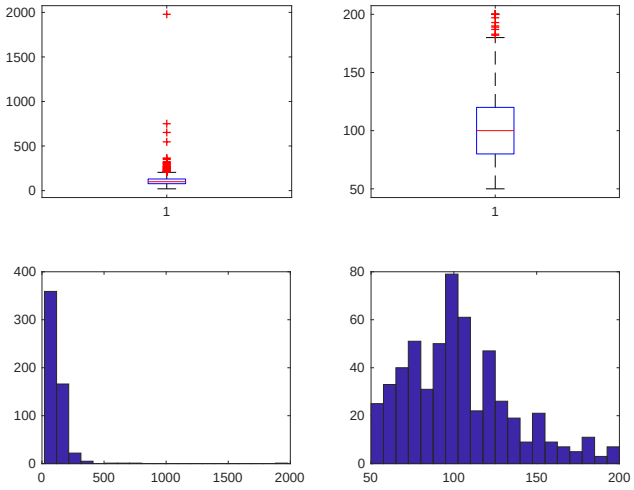

Figura 80. Superficie PVC-VETRO NON ESISTENTE

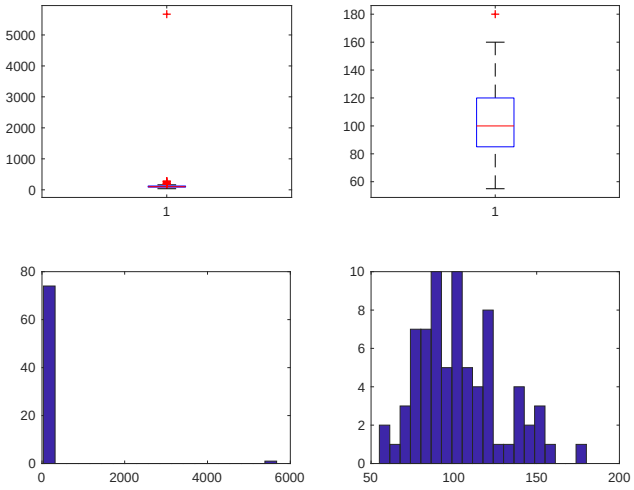

Figura 81. Superficie MISTO-SINGOLO

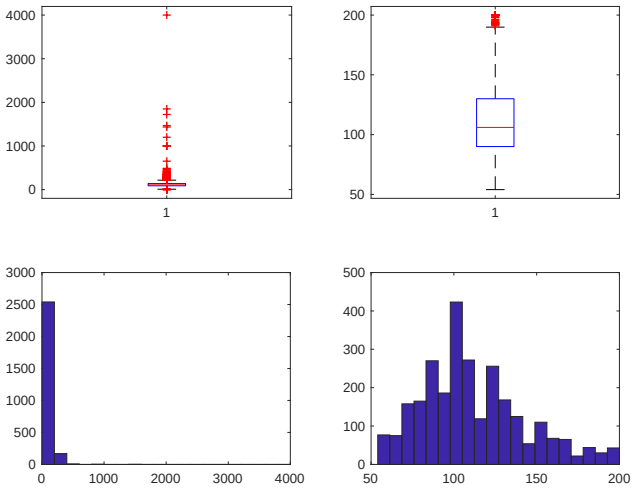

Figura 82. Superficie MISTO-DOPPIO

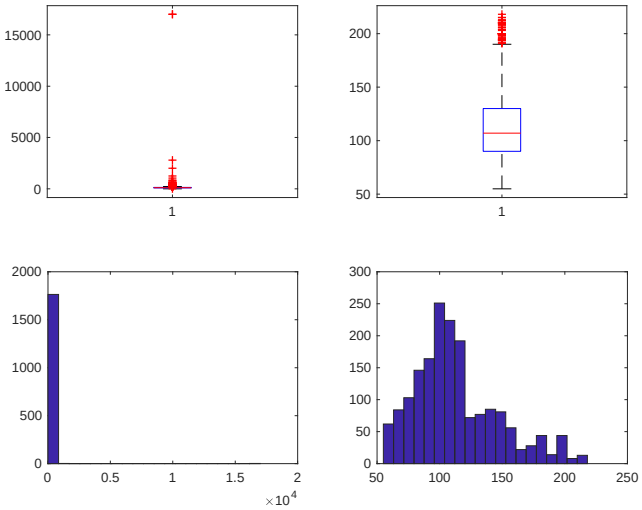

Figura 83. Superficie MISTO-TRIPLO

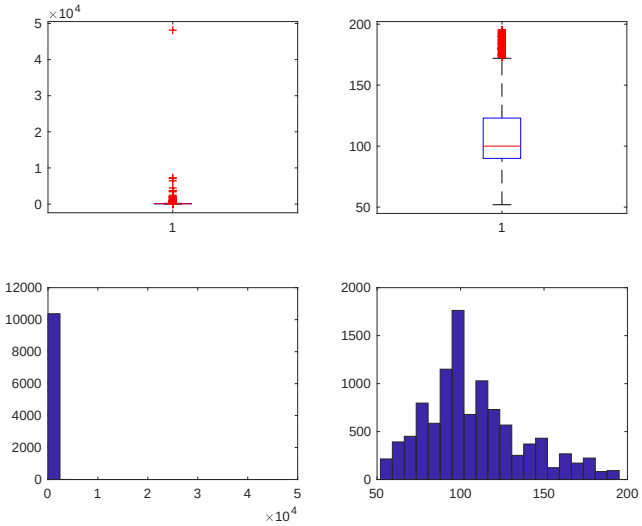

Figura 84. Superficie MISTO-VETRO A BASSA EMISSIONE

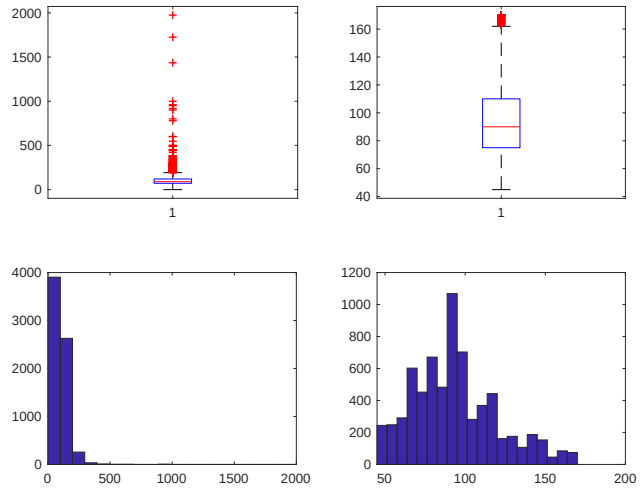

Figura 85. Superficie MISTO-VETRO NON ESISTENTE

Per ogni sottopopolazione sono state prese in considerazione le variabili RISPARMIO, COSTO (Costo intervento + Costo professionale), COSTO/RISPARMIO per verificare ulteriormente eventuali casi anomali sulla variabile COSTO e DETRAZIONE.
Tutte le variabili sono state studiate dopo averle normalizzate per "numero di unità immobiliare" e per "superficie".
Per le variabili RISPARMIO, COSTO e COSTO/RISPARMIO, i dati anomali sono stati individuati ed imputati tramite due programmi MATLAB simili a quello utilizzato per la variabile "superficie" in modo da rispettare sia la natura delle variabili stesse che lo scopo finale dell'analisi.

I risultati più immediati sono:

- RISPARMIO: 56 imputazioni (caso 1;1), 1183 imputazioni (caso 1;2), 399 imputazioni (caso 1;3), 3001 imputazioni (caso 1;4), 308 imputazioni (caso 1;5), 5 imputazioni (caso 2;1), 218 imputazioni (caso 2;2), 8 imputazioni (caso 2;3), 81 imputazioni (caso 2;4), 110 imputazioni (caso 2;5), 40 imputazioni (caso 3;1), 1357 imputazioni (caso 3;2), 216 imputazioni (caso 3;3), 3698 imputazioni (caso 3;4), 326 imputazioni (caso 3;5), 52 imputazioni (caso 4;1), 3205 imputazioni (caso 4;2), 687 imputazioni (caso 4;3), 8518 imputazioni (caso 4;4), 97 imputazioni (caso 4;5), 15 imputazioni (caso 5;1), 440 imputazioni (caso 5;2), 278 imputazioni (caso 5;3), 1610 imputazioni (caso 5;4), 755 imputazioni (caso 5;5).

- COSTO: 30 imputazioni (caso 1;1), 894 imputazioni (caso 1;2), 233 imputazioni (caso 1;3), 2098 imputazioni (caso 1;4), 283 imputazioni (caso 1;5), 2 imputazioni (caso 2;1), 33 imputazioni (caso 2;2), 6 imputazioni (caso 2;3), 61 imputazioni (caso 2;4), 91 imputazioni (caso 2;5), 19 imputazioni (caso 3;1), 974 imputazioni (caso 3;2), 155 imputazioni (caso 3;3), 2939 imputazioni (caso 3;4), 241 imputazioni (caso 3;5), 34 imputazioni (caso 4;1), 2756 imputazioni (caso 4;2), 531 imputazioni (caso 4;3), 7526 imputazioni (caso 4;4), 45 imputazioni (caso 4;5), 12 imputazioni (caso 5;1), 314 imputazioni (caso 5;2), 197 imputazioni (caso 5;3), 1232 imputazioni (caso 5;4), 641 imputazioni (caso 5;5).

- COSTO/RISPARMIO per imputare COSTO: 50 imputazioni (caso 1;1), 899 imputazioni (caso 1;2), 227 imputazioni (caso 1;3), 2315 imputazioni (caso 1;4), 163 imputazioni (caso 1;5), 2 imputazioni (caso 2;1), 36 imputazioni (caso 2;2), 3 imputazioni (caso 2;3), 50 imputazioni (caso 2;4), 112 imputazioni (caso 2;5), 34 imputazioni (caso 3;1), 895 imputazioni (caso 3;2), 133 imputazioni (caso 3;3), 2316 imputazioni (caso 3;4), 280 imputazioni (caso 3;5), 48 imputazioni (caso 4;1), 2966 imputazioni (caso 4;2), 491 imputazioni (caso 4;3), 7581 imputazioni (caso 4;4), 63 imputazioni (caso 4;5), 10 imputazioni (caso 5;1), 272 imputazioni (caso 5;2), 149 imputazioni (caso 5;3), 898 imputazioni (caso 5;4), 940 imputazioni (caso 5;5).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

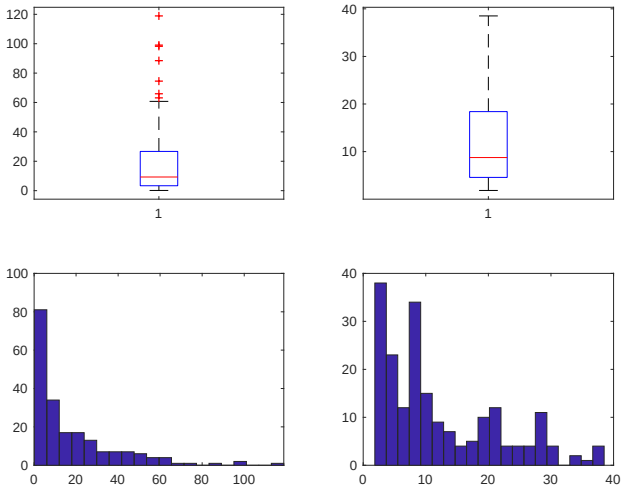

Figura 86. Risparmio LEGNO-SINGOLO

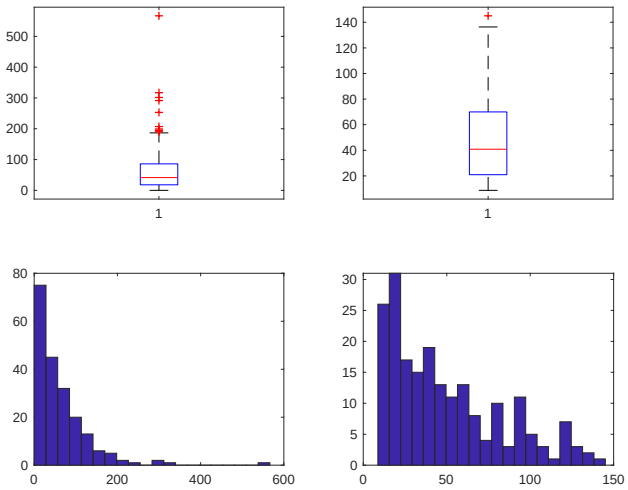

Figura 87. Costo LEGNO-SINGOLO

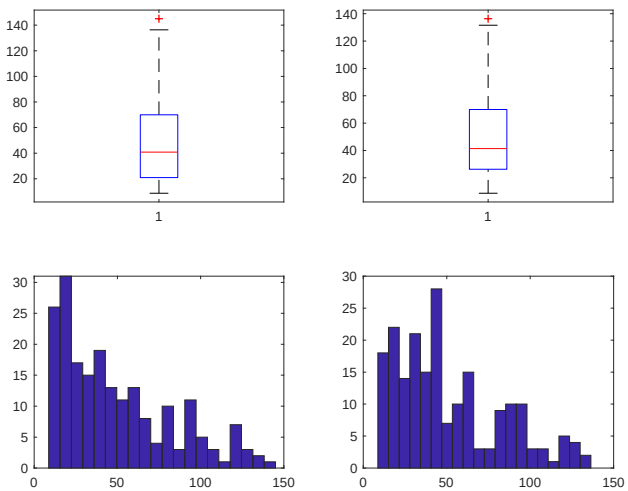

Figura 88. Costo dopo studio Costo/Risparmio LEGNO-SINGOLO

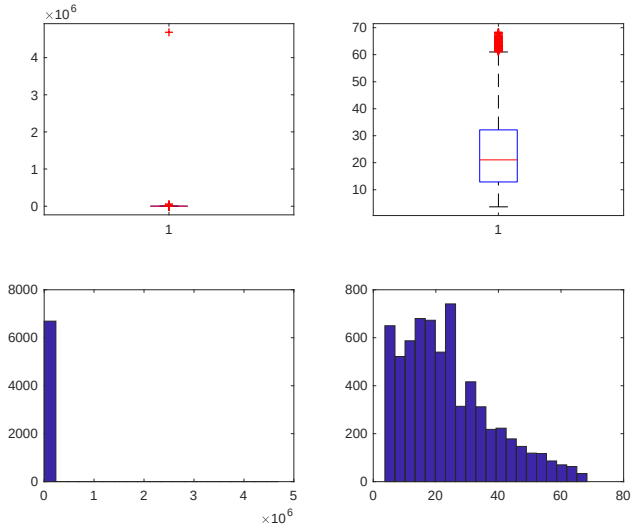

Figura 89. Risparmio LEGNO-DOPPIO

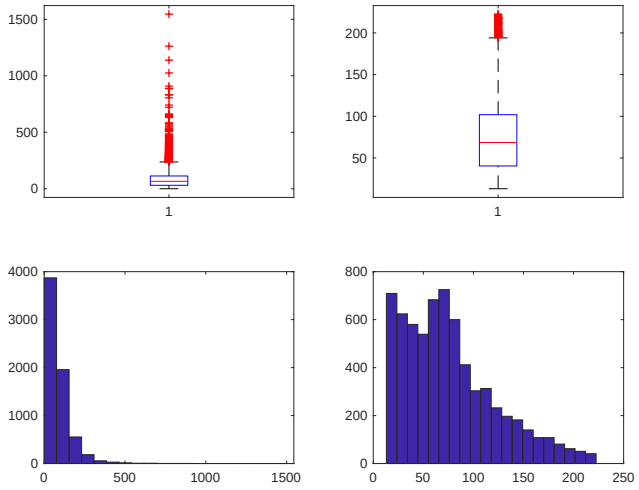

Figura 90. Costo LEGNO-DOPPIO

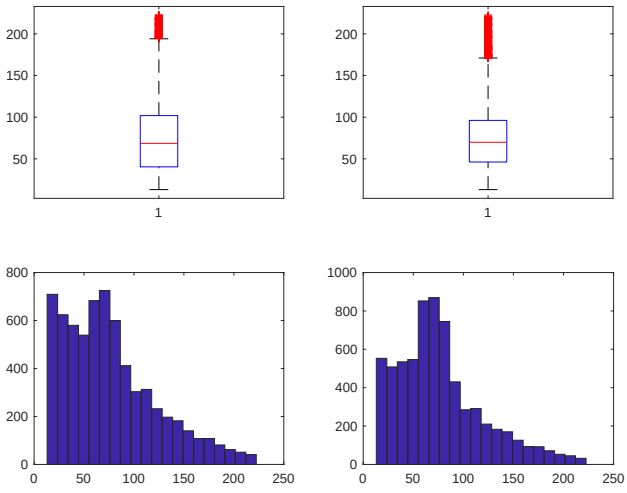

Figura 91. Costo dopo studio Costo/Risparmio LEGNO-DOPPIO

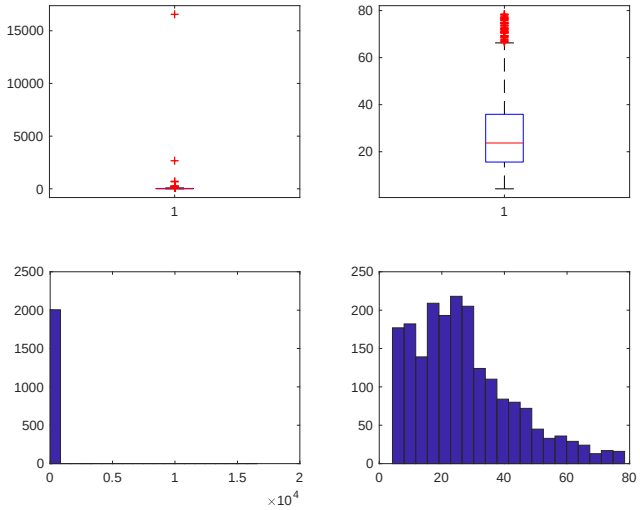

Figura 92. Risparmio LEGNO-TRIPLO

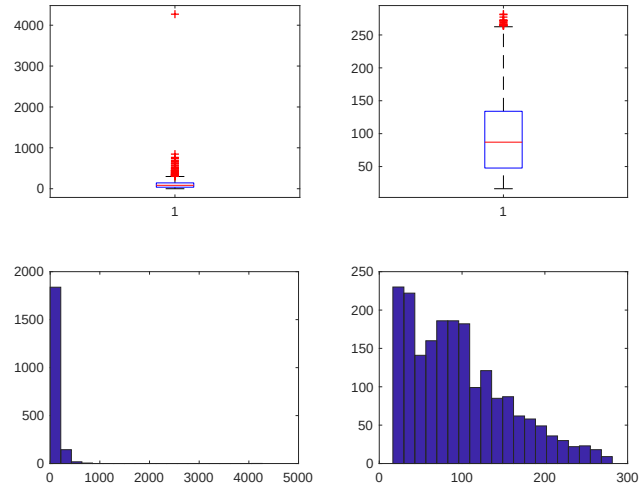

Figura 93. Costo LEGNO-TRIPLO

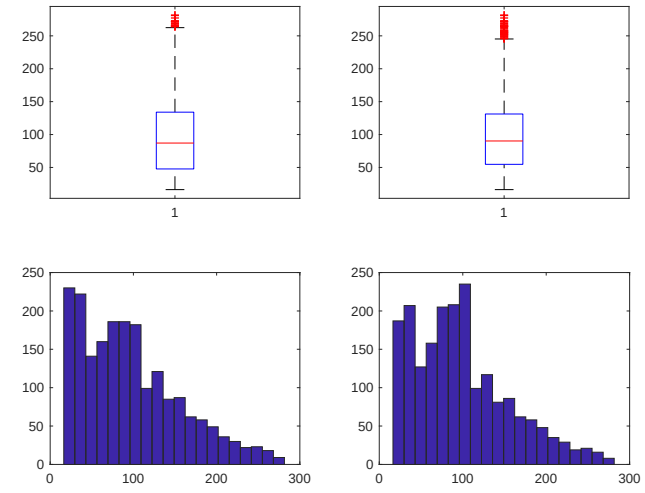

Figura 94. Costo dopo studio Costo/Risparmio LEGNO-TRIPLO

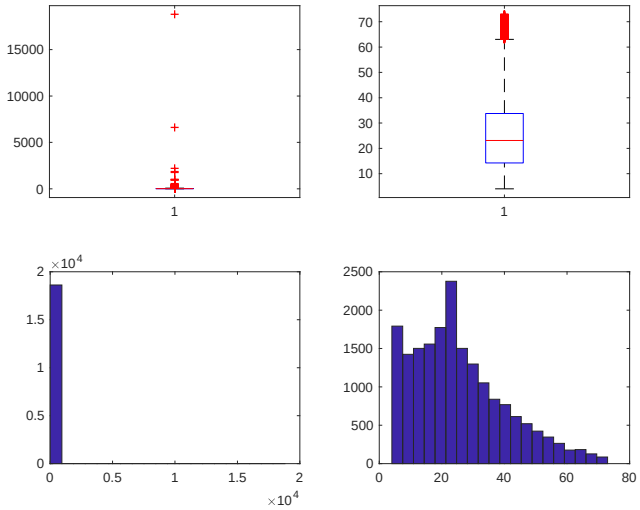

Figura 95. Risparmio LEGNO-VETRO A BASSA EMISSIONE

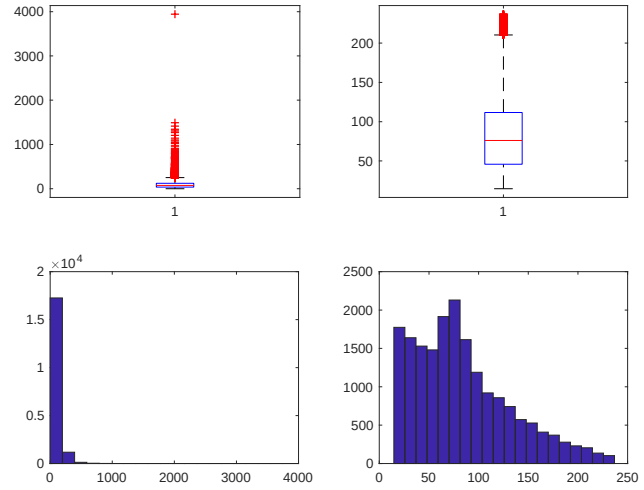

Figura 96. Costo LEGNO-VETRO A BASSA EMISSIONE

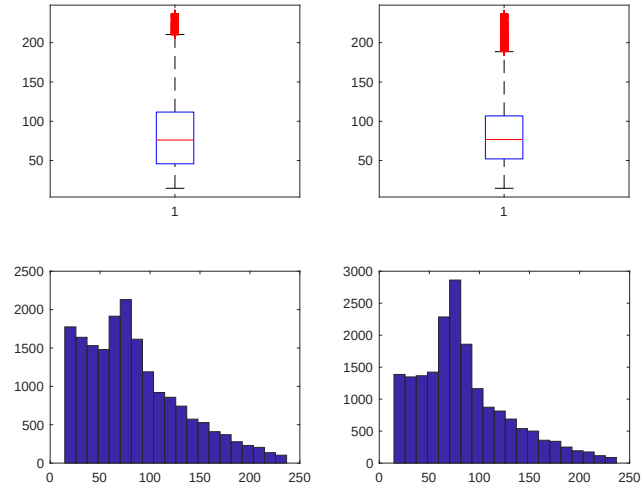

Figura 97. Costo dopo studio Costo/Risparmio LEGNO-VETRO A BASSA EMISSIONE

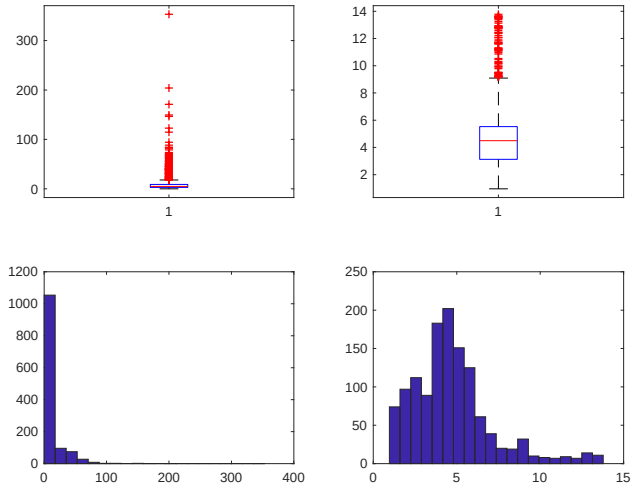

Figura 98. Risparmio LEGNO-VETRO NON ESISTENTE

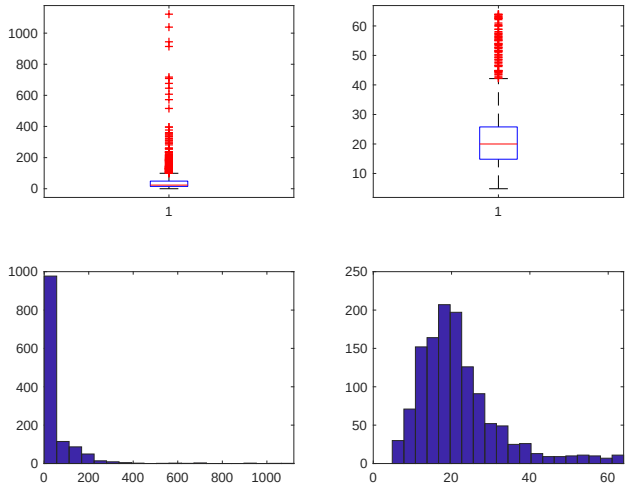

Figura 99. Costo LEGNO-VETRO NON ESISTENTE

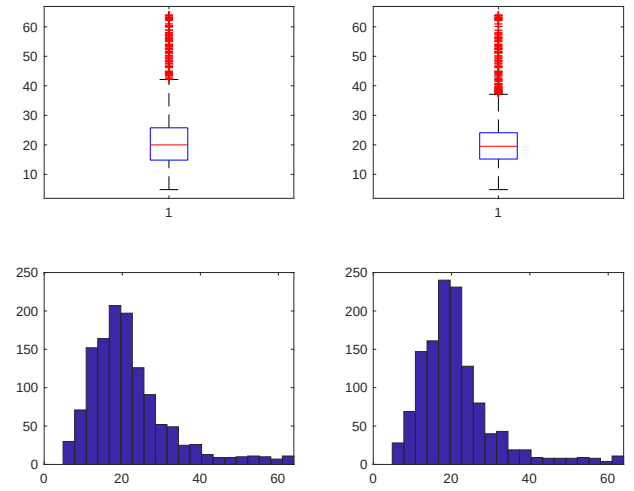

Figura 100. Costo dopo studio Costo/Risparmio LEGNO-VETRO NON ESISTENTE

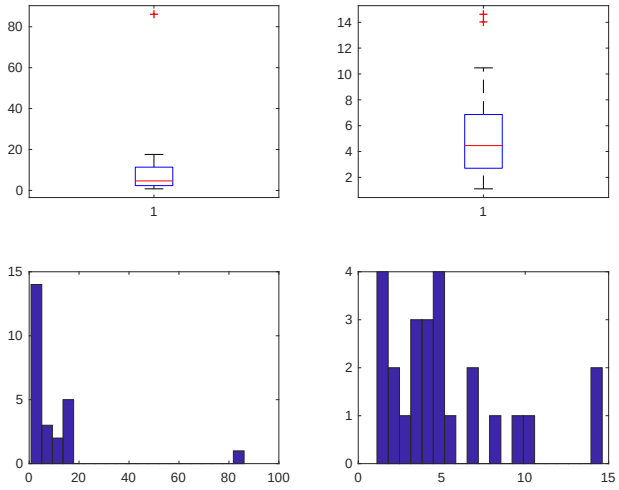

Figura 101. Risparmio METALLO NO TERMICO-SINGOLO

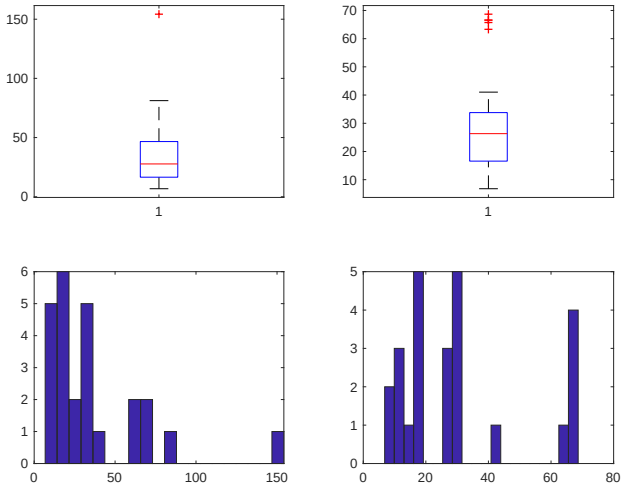

Figura 102. Costo METALLO NO TERMICO-SINGOLO

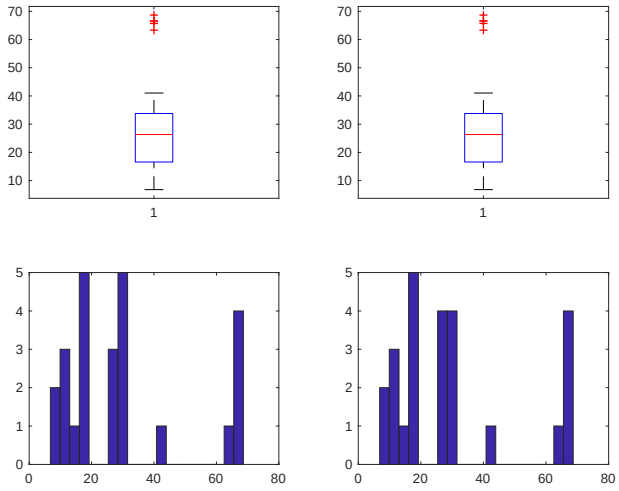

Figura 103. Costo dopo studio Costo/Risparmio METALLO NO TERMICO-SINGOLO

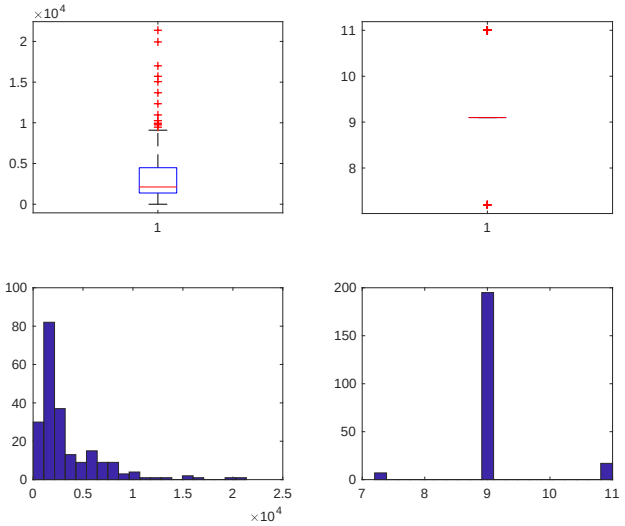

Figura 104. Risparmio METALLO NO TERMICO-DOPPIO

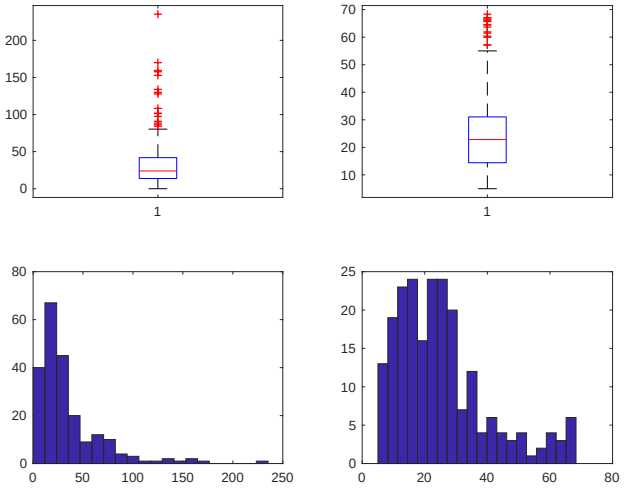

Figura 105. Costo METALLO NO TERMICO-DOPPIO

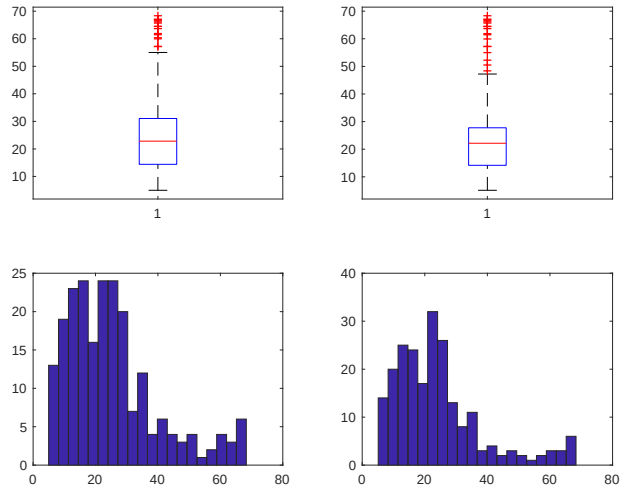

Figura 106. Costo dopo studio Costo/Risparmio METALLO NO TERMICO-DOPPIO

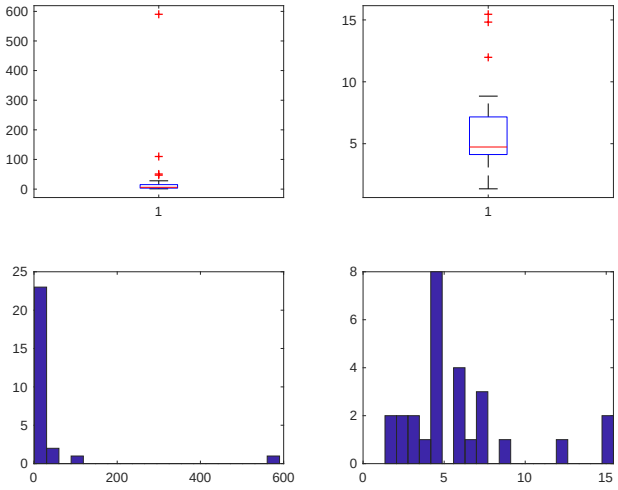

Figura 107. Risparmio METALLO NO TERMICO-TRIPLO

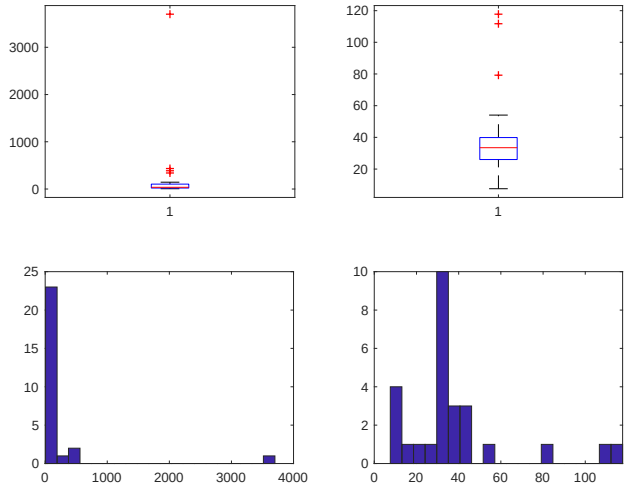

Figura 108. Costo METALLO NO TERMICO-TRIPLO

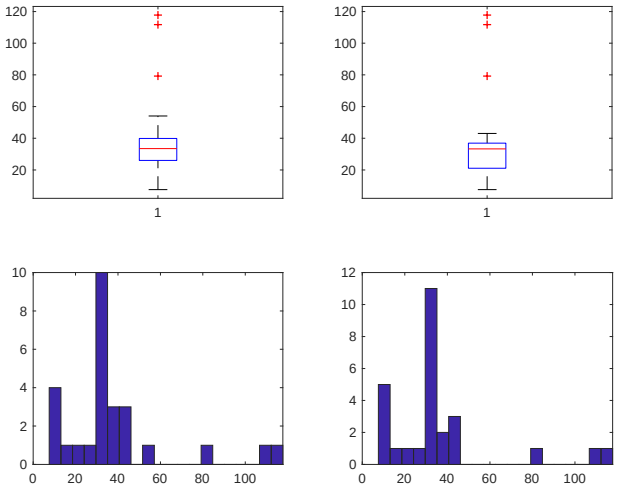

Figura 109. Costo dopo studio Costo/Risparmio METALLO NO TERMICO-TRIPLO

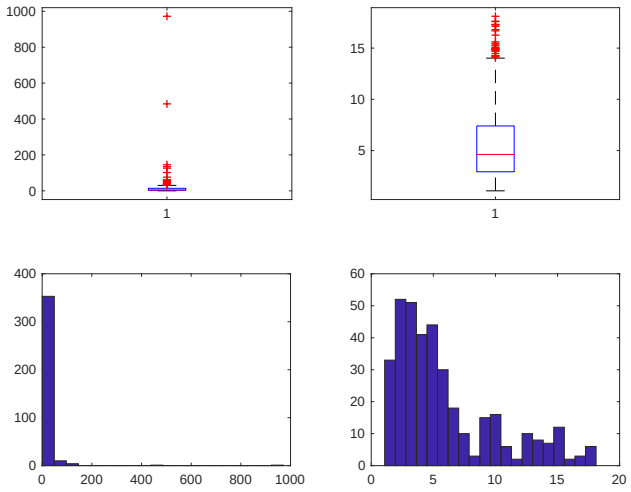

Figura 110. Risparmio METALLO NO TERMICO-VETRO A BASSA EMISSIONE

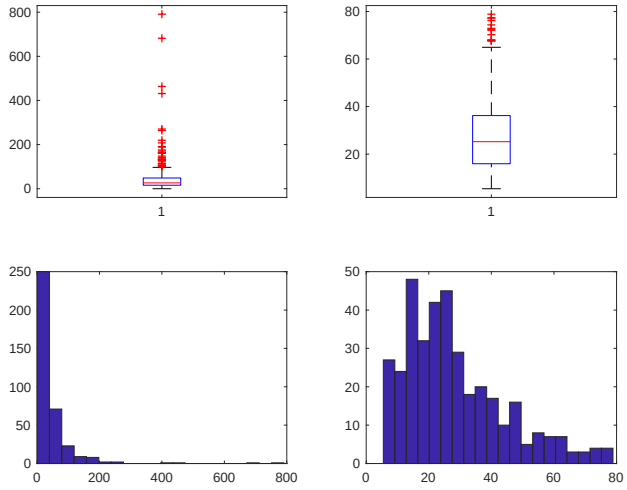

Figura 111. Costo METALLO NO TERMICO-VETRO A BASSA EMISSIONE

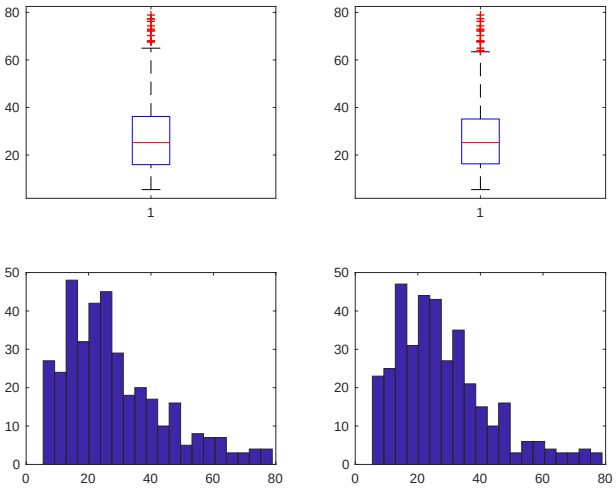

Figura 112. Costo dopo studio Costo/Risparmio METALLO NO TERMICO-VETRO A BASSA EMISSIONE

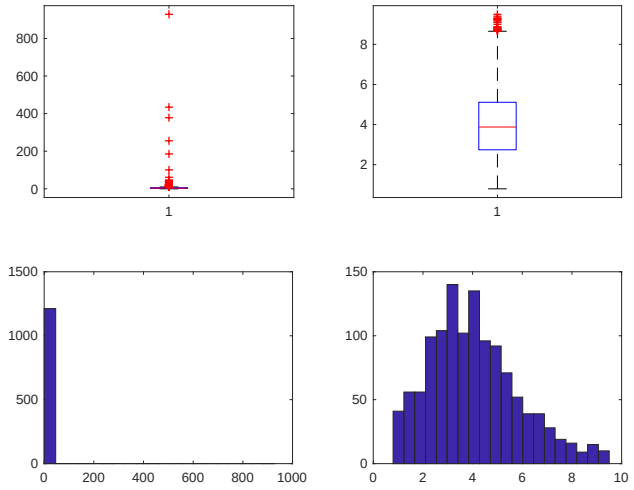

Figura 113. Risparmio METALLO NO TERMICO-VETRO NON ESISTENTE

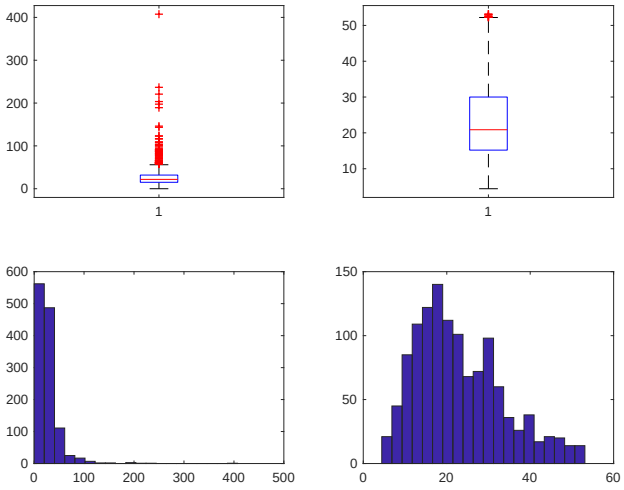

Figura 114. Costo METALLO NO TERMICO-VETRO NON ESISTENTE

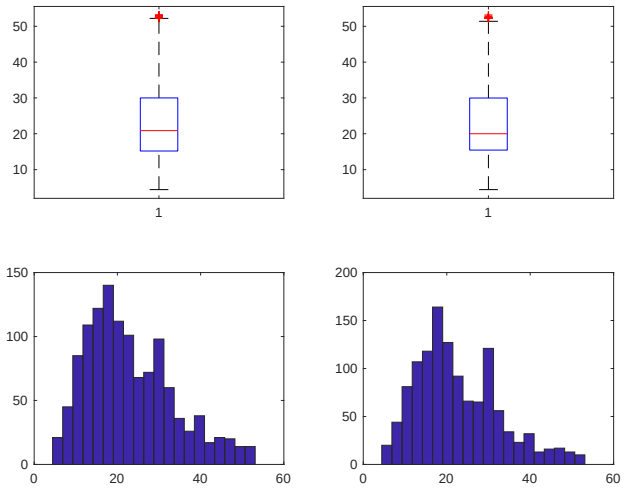

Figura 115. Costo dopo studio Costo/Risparmio METALLO NO TERMICO-VETRO NON ESISTENTE

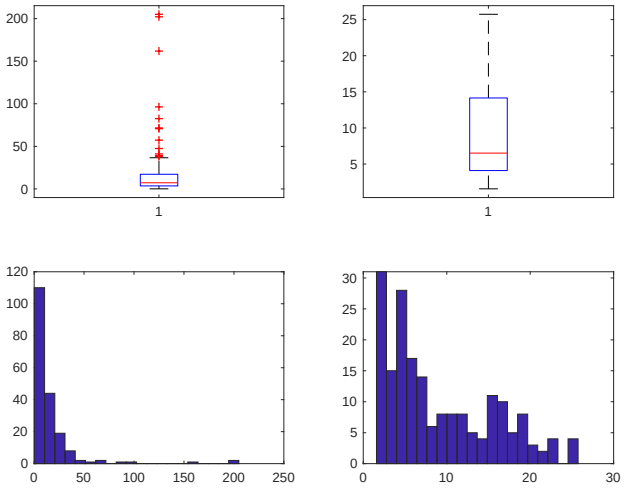

Figura 116. Risparmio METALLO TERMICO-SINGOLO

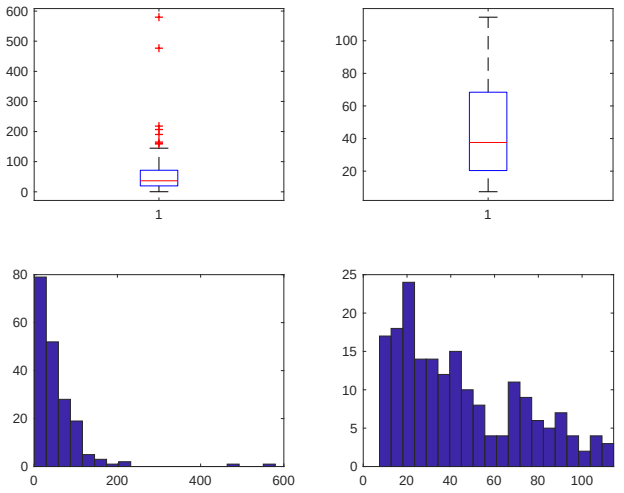

Figura 117. Costo METALLO TERMICO-SINGOLO

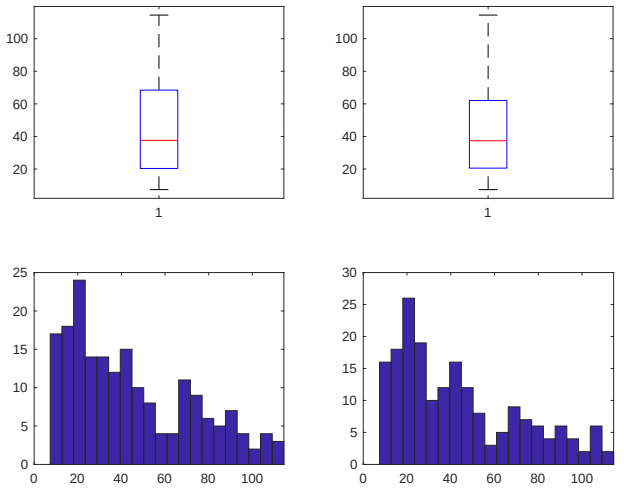

Figura 118. Costo dopo studio Costo/Risparmio METALLO TERMICO-SINGOLO

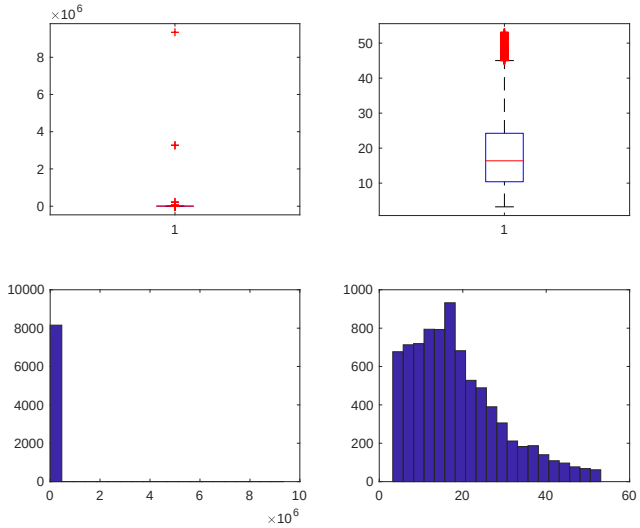

Figura 119. Risparmio METALLO TERMICO-DOPPIO

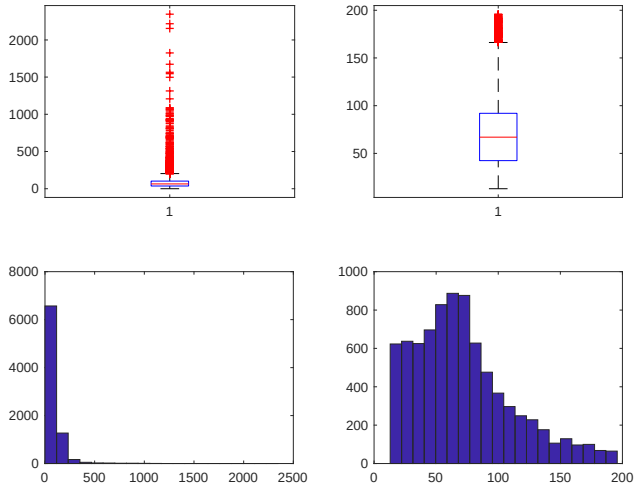

Figura 120. Costo METALLO TERMICO-DOPPIO

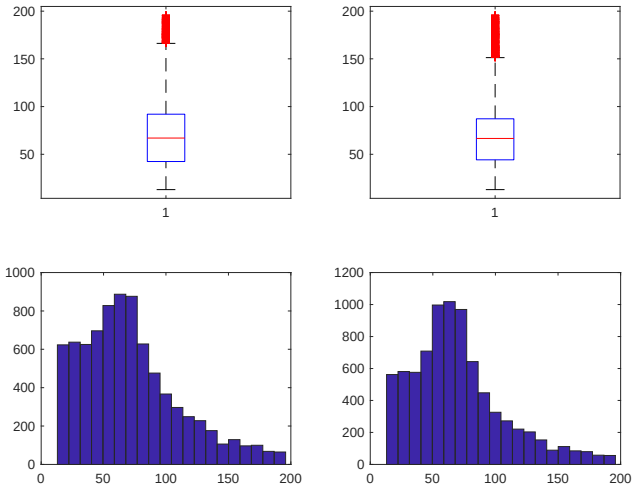

Figura 121. Costo dopo studio Costo/Risparmio METALLO TERMICO-DOPPIO

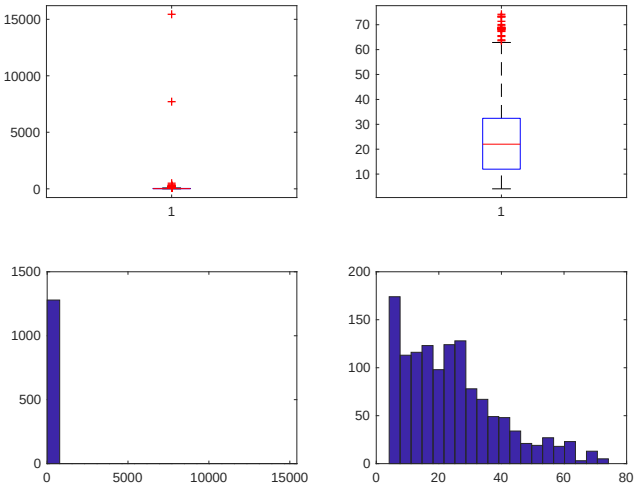

Figura 122. Risparmio METALLO TERMICO-TRIPLO

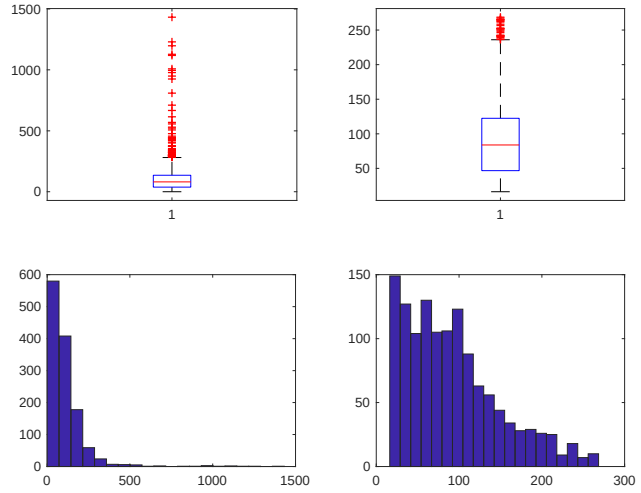

Figura 123. Costo METALLO TERMICO-TRIPLO

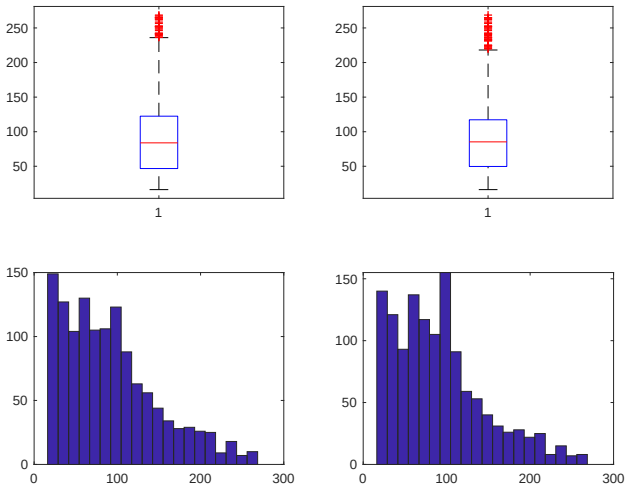

Figura 124. Costo dopo studio su Costo/Risparmio METALLO TERMICO-TRIPLO

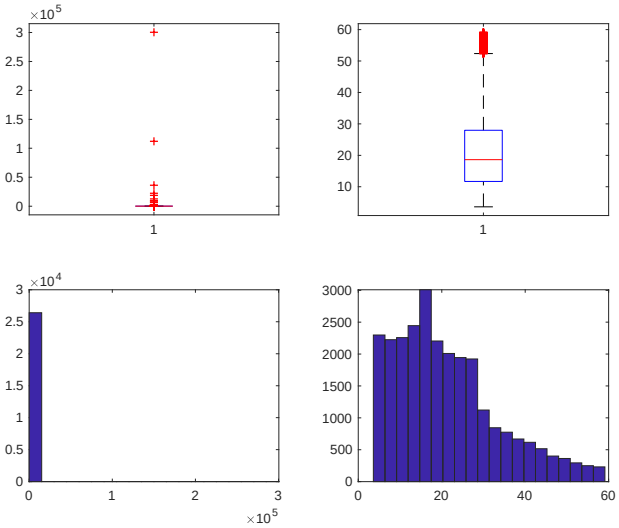

Figura 125. Risparmio METALLO TERMICO-VETRO A BASSA EMISSIONE

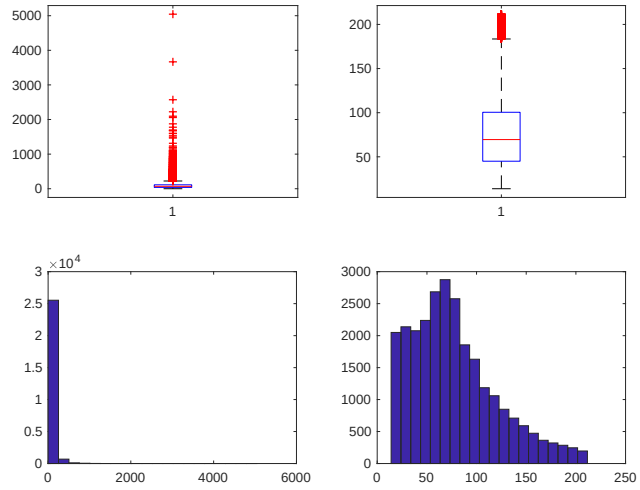

Figura 126. Costo METALLO TERMICO-VETRO A BASSA EMISSIONE

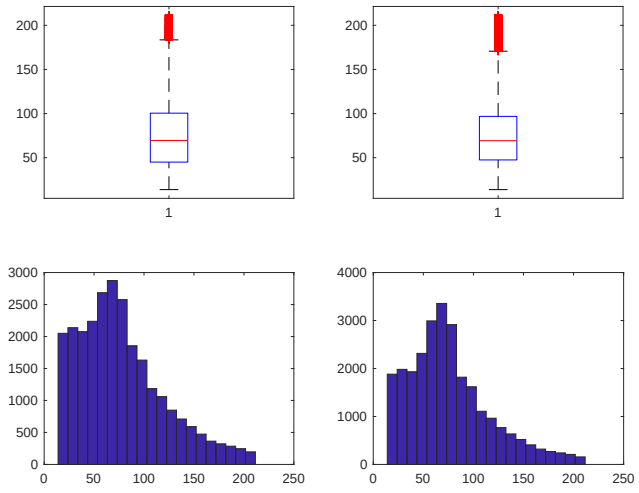

Figura 127. Costo dopo studio su Costo/Risparmio METALLO TERMICO-VETRO A BASSA EMISSIONE

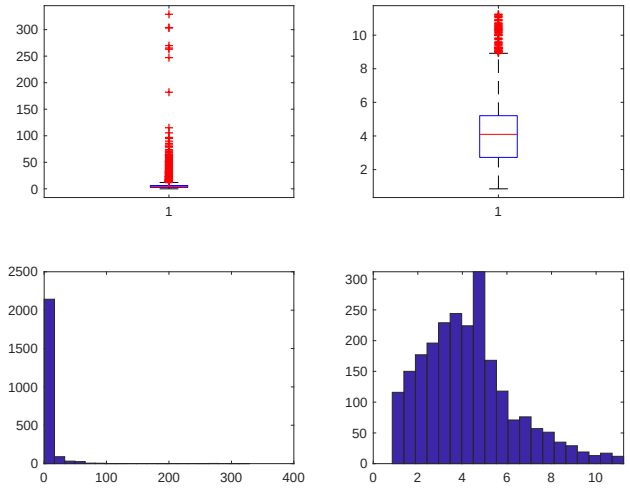

Figura 128. Risparmio METALLO TERMICO-VETRO NON ESISTENTE

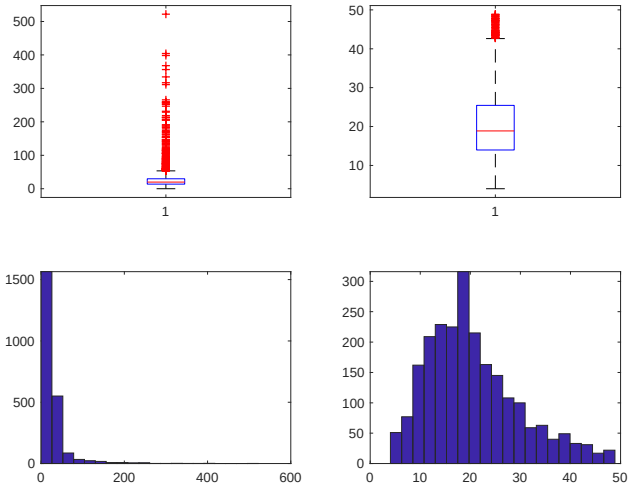

Figura 129. Costo METALLO TERMICO-VETRO NON ESISTENTE

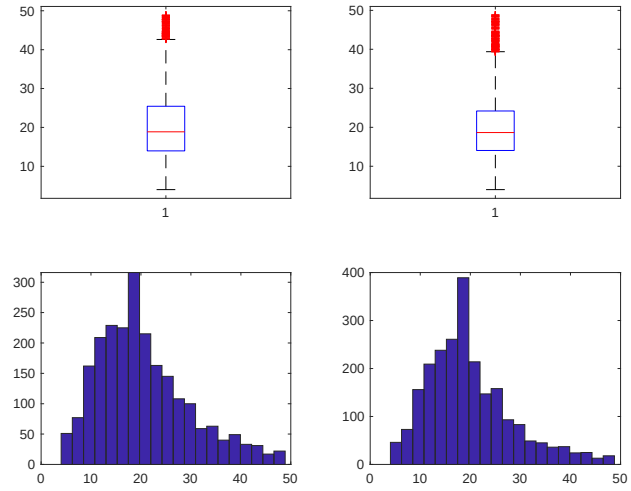

Figura 130. Costo dopo studio su Costo/Risparmio METALLO TERMICO-VETRO NON ESISTENTE

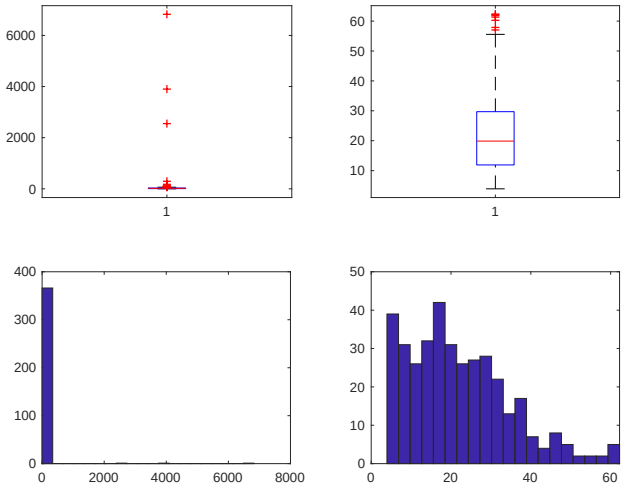

Figura 131. Risparmio PVC-SINGOLO

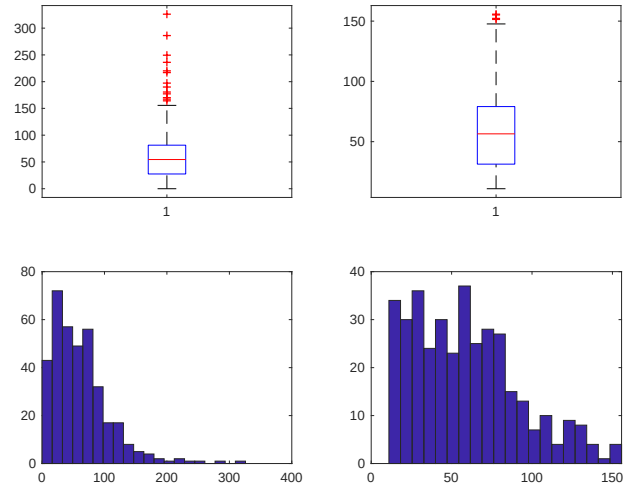

Figura 132. Costo PVC-SINGOLO

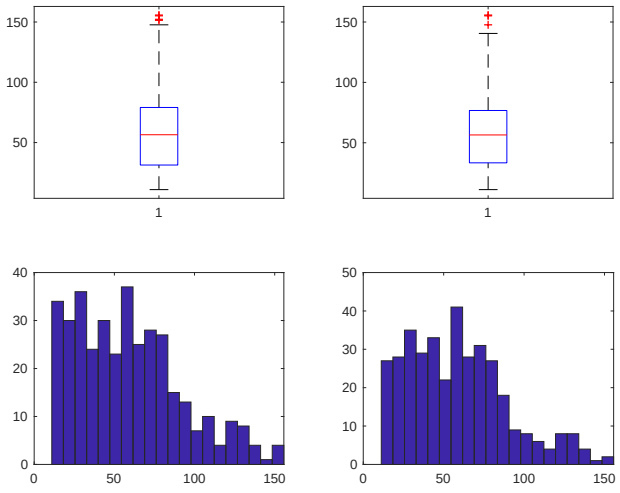

Figura 133. Costo dopo studio su Costo/Risparmio PVC-SINGOLO

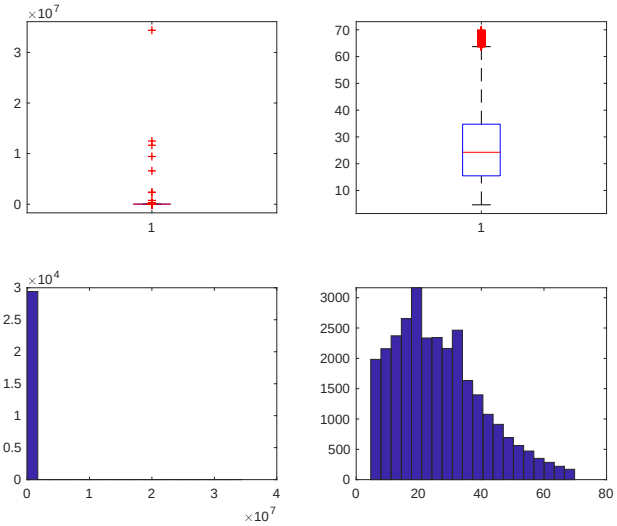

Figura 134. Risparmio PVC-DOPPIO

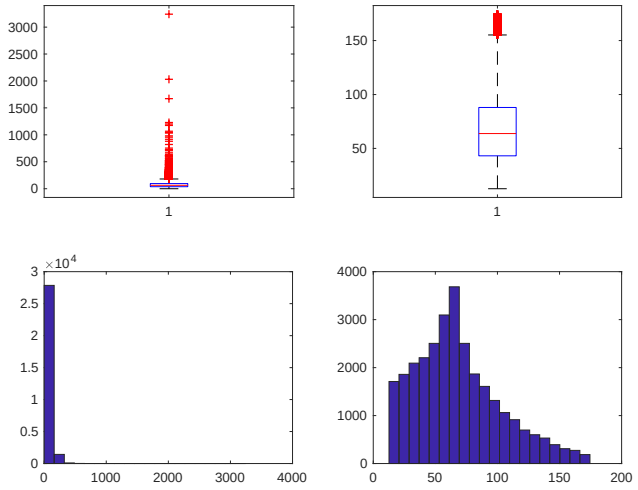

Figura 135. Costo PVC-DOPPIO

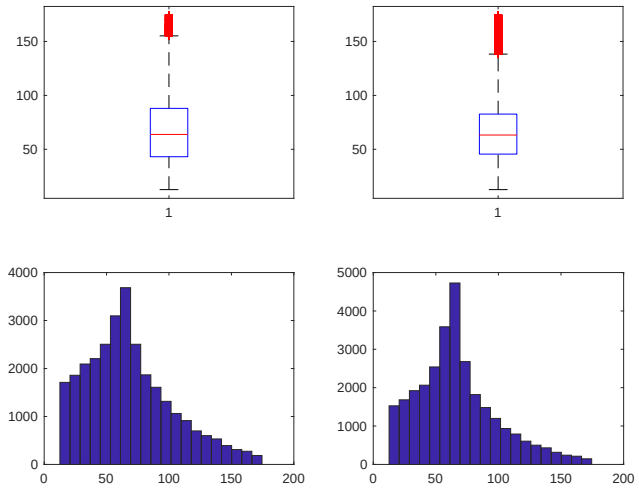

Figura 136. Costo dopo studio Costo/Risparmio PVC-DOPPIO

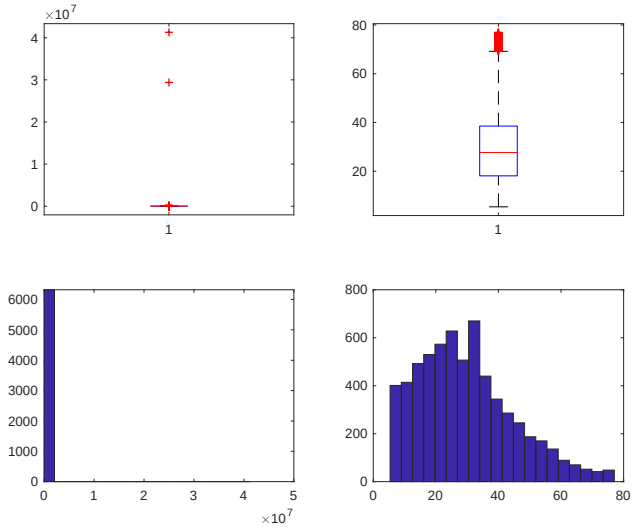

Figura 137. Risparmio PVC-TRIPLO

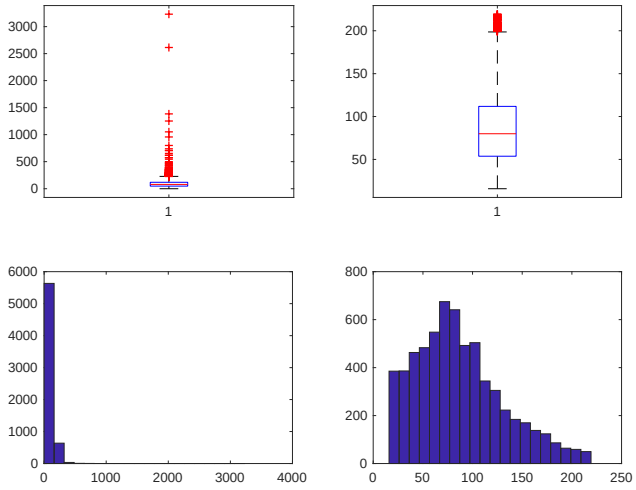

Figura 138. Costo PVC-TRIPLO

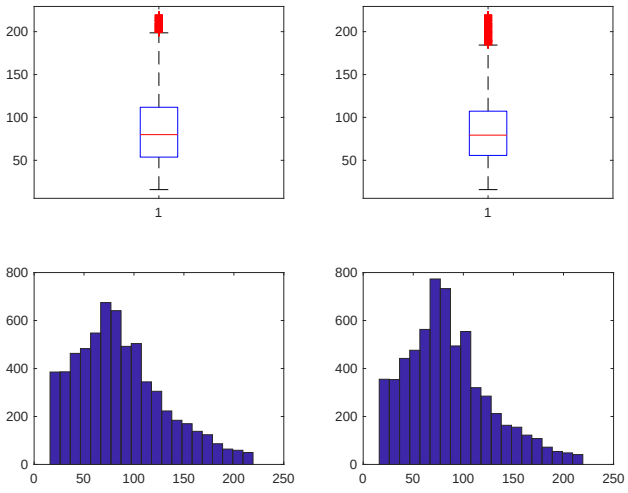

Figura 139. Costo dopo studio Costo/Risparmio PVC-TRIPLO

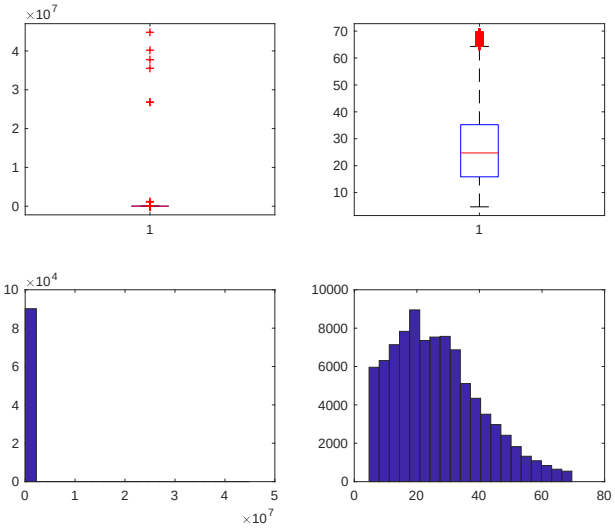

Figura 140. Risparmio PVC-VETRO A BASSA EMISSIONE

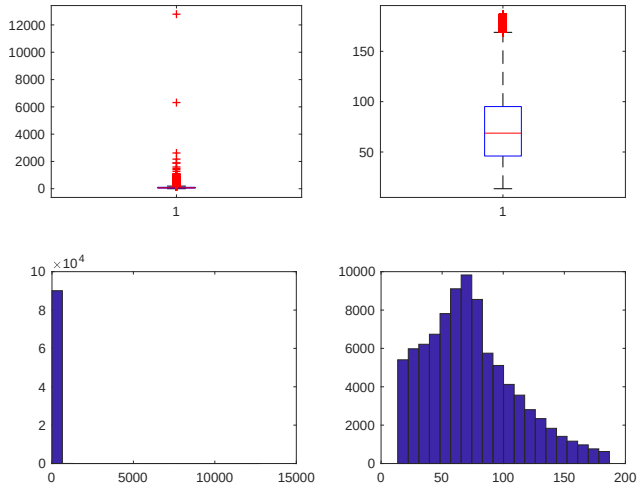

Figura 141. Costo PVC-VETRO A BASSA EMISSIONE

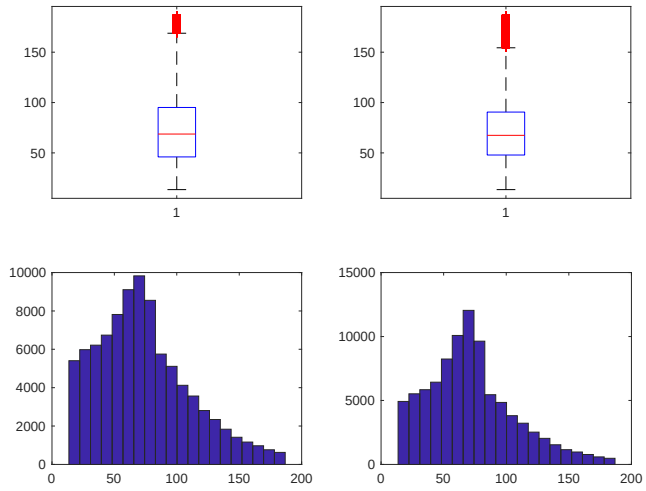

Figura 142. Costo dopo studio Costo/Risparmio PVC-VETRO A BASSA EMISSIONE

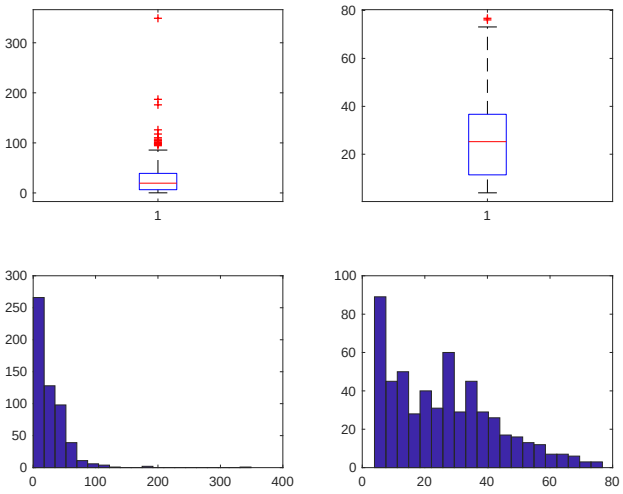

Figura 143. Risparmio PVC-VETRO NON ESISTENTE

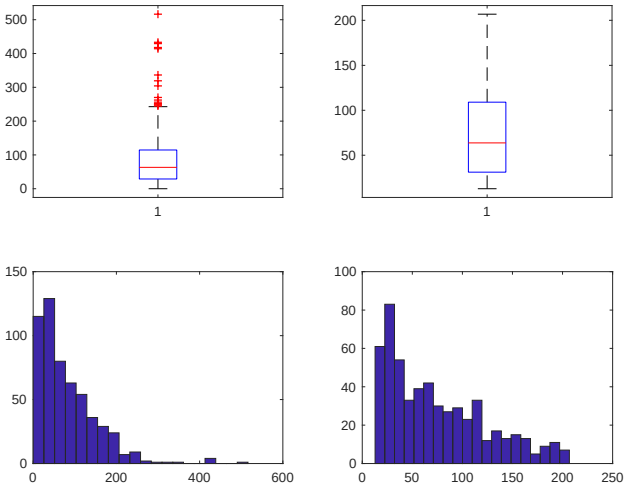

Figura 144. Costo PVC-VETRO NON ESISTENTE

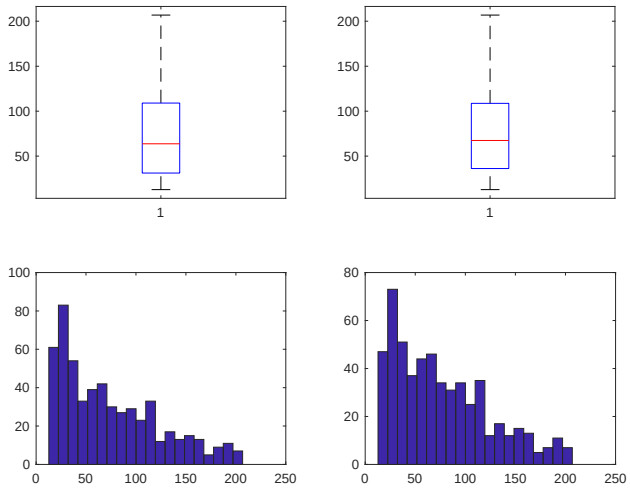

Figura 145. Costo dopo studio Costo/Risparmio PVC-VETRO NON ESISTENTE

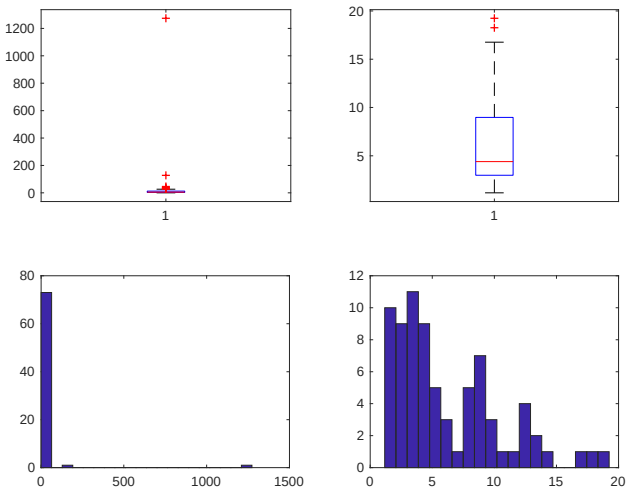

Figura 146. Risparmio MISTO-SINGOLO

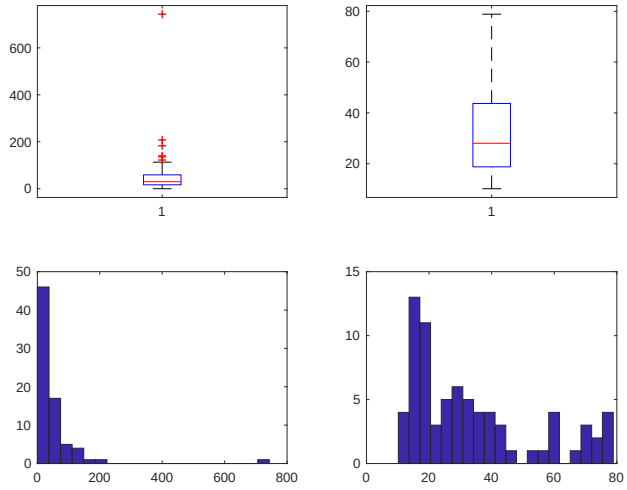

Figura 147. Costo MISTO-SINGOLO

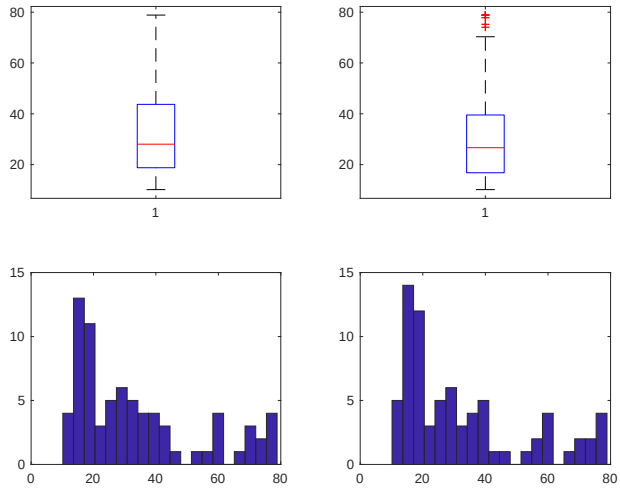

Figura 148. Costo dopo studio Costo/Risparmio MISTO-SINGOLO

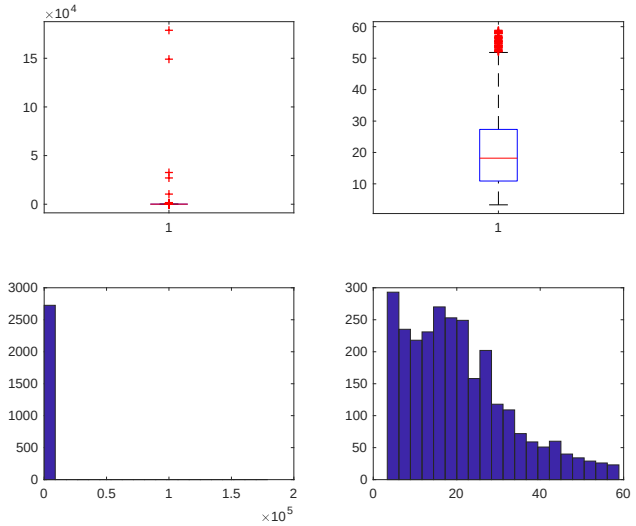

Figura 149. Risparmio MISTO-DOPPIO

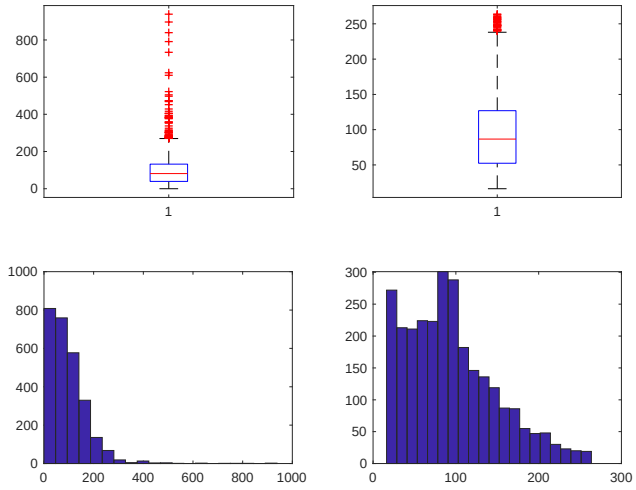

Figura 150. Costo MISTO-DOPPIO

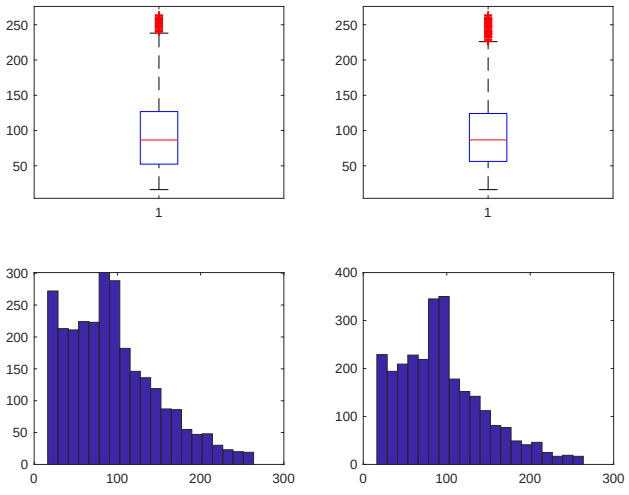

Figura 151. Costo dopo studio Costo/Risparmio MISTO-DOPPIO

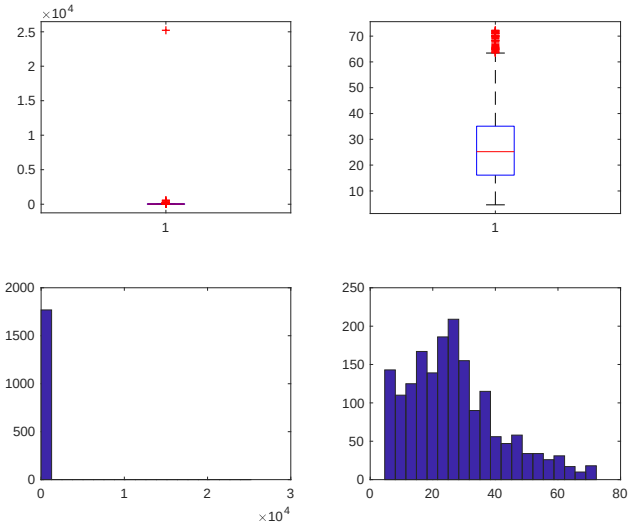

Figura 152. Risparmio MISTO-TRIPLO

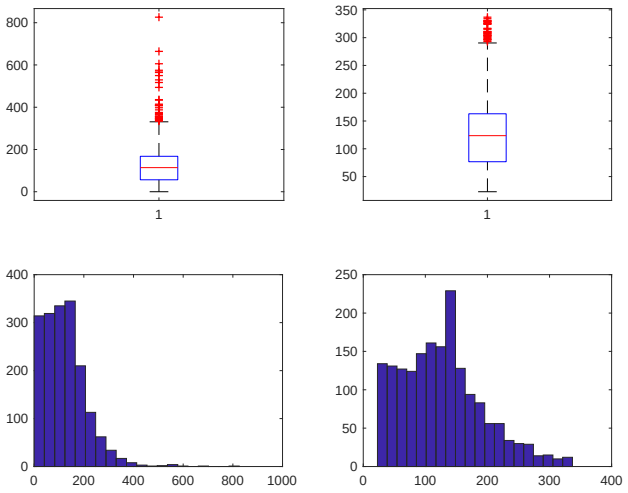

Figura 153. Costo MISTO-TRIPLO

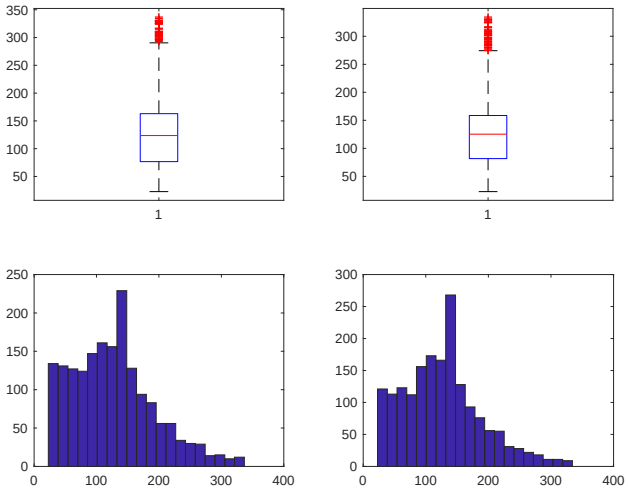

Figura 154. Costo dopo studio Costo/Risparmio MISTO-TRIPLO

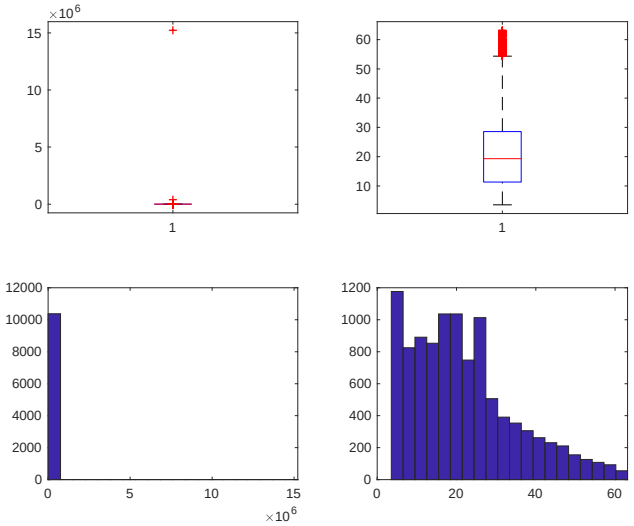

Figura 155. Risparmio MISTO-VETRO A BASSA EMISSIONE

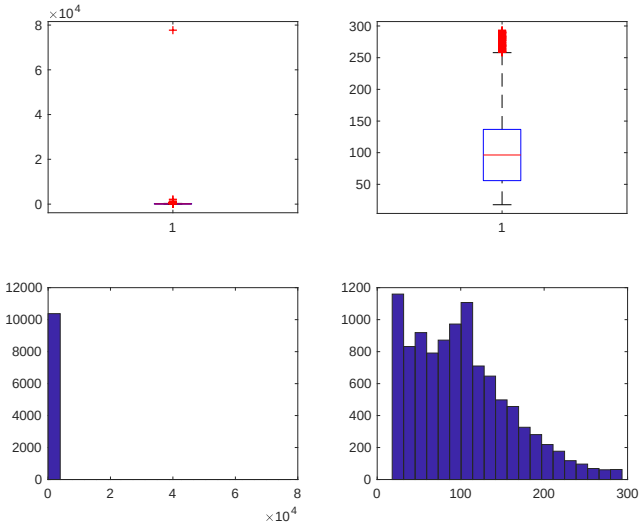

Figura 156. Costo MISTO-VETRO A BASSA EMISSIONE

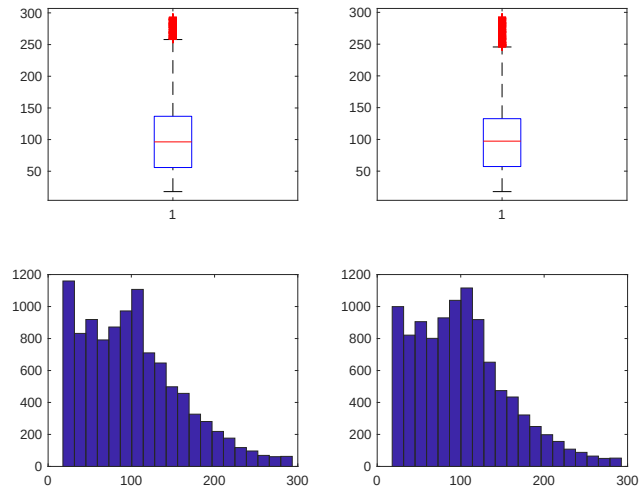

Figura 157. Costo dopo studio Costo/Risparmio MISTO-VETRO A BASSA EMISSIONE

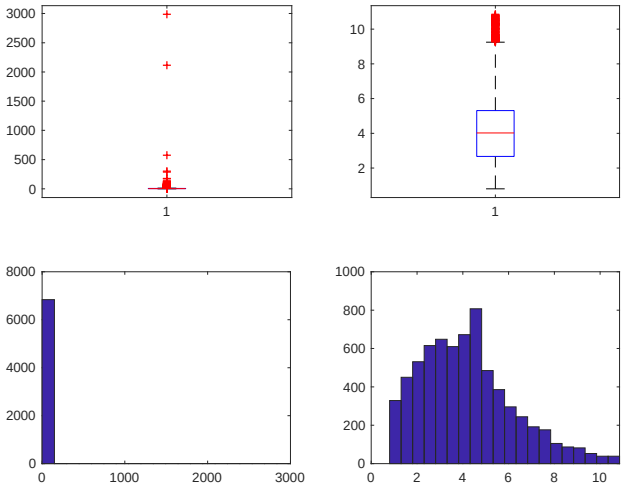

Figura 158. Risparmio MISTO-VETRO NON ESISTENTE

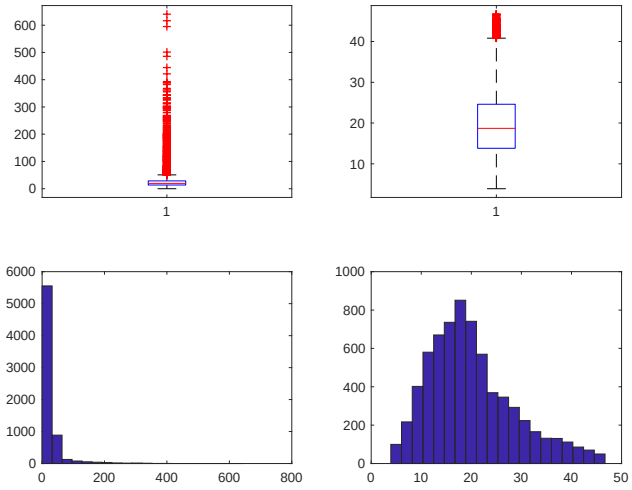

Figura 159. Costo MISTO-VETRO NON ESISTENTE

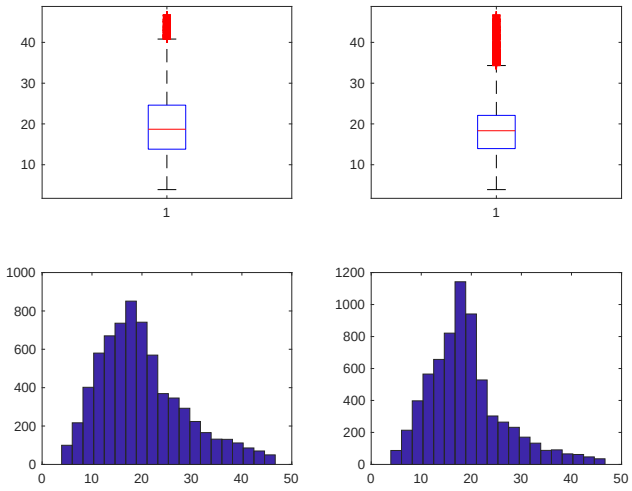

Figura 160. Costo dopo studio Costo/Risparmio MISTO-VETRO NON ESISTENTE

La variabile DETRAZIONE è stata imputata secondo la regola:
detrazione= 0.65*costo,
e sono stati imputati:

- 35 dati anomali (caso 1;1) di cui 30 per imputazioni svolte su COSTO,
- 684 dati anomali (caso 1;2) di cui 567 per imputazioni svolte su COSTO,
- 133 dati anomali (caso 1;3) di cui 110 per imputazioni svolte su COSTO,
- 1546 dati anomali (caso 1;4) di cui 1418 per imputazioni svolte su COSTO,
- 363 dati anomali (caso 1;5) di cui 355 per imputazioni svolte su COSTO,
- 6 dati anomali (caso 2;1) di cui 4 per imputazioni svolte su COSTO,
- 48 dati anomali (caso 2;2) di cui 41 per imputazioni svolte su COSTO,
- 7 dati anomali (caso 2;3) di cui 7 per imputazioni svolte su COSTO,
- 73 dati anomali (caso2;4) di cui 65 per imputazioni svolte su COSTO,
- 136 dati anomali (caso 2;5) di cui 130 per imputazioni svolte su COSTO,
- 33 dati anomali (caso 3;1) di cui 26 per imputazioni svolte su COSTO,
- 1030 dati anomali (caso 3;2) di cui 858 per imputazioni svolte su COSTO,
- 127 dati anomali (caso 3;3) di cui 108 per imputazioni svolte su COSTO,
- 2764 dati anomali (caso 3;4) di cui 2450 per imputazioni svolte su COSTO,
- 399 dati anomali (caso 3;5) di cui 378 per imputazioni svolte su COSTO,
- 37 dati anomali (caso 4;1) di cui 30 per imputazioni svolte su COSTO,
- 3081 dati anomali (caso 4;2) di cui 2531 per imputazioni svolte su COSTO,
- 529 dati anomali (caso 4;3) di cui 423 per imputazioni svolte su COSTO,
- 7723 dati anomali (caso 4;4) di cui 6989 per imputazioni svolte su COSTO,
- 40 dati anomali (caso 4;5) di cui 37 per imputazioni svolte su COSTO,
- 17 dati anomali (caso 5;1) di cui 14 per imputazioni svolte su COSTO,
- 199 dati anomali (caso 5;2) di cui 149 per imputazioni svolte su COSTO,
- 113 dati anomali (caso 5;3) di cui 89 per imputazioni svolte su COSTO,
- 661 dati anomali (caso 5;4) di cui 576 per imputazioni svolte su COSTO,
- 1255 dati anomali (caso 5;5) di cui 1225 per imputazioni svolte su COSTO.

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

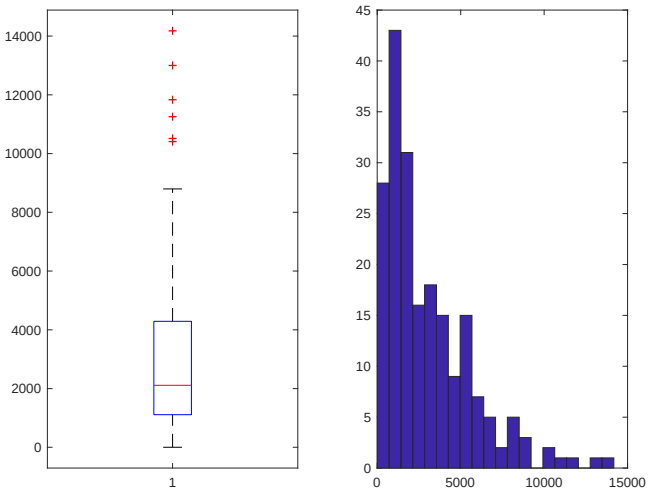

Figura 161. Detrazione LEGNO-SINGOLO

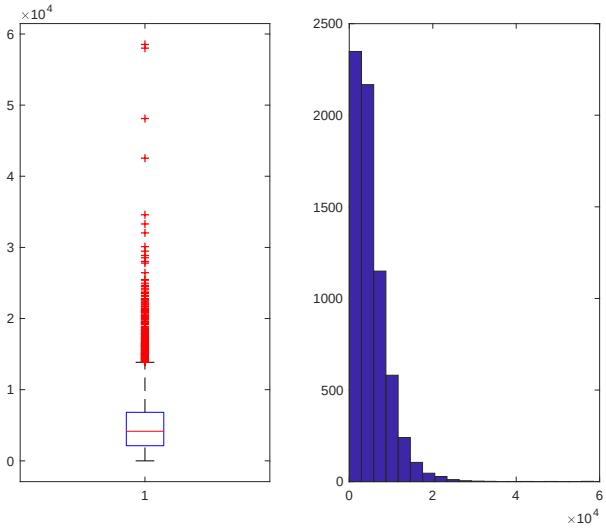

Figura 162. Detrazione LEGNO-DOPPIO

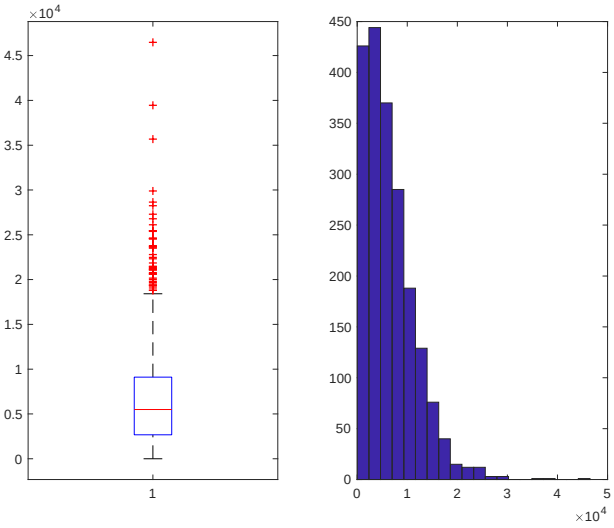

Figura 163. Detrazione LEGNO-TRIPLO

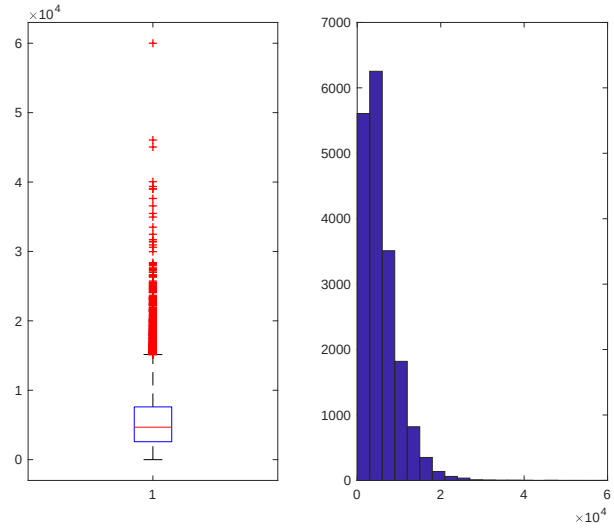

Figura 164. Detrazione LEGNO-VETRO A BASSA EMISSIONE

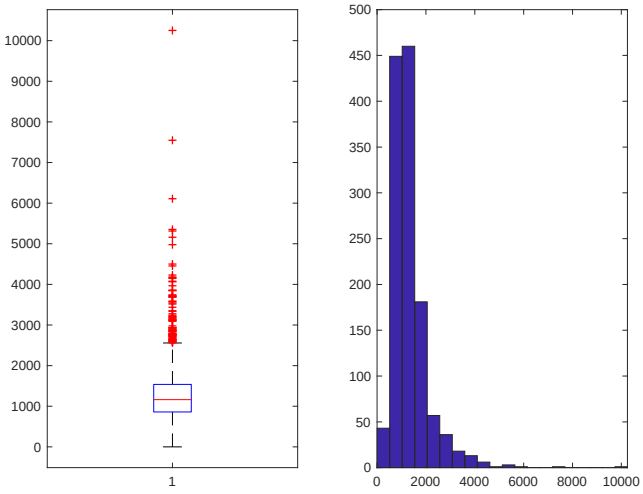

Figura 165. Detrazione LEGNO-VETRO NON ESISTENTE

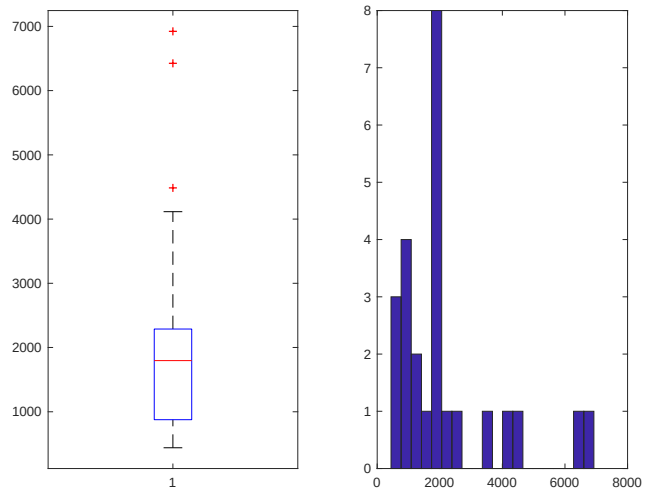

Figura 166. Detrazione METALLO NO TERMICO-SINGOLO

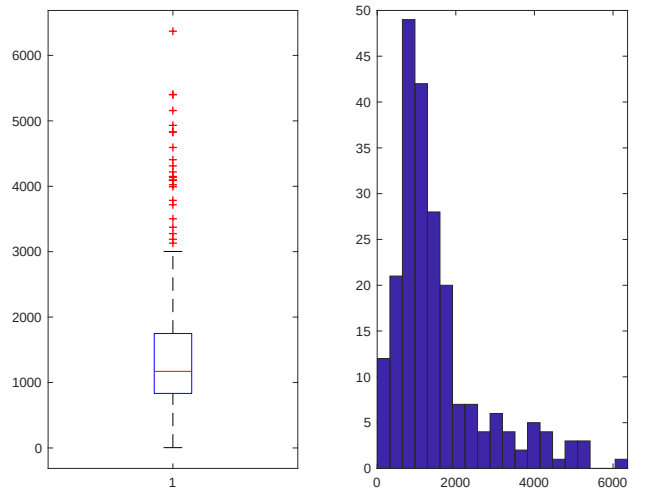

Figura 167. Detrazione METALLO NO TERMICO-DOPPIO

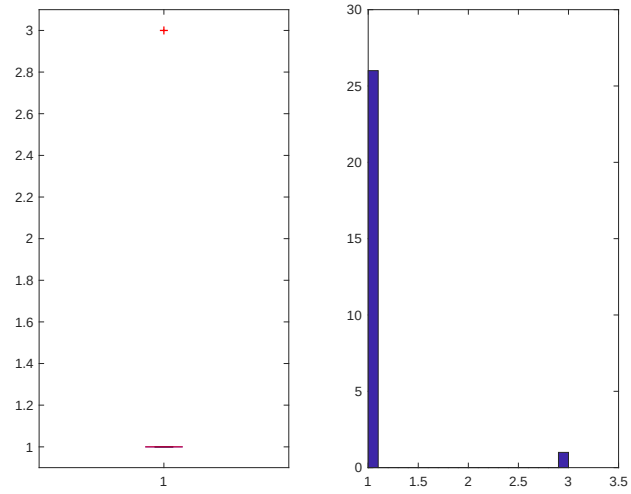

Figura 168. Detrazione METALLO NO TERMICO-TRIPLO

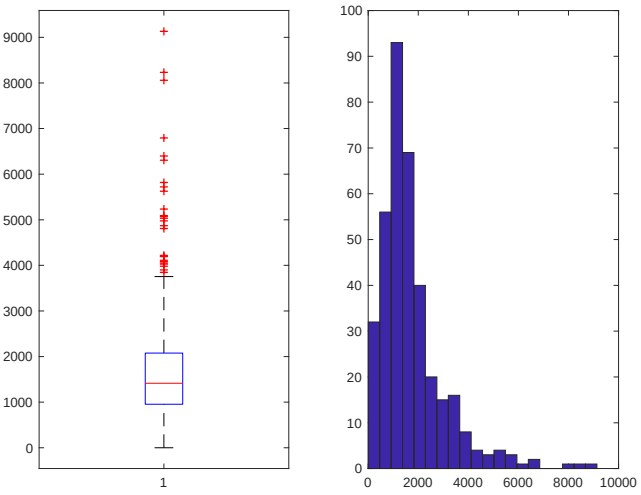

Figura 169. Detrazione METALLO NO TERMICO-VETRO A BASSA EMISSIONE

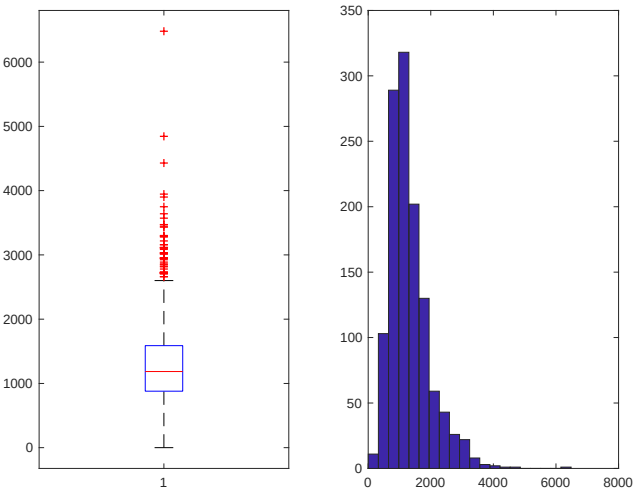

Figura 170. Detrazione METALLO NO TERMICO-VETRO NON ESISTENTE

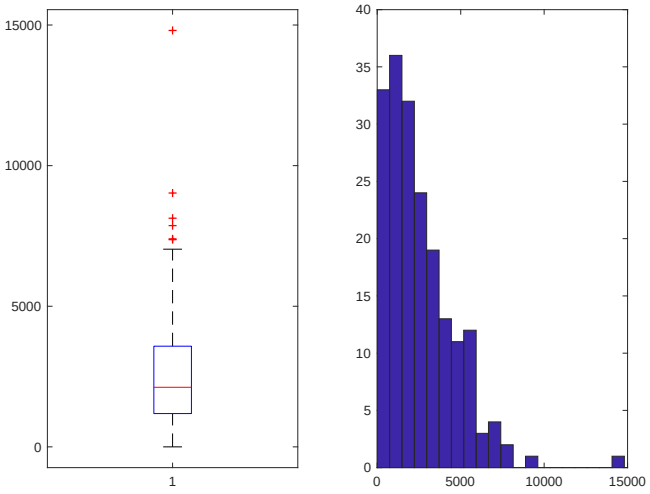

Figura 171. Detrazione METALLO TERMICO-SINGOLO

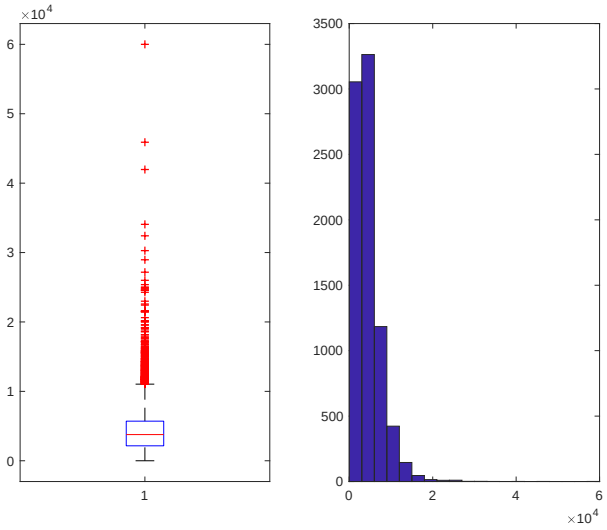

Figura 172. Detrazione METALLO TERMICO-DOPPIO

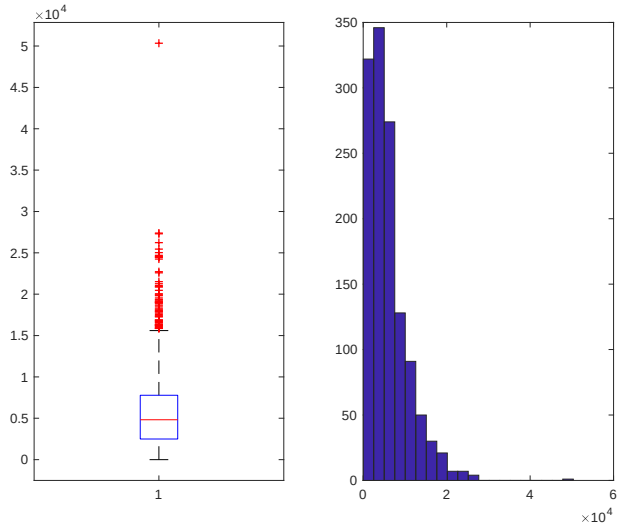

Figura 173. Detrazione METALLO TERMICO-TRIPLO

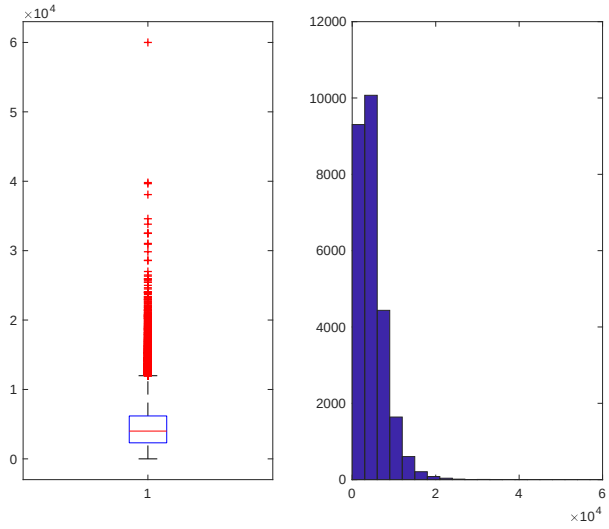

Figura 174. Detrazione METALLO TERMICO-VETRO A BASSA EMISSIONE

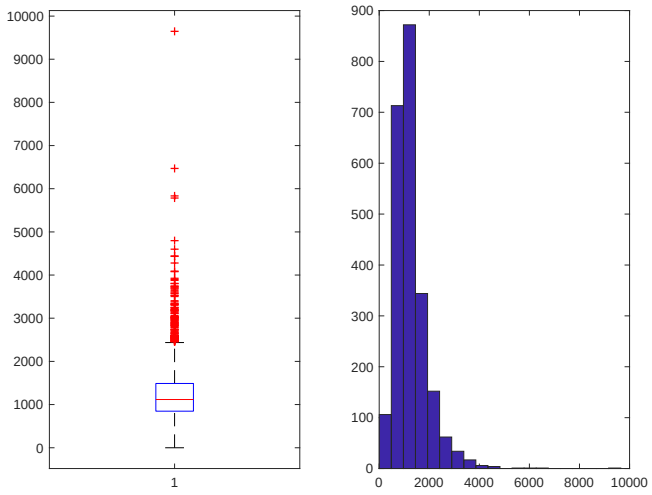

Figura 175. Detrazione METALLO TERMICO-VETRO NON ESISTENTE

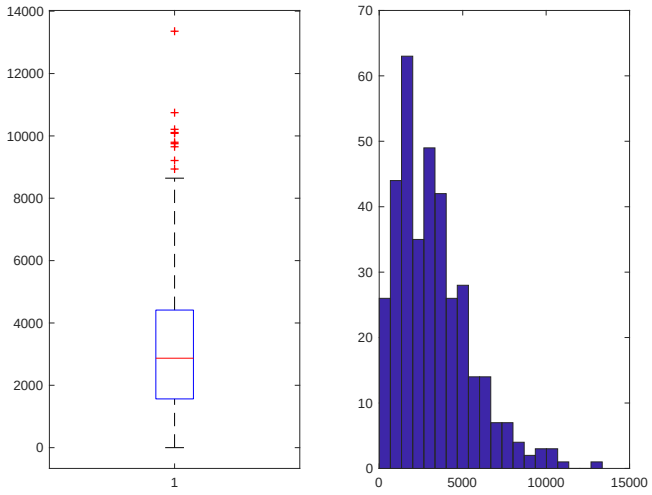

Figura 176. Detrazione PVC-SINGOLO

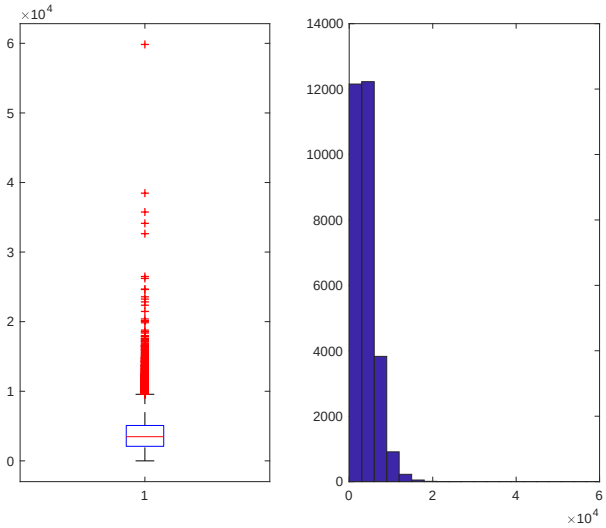

Figura 177. Detrazione PVC-DOPPIO

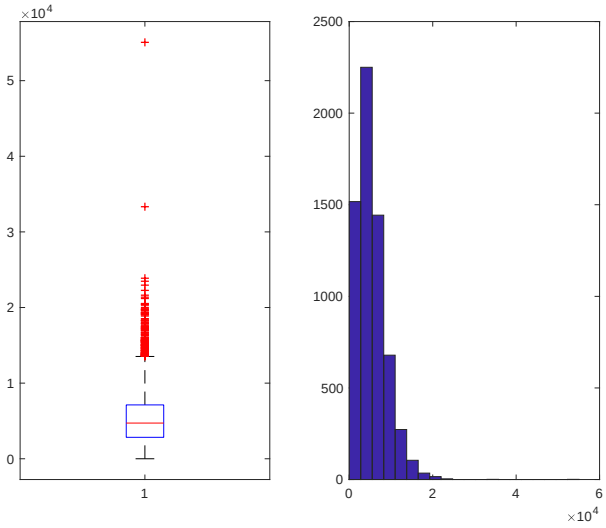

Figura 178. Detrazione PVC-TRIPLO

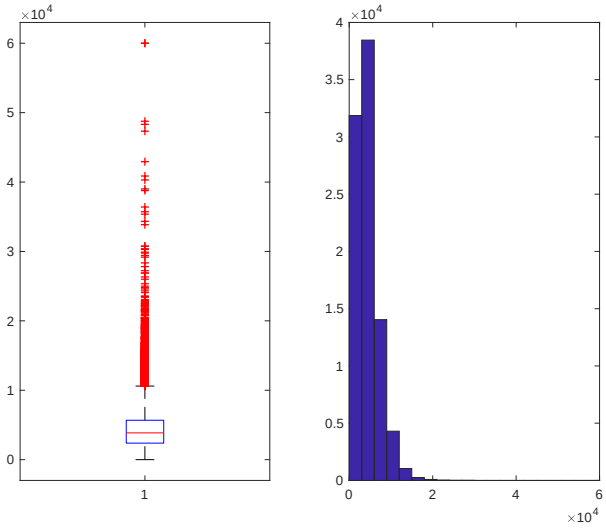

Figura 179. Detrazione PVC-VETRO A BASSA EMISSIONE

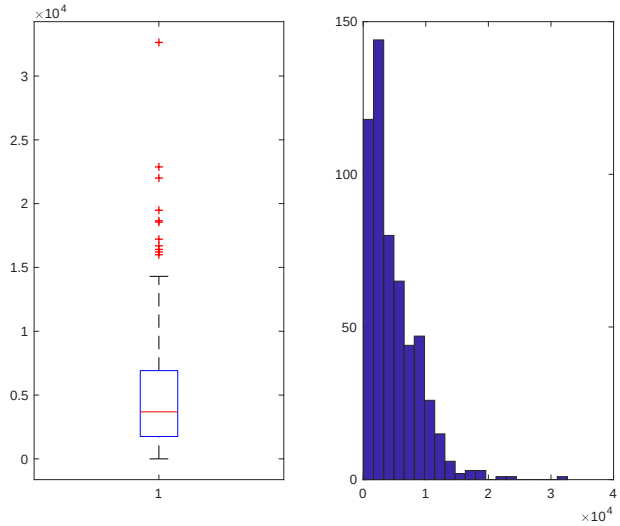

Figura 180. Detrazione PVC-VETRO NON ESISTENTE

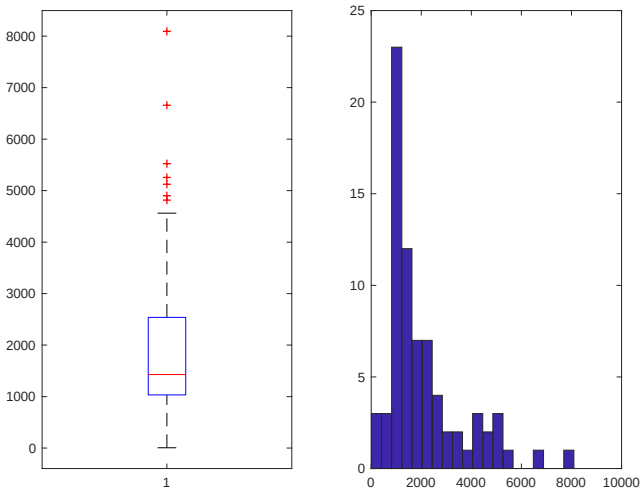

Figura 181. Detrazione MISTO-SINGOLO

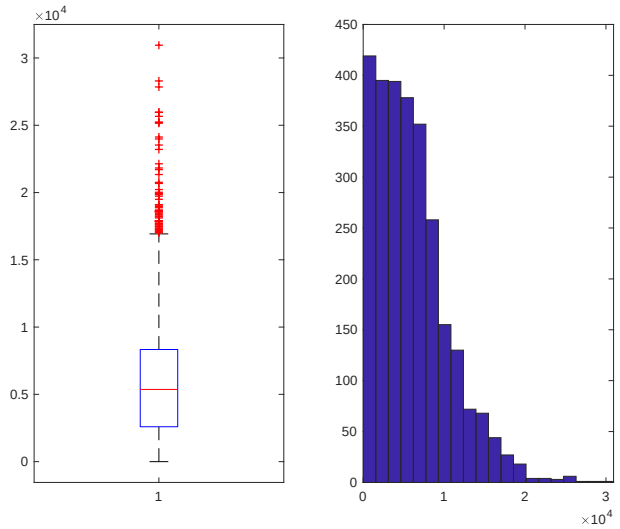

Figura 182. Detrazione MISTO-DOPPIO

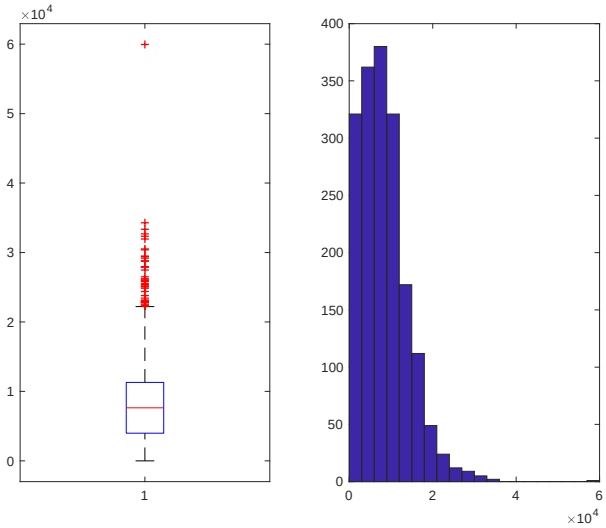

Figura 183. Detrazione MISTO-TRIPLO

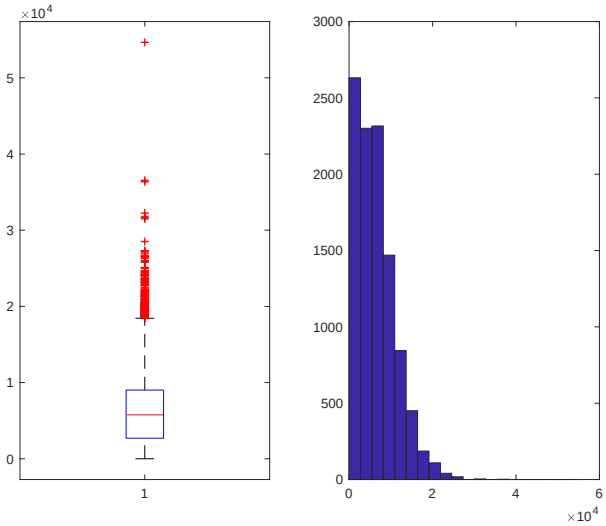

Figura 184. Detrazione MISTO-VETRO A BASSA EMISSIONE

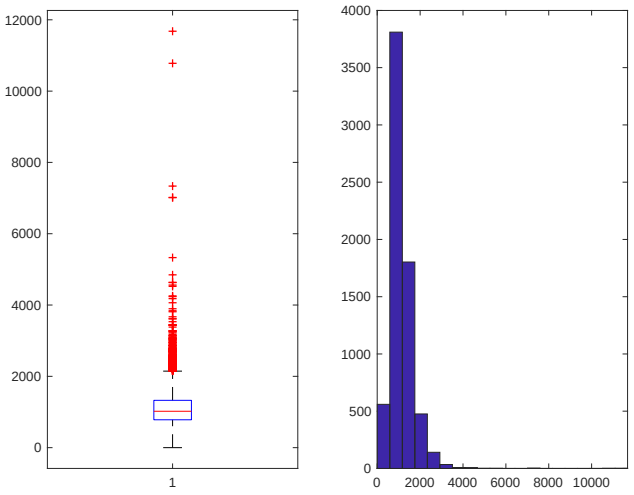

Figura 185. Detrazione MISTO-VETRO NON ESISTENTE

Una rapida considerazione finale arriva dal confronto della somma iniziale del Risparmio e del Costo con la somma delle medesime variabili dopo tutte le imputazioni svolte:

- Risparmio Iniziale: 39159034906.1904 kWh/anno
- Risparmio Finale: 613493437.013551 kWh/anno
- Costo Iniziale: 1834856757.00634 €
- Costo Finale: 1870107289.67201 €

1.4.4 Comma 345c

Dopo la fase di pulizia e ricodifica del database, la procedura di individuazione e correzione dei dati mancanti e dei dati anomali ha visto la creazione di quattro programmi MATLAB, uno che è alla base di tutto lo studio e da dove vengono richiamati gli altri tre che invece si occupano della fase di imputazione (ognuno di essi in base alla natura della variabile studiata).

Le prime operazioni svolte hanno visto la creazione di indicatori della tipologia di intervento e sulla base di questi sono state individuate alcune sottopopolazioni per rendere l'imputazione dei dati anomali più precisa ed accurata. I primi risultati ottenuti sono:

- 7955 imputazioni per la sottopopolazione RESIDENZIALE (numerosità della popolazione 80863);
- 1106 imputazioni per la sottopopolazione NON RESIDENZIALE (numerosità 4090).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

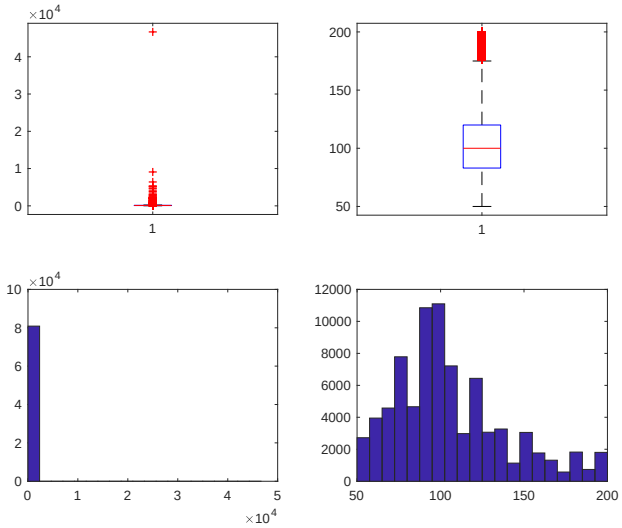

Figura 186. Superficie RES

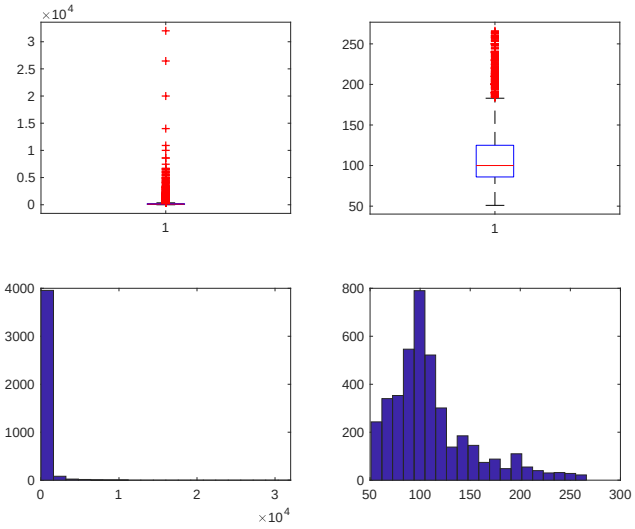

Figura 187. Superficie NORES

Per ogni sottopopolazione sono state prese in considerazione le variabili RISPARMIO, COSTO (Costo intervento + Costo professionale), COSTO/RISPARMIO per verificare ulteriormente eventuali casi anomali sulla variabile COSTO e DETRAZIONE.
Tutte le variabili sono state studiate dopo averle normalizzate per "numero di unità immobiliare" e per la variabile "superficie".
Per le variabili RISPARMIO, COSTO e COSTO/RISPARMIO, i dati anomali sono stati individuati ed imputati tramite due programmi MATLAB simili a quello utilizzato per la variabile "superficie" in modo da rispettare sia la natura delle variabili stesse che lo scopo finale dell'analisi.
Per quanto riguarda la variabile RISPARMIO i dati presentavano l'anomala presenza del valore "0" per 42110 unità su 47674, questo ha portato alla necessità di effetturare un studio preliminare sulla variabile in questione per riportare i valori all'interno di un intervallo reale. Per questa operazione di imputazione si è considerata una procedura di tipo "donatore" restringendoci ai soli valori positivi per ogni sottopopolazione già individuata.

I risultati qui proposti sono da riferirsi alla fase d'imputazione svolta sul dataset già pulito da questa anomalia:

- RISPARMIO: 4362 imputazioni (RES) e 221 imputazioni (NORES);
- COSTO: 12295 imputazioni (RES) e 815 imputazioni (NORES);
- COSTO/RISPARMIO per imputare COSTO: 11416 imputazioni (RES) e 845 imputazioni (NORES).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

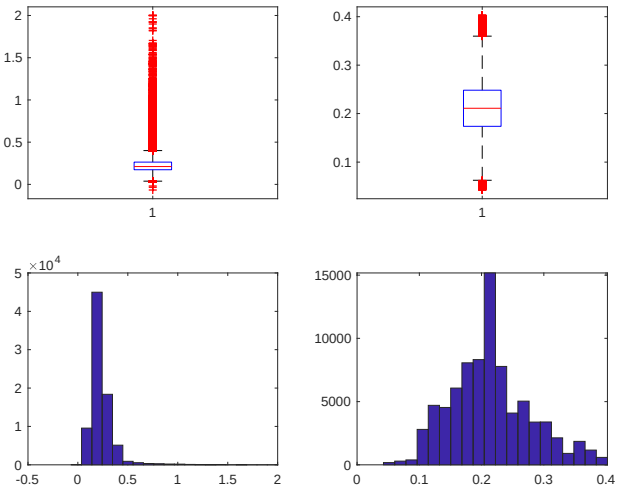

Figura 188. Risparmio RES (Normalizzato per Superficie)

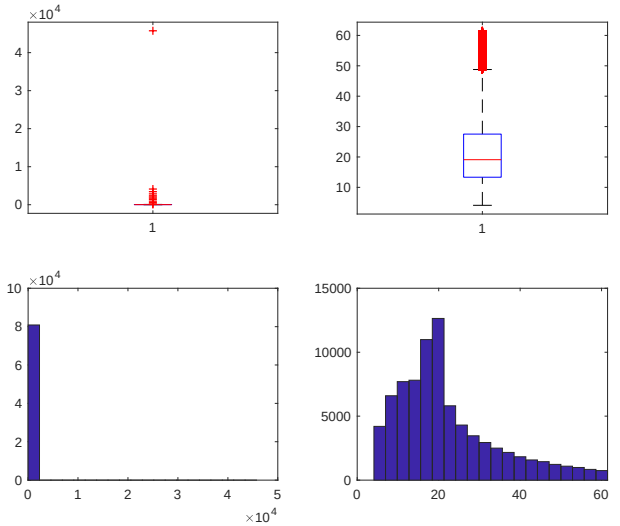

Figura 189. Costo RES

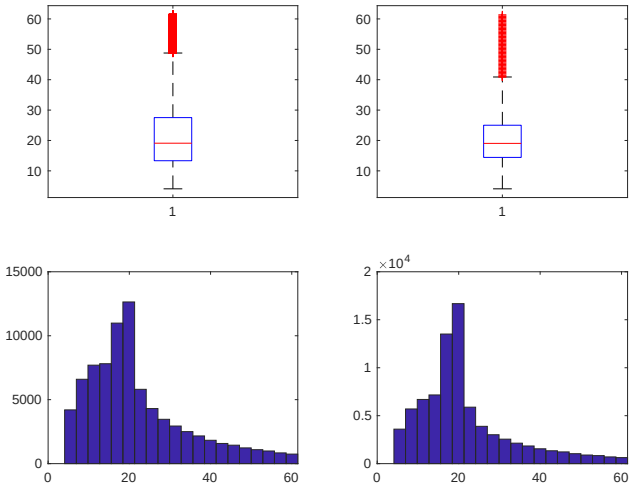

Figura 190. Costo dopo studio su Costo/Risparmio RES

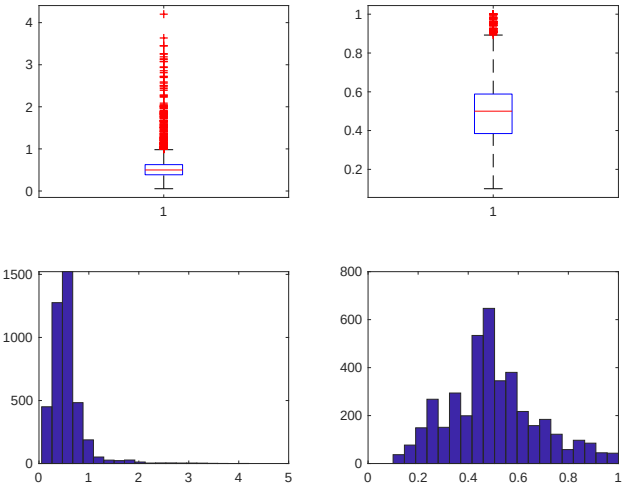

Figura 191. Risparmio NORES (Normalizzato per superficie)

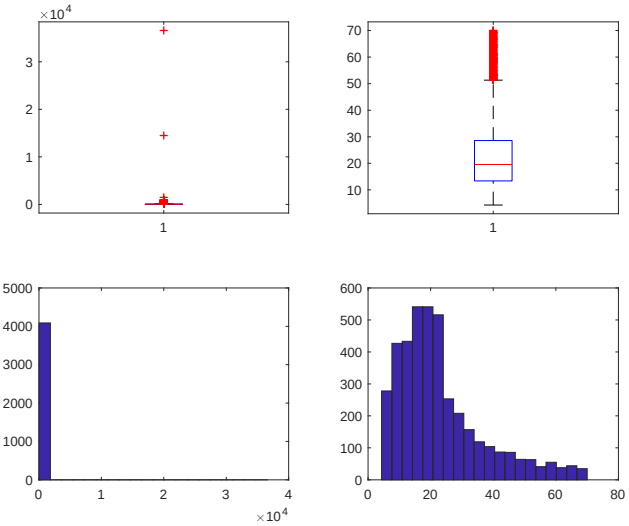

Figura 192. Costo NORES

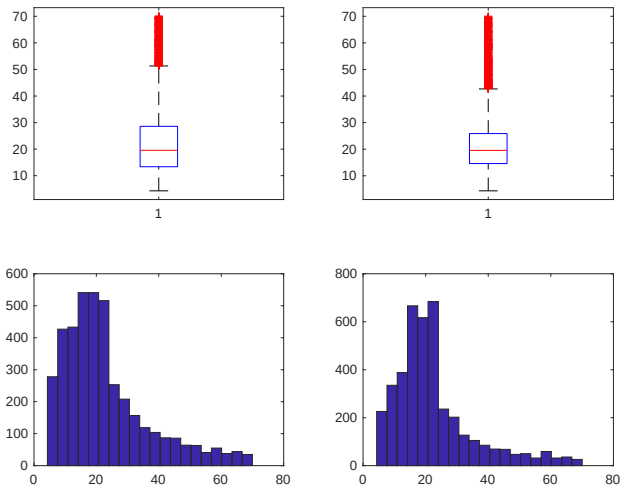

Figura 193. Costo dopo studio su Costo/Risparmio NORES

Ai fini di una corretta analisi è stato necessario effettuare una terza fase di controllo che ha visto come protagoniste le variabili d'interesse normalizzate per superficie.

La variabile DETRAZIONE è stata imputata secondo la regola:

detrazione= 0.65*costo

e sono stati individuate e conseguentemente imputati:

- 15021 dati anomali (RES) di cui 13500 per imputazioni svolte su COSTO,
- 939 dati anomali (NORES) di cui 870 per imputazioni svolte su COSTO.

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

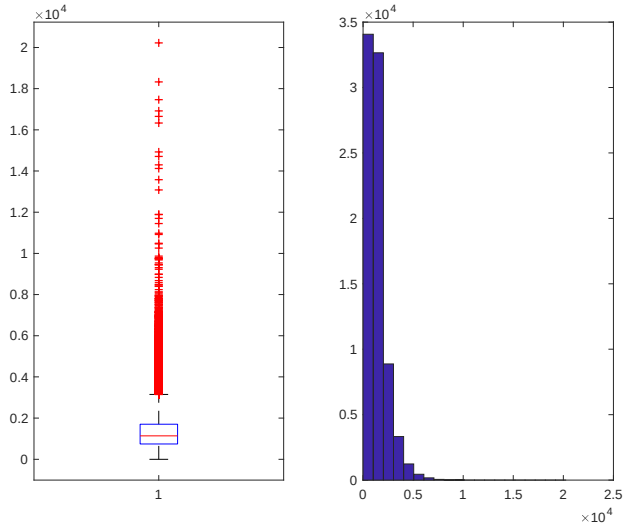

Figura 194. Detrazione RES

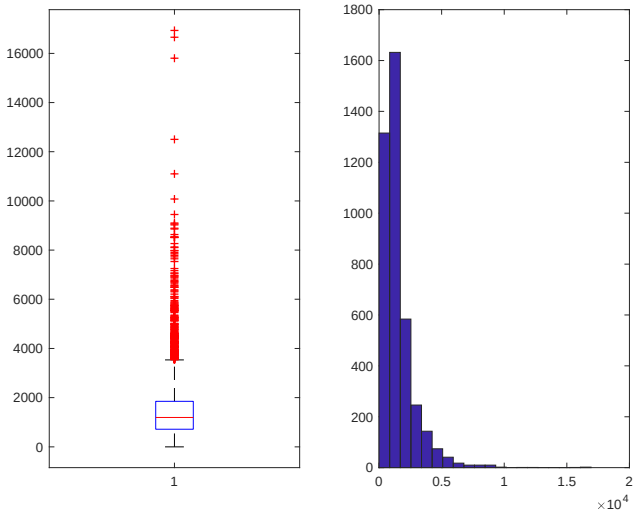

Figura 195. Detrazione NORES

Una rapida considerazione finale arriva dal confronto della somma iniziale del Risparmio e del Costo con la somma delle medesime variabili dopo tutte le imputazioni svolte:

- Risparmio Iniziale: 17356981.1595999 kWh/anno
- Risparmio Finale: 21110915.3766636 kWh/anno
- Costo Iniziale: 294600993.772395 €
- Costo Finale: 211105235.596634 €

1.4.5 Comma 346

Dopo la fase di pulizia e ricodifica del database, la procedura di individuazione e correzione dei dati mancanti e dei dati anomali ha visto la creazione di cinque programmi MATLAB, uno che è alla base di tutto lo studio e da dove vengono richiamati gli altri tre che invece si occupano della fase di imputazione (ognuno di essi in base alla natura della variabile studiata).

Come in ogni altro comma si sono per primi creati gli indicatori per individuare le sottopopolazioni oggette dello studio.

Per quanto riguarda la variabile superficie, il primo risultato importante è:

- 873 imputazioni per la sottopopolazione RESIDENZIALE (numerosità della popolazione 7889);
- 142 imputazioni per la sottopopolazione NON RESIDENZIALE (numerosità 347).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione della variabile "superficie" per le 4 sottopopolazioni cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

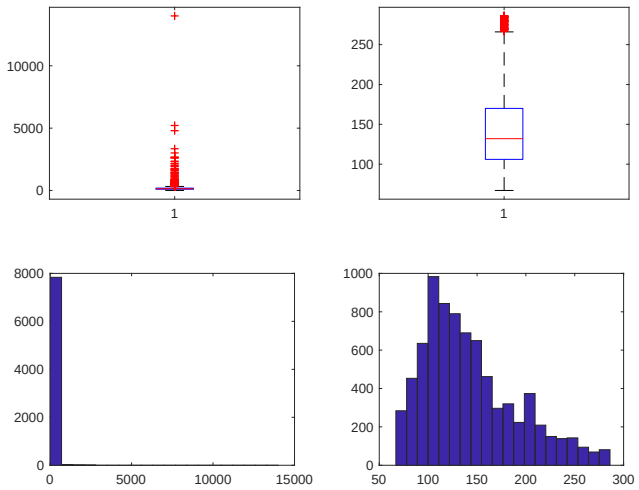

Figura 196. Superficie RES

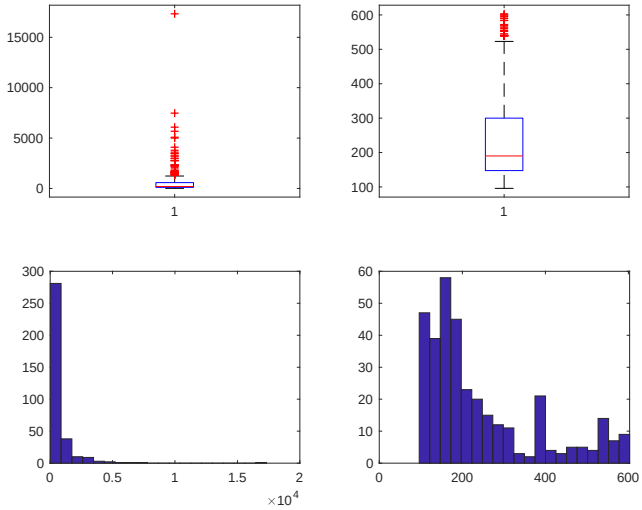

Figura 197. Superficie NORES

Per ogni sottopopolazione sono state prese in considerazione le variabili RISPARMIO, COSTO (Costo intervento + Costo professionale), COSTO/RISPARMIO per verificare ulteriormente eventuali casi anomali sulla variabile COSTO e DETRAZIONE.
Tutte le variabili sono state studiate dopo averle normalizzate per "numero di unità immobiliare" e per "superficie".
Per le variabili RISPARMIO, COSTO e COSTO/RISPARMIO, i dati anomali sono stati individuati ed imputati tramite due programmi MATLAB simili a quello utilizzato per la variabile "superficie" in modo da rispettare sia la natura delle variabili stesse che lo scopo finale dell'analisi.

I risultati più immediati sono:

- RISPARMIO: 784 imputazioni (RES) e 48 imputazioni (NORES);
- COSTO: 770 imputazioni (RES) e 50 imputazioni (NORES);
- COSTO/RISPARMIO per imputare COSTO: 893 imputazioni (RES) e 43 imputazioni (NORES).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

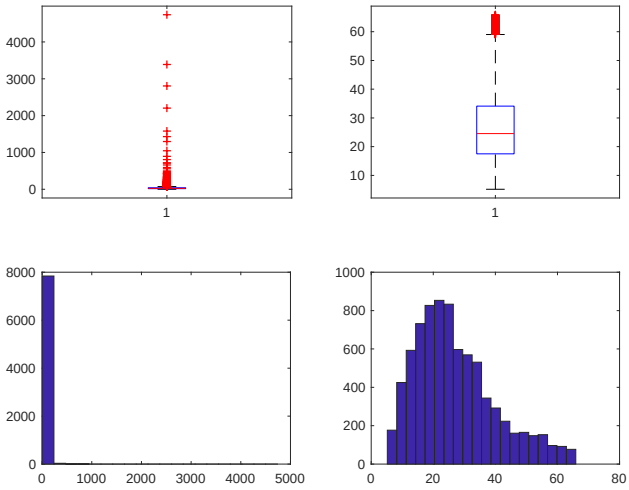

Figura 198. Risparmio RES

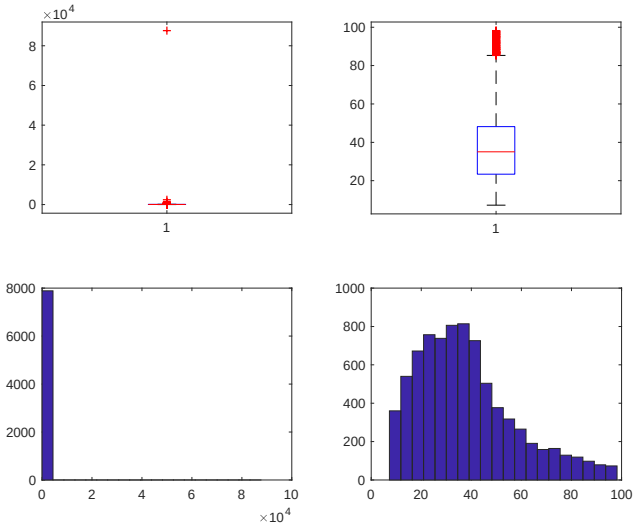

Figura 199. Costo RES

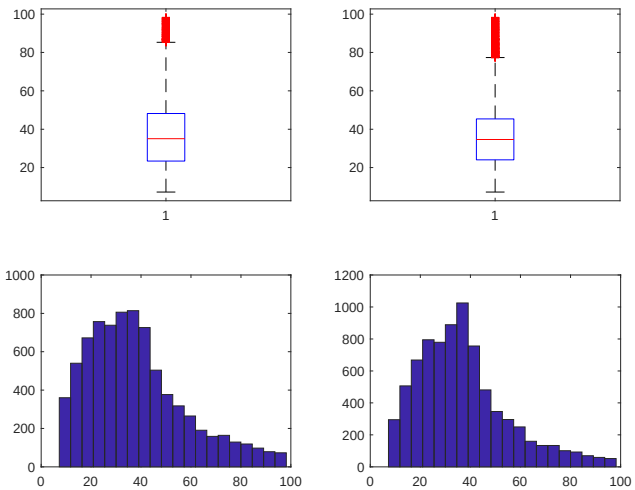

Figura 200. Costo dopo studio su Costo/Risparmio RES

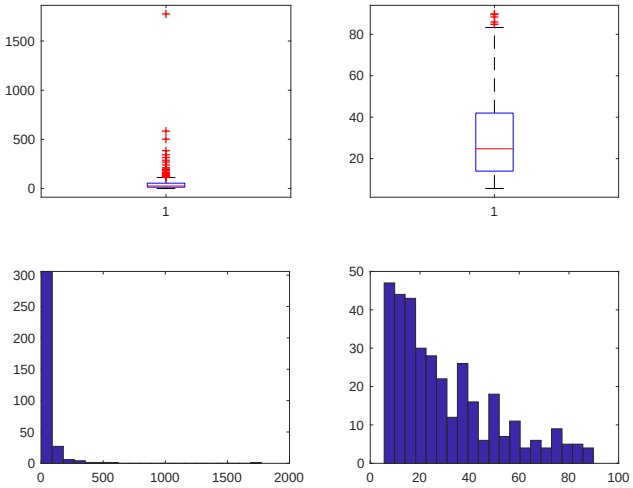

Figura 201. Risparmio NORES

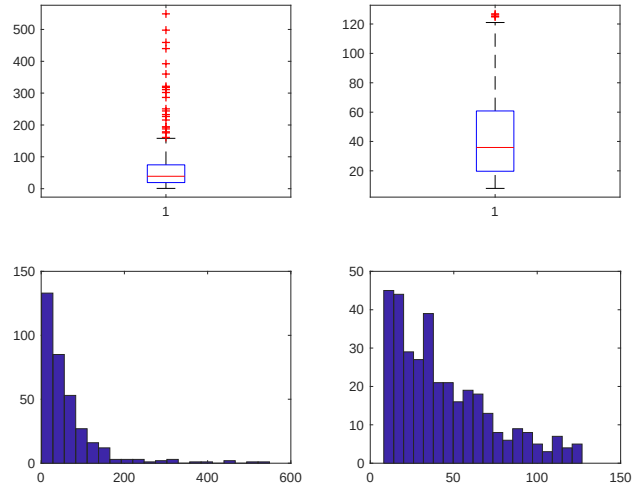

Figura 202. Costo NORES

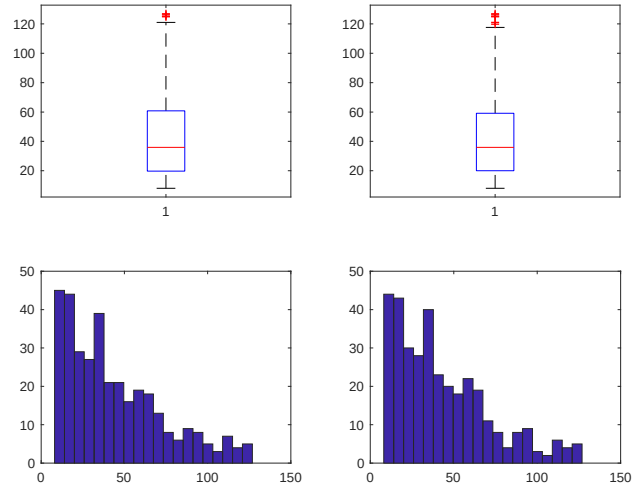

Figura 203. Costo dopo studio su Costo/Risparmio NORES

La variabile DETRAZIONE è stata imputata secondo la regola:

detrazione= 0.65*costo

e sono stati individuate e conseguentemente imputati:

- 1074 dati anomali (RES) di cui 986 per imputazioni svolte su COSTO,
- 55 dati anomali (NORES) di cui 52 per imputazioni svolte su COSTO.

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

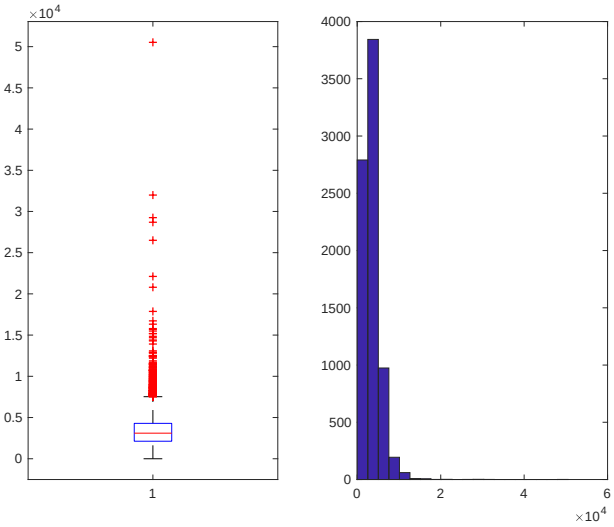

Figura 204. Detrazione RES

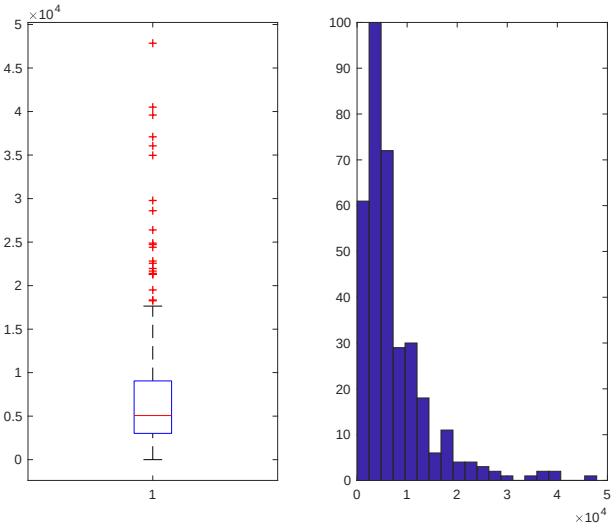

Figura 205. Detrazione NORES

Una rapida considerazione finale arriva dal confronto della somma iniziale del Risparmio e del Costo con la somma delle medesime variabili dopo tutte le imputazioni svolte:

- Risparmio Iniziale: 46377197.8076993 kWh/anno

  Risparmio Finale: 35453724.8298168 kWh/anno

- Costo Iniziale: 71389233.3715999 €

Costo Finale: 49166804.8916269 €

1.4.6 Comma 347

Dopo la fase di pulizia e ricodifica del database, la procedura di individuazione e correzione dei dati mancanti e dei dati anomali ha visto la creazione di tre programmi MATLAB, uno che è alla base di tutto lo studio e da dove vengono richiamati gli altri due che invece si occupano della fase di imputazione (ognuno di essi in base alla natura della variabile studiata).

Sono state individuate 5 sottopopolazioni sulla base della tipologia d'intervento

- 3514 imputazioni per la sottopopolazione POMPC (pompa di calore, numerosità 15439 casi),
- 9699 imputazioni per la sottopopolazione COND (impianto a caldaia a condensazione, numerosità 65297 casi);
- 354 imputazioni sulla "superficie totale" per la sottopopolazione ALTRO (tipo di impianto non specificato, numerosità 2118 casi),
- 550 imputazioni sulla "superficie totale" per la sottopopolazione BIO (tipo di impianto non specificato, numerosità 4696 casi),
- 8 imputazioni sulla "superficie totale" per la sottopopolazione GEO (tipo di impianto non specificato, numerosità 60 casi).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

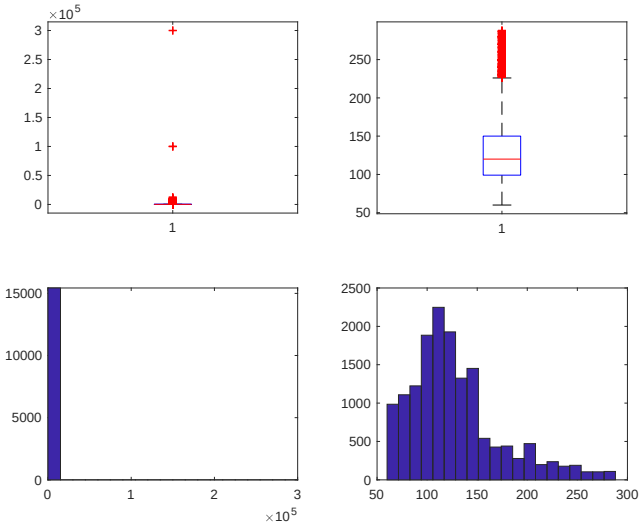

Figura 206. Superficie POMPC

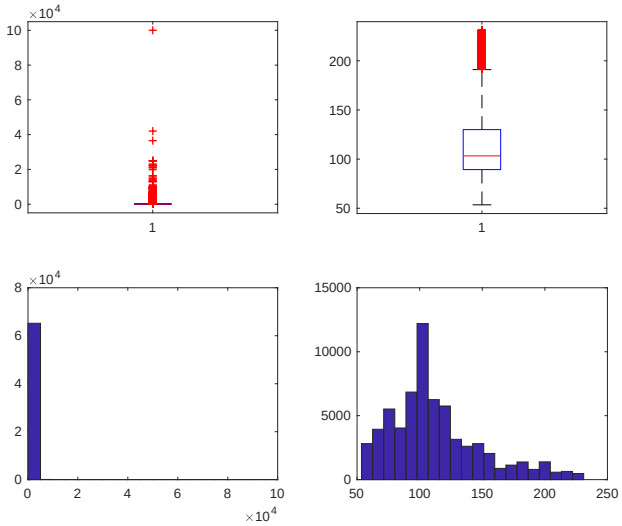

Figura 207. Superficie COND

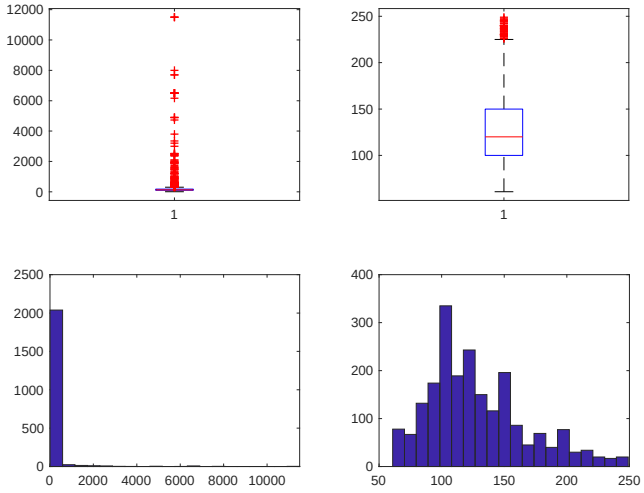

Figura 208. Superficie ALTRO

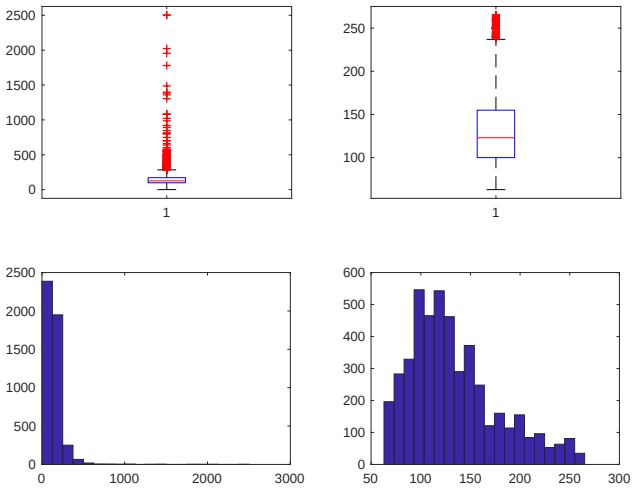

Figura 209. Superficie BIO

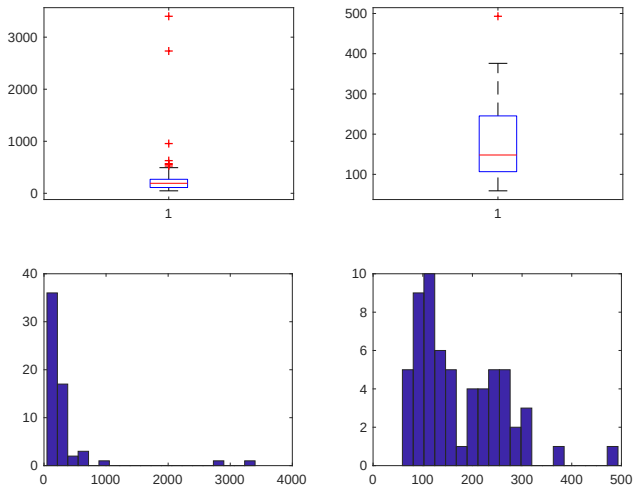

Figura 210. Superficie GEO

Per ogni sottopopolazione sono state prese in considerazione le variabili RISPARMIO, COSTO (Costo intervento + Costo professionale) e DETRAZIONE.
Tutte le variabili sono state studiate dopo averle normalizzate per "numero di unità immobiliare" E per la variabile "superficie".
Per le variabili RISPARMIO e COSTO, i dati anomali sono stati individuati ed imputati tramite due programmi MATLAB simili a quello utilizzato per la variabile "superficie" in modo da rispettare sia la natura delle variabili stesse che lo scopo finale dell'analisi.

I risultati più immediati sono:

- RISPARMIO: 4958 imputazioni (POMPC), 13900 imputazioni (COND), 863 imputazioni (ALTRO), 1784 imputazioni (BIO) e 26 imputazioni (GEO);
- COSTO: 3934 imputazioni (POMPC), 10860 imputazioni (COND), 534 imputazioni (ALTRO), 618 imputazioni (BIO) e 15 imputazioni (GEO);
- COSTO/RISPARMIO per imputare COSTO: 3692 imputazioni (POMPC), 12874 imputazioni (COND), 756 imputazioni (ALTRO), 1643 imputazioni (BIO) e 16 imputazioni (GEO).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

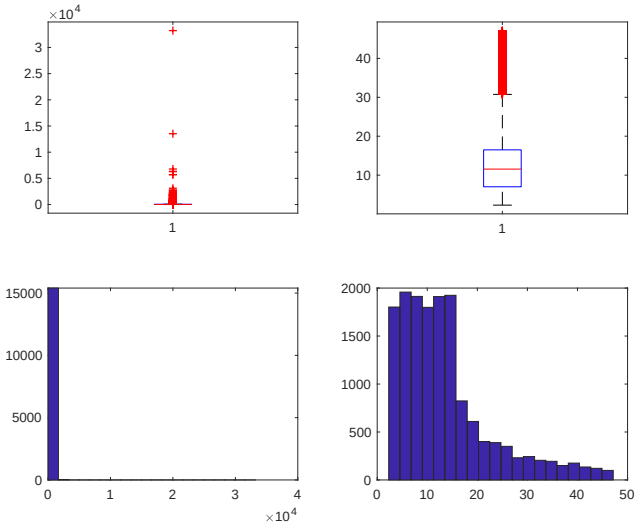

Figura 211. Risparmio POMPC

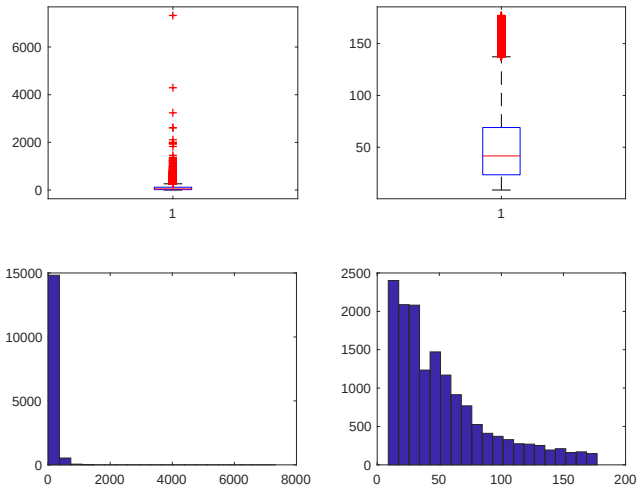

Figura 212. Costo POMPC

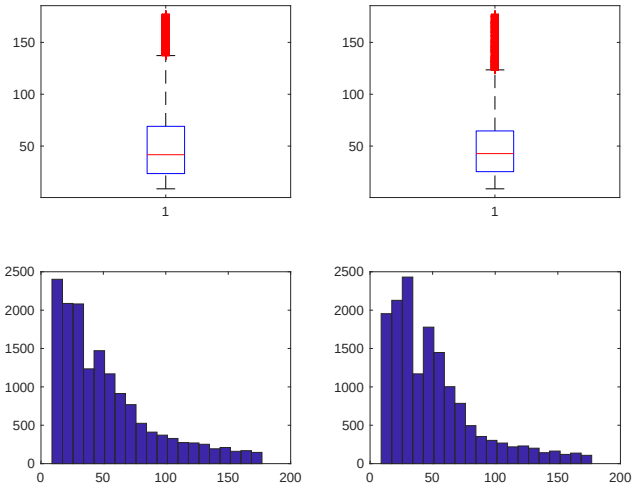

Figura 213. Costo dopo studio su Costo/Risparmio POMPC

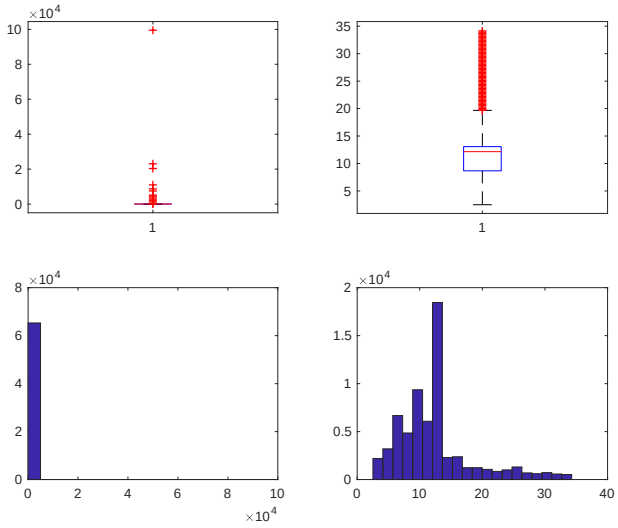

Figura 214. Risparmio COND

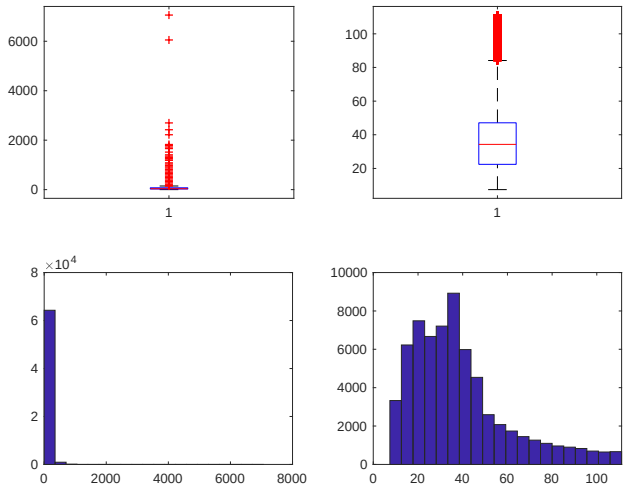

Figura 215. Costo COND

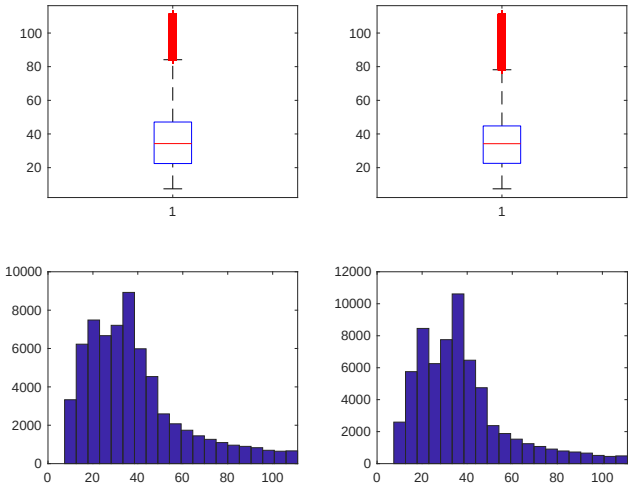

Figura 216. Costo dopo studio su Costo/Risparmio COND

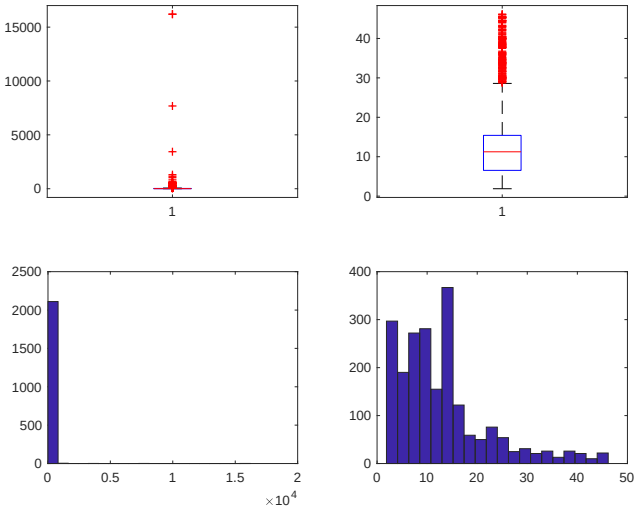

Figura 217. Risparmio ALTRO

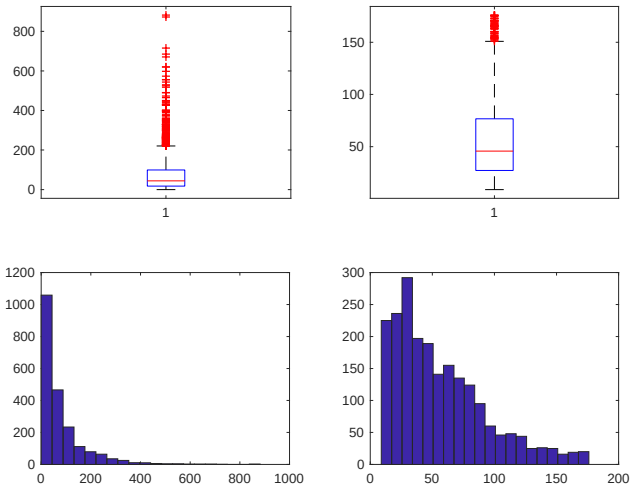

Figura 218. Costo ALTRO

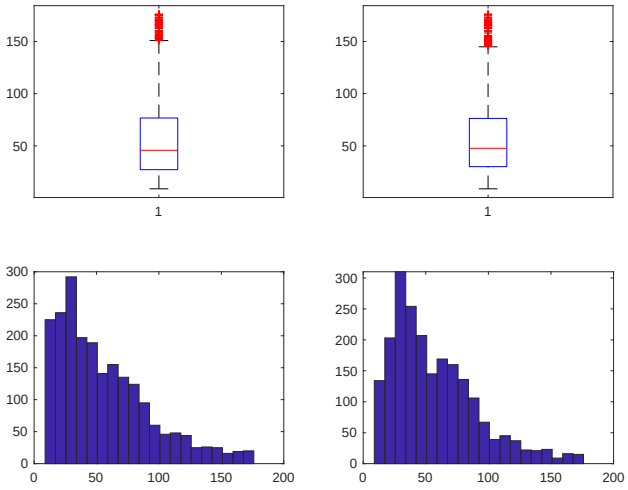

Figura 219. Costo dopo studio su Costo/Risparmio ALTRO

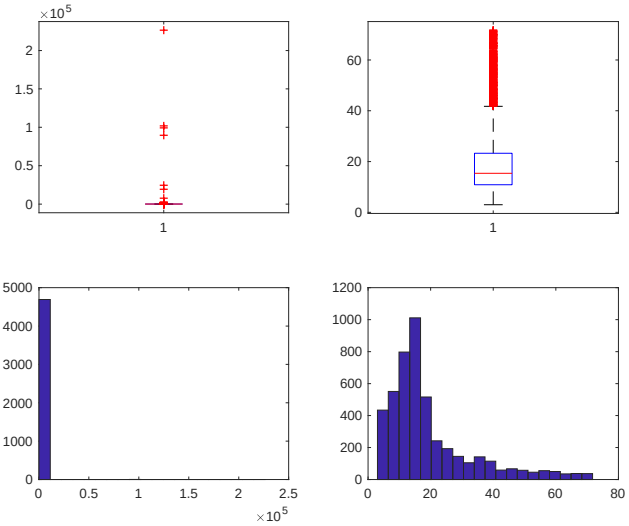

Figura 220. Risparmio BIO

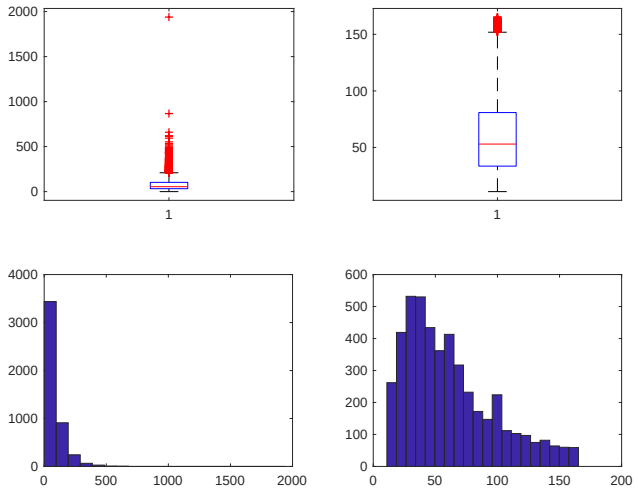

Figura 221. Costo BIO

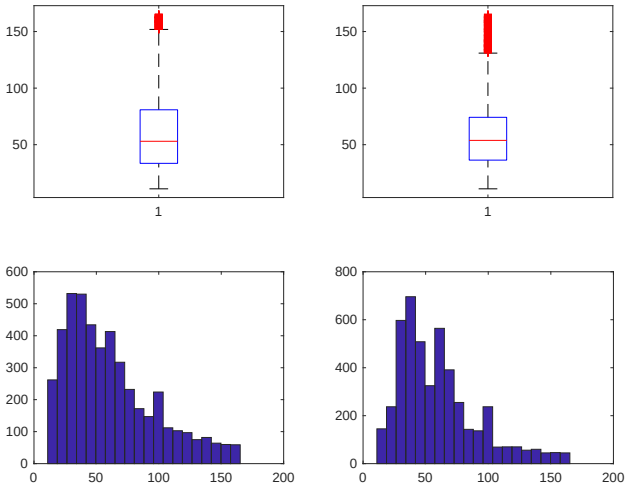

Figura 222. Costo dopo studio su Costo/Risparmio BIO

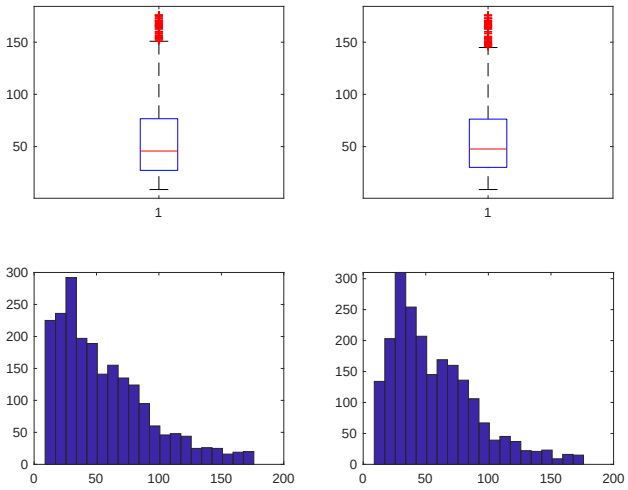

Figura 223. Risparmio GEO

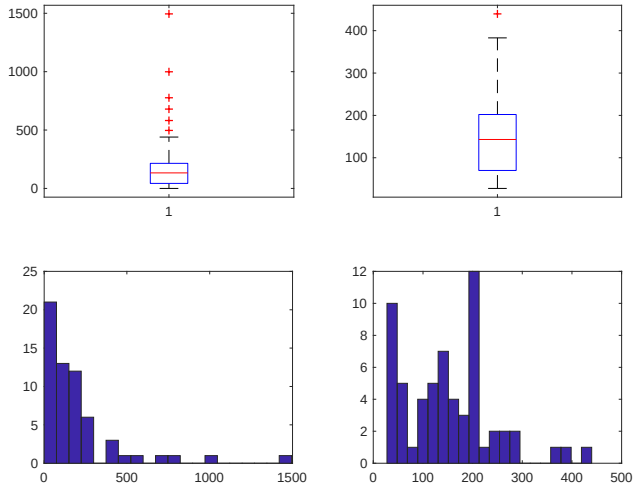

Figura 224. Costo GEO

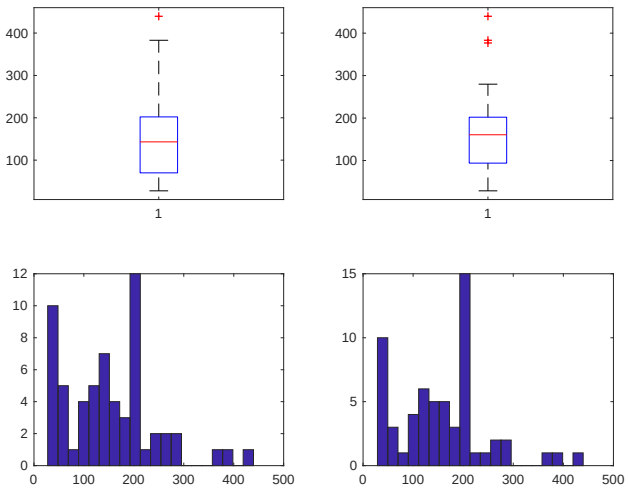

Figura 225. Costo dopo studio su Costo/Risparmio GEO

La variabile DETRAZIONE non è stata condotta nessuna analisi per questo comma.

Una rapida considerazione finale arriva dal confronto della somma iniziale del Risparmio e del Costo con la somma delle medesime variabili dopo tutte le imputazioni svolte:

- Risparmio Iniziale: 623568390.222525 kWh/anno
- Risparmio Finale: 268368649.960858 kWh/anno
- Costo Iniziale: 967413768.940323 €

  Costo Finale: 740011961.848018 €

1.4.7 Comma BA

Dopo la fase di pulizia e ricodifica del database, la procedura di individuazione e correzione dei dati mancanti e dei dati anomali ha visto la creazione di cinque programmi MATLAB, uno che è alla base di tutto lo studio e da dove vengono richiamati gli altri tre che invece si occupano della fase di imputazione (ognuno di essi in base alla natura della variabile studiata).

Come in ogni altro comma si sono per primi creati gli indicatori per individuare le sottopopolazioni oggette dello studio.

Per quanto riguarda la variabile superficie, il primo risultato importante è:

- 449 imputazioni per la sottopopolazione RESIDENZIALE (numerosità della popolazione 1595);
- 105 imputazioni per la sottopopolazione NON RESIDENZIALE (numerosità 317).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione della variabile "superficie" per le 2 sottopopolazioni cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

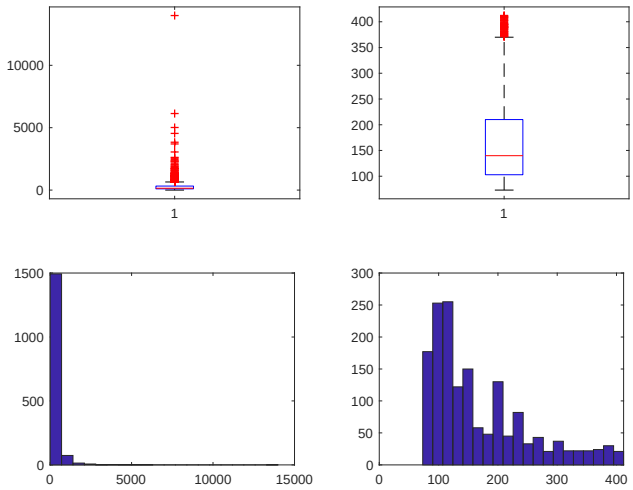

Figura 226. Superficie RES

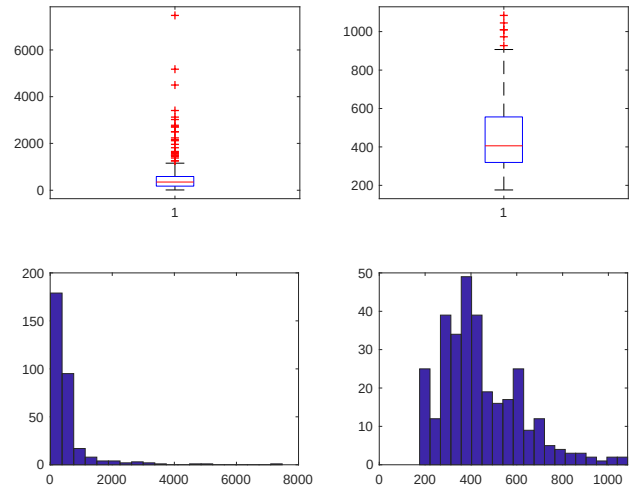

Figura 227. Superficie NORES

Per ogni sottopopolazione sono state prese in considerazione le variabili RISPARMIO, COSTO (Costo intervento + Costo professionale), COSTO/RISPARMIO per verificare ulteriormente eventuali casi anomali sulla variabile COSTO e DETRAZIONE.
Tutte le variabili sono state studiate dopo averle normalizzate per "numero di unità immobiliare" e per "superficie".
Per le variabili RISPARMIO, COSTO e COSTO/RISPARMIO, i dati anomali sono stati individuati ed imputati tramite due programmi MATLAB simili a quello utilizzato per la variabile "superficie" in modo da rispettare sia la natura delle variabili stesse che lo scopo finale dell'analisi.

I risultati più immediati sono:

- RISPARMIO: 402 imputazioni (RES) e 49 imputazioni (NORES);
- COSTO: 396 imputazioni (RES) e 40 imputazioni (NORES);

- COSTO/RISPARMIO per imputare COSTO: 470 imputazioni (RES) e 67 imputazioni (NORES).

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

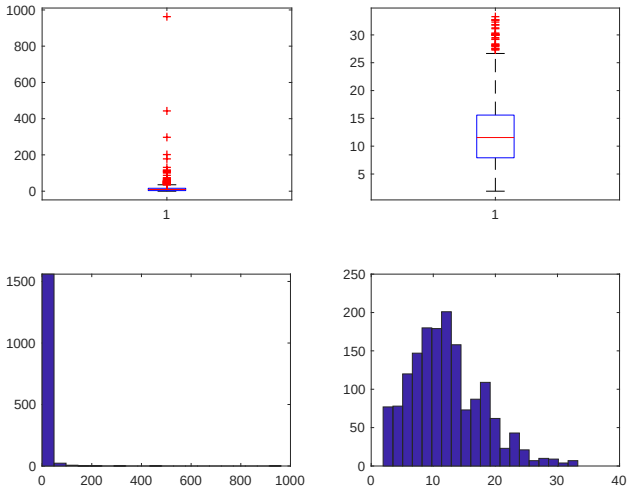

Figura 228. Risparmio RES

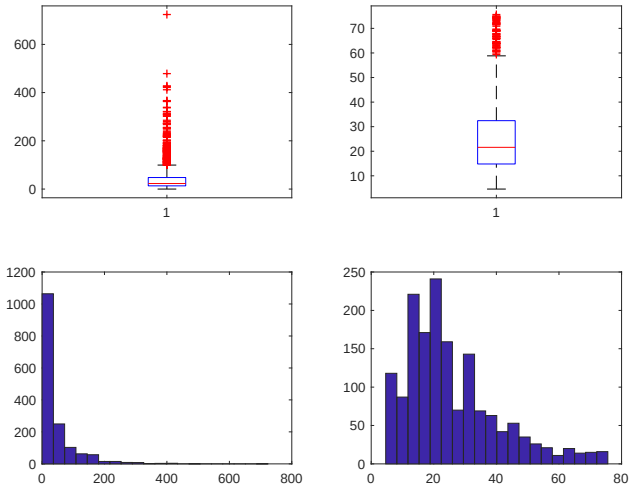

Figura 229. Costo RES

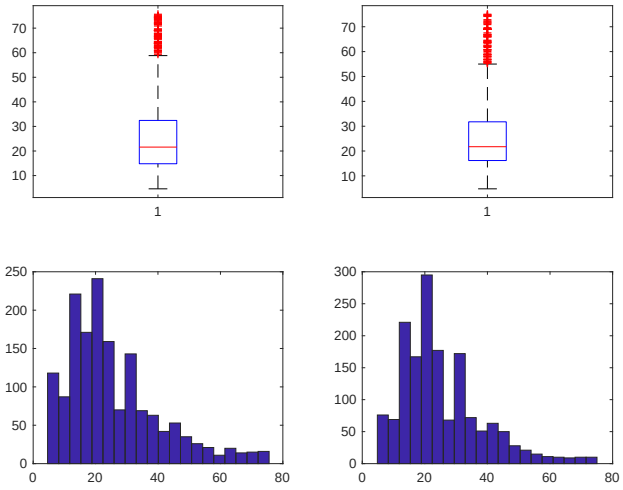

Figura 230. Costo dopo studio su Costo/Risparmio RES

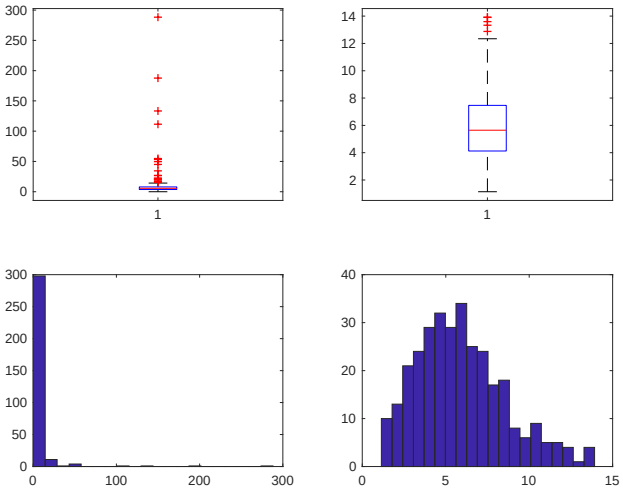

Figura 231. Risparmio NORES

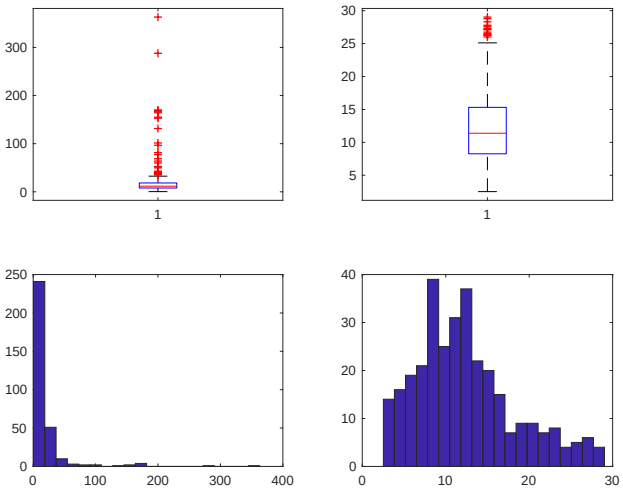

Figura 232. Costo NORES

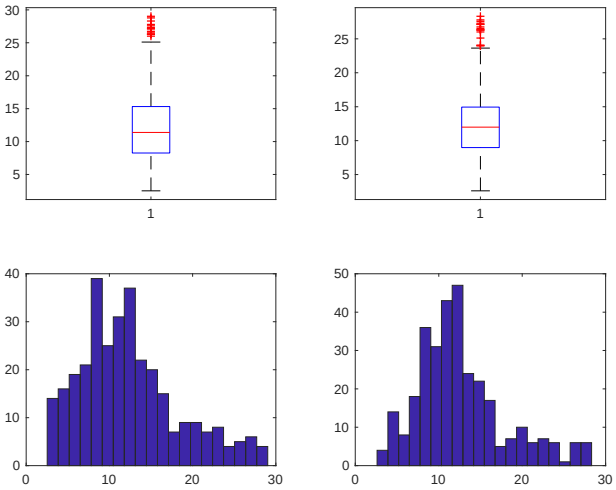

Figura 233. Costo dopo studio su Costo/Risparmio NORES

La variabile DETRAZIONE è stata imputata secondo la regola:

detrazione= 0.65*costo

e sono stati individuate e conseguentemente imputati:

- 414 dati anomali (RES) di cui 399 per imputazioni svolte su COSTO,
- 42 dati anomali (NORES) di cui 41 per imputazioni svolte su COSTO.

I grafici (Box-Plot ed Istogramma) qui riportati evidenziano come la distribuzione delle variabili studiate cambi radicalmente prima (Sinistra) e dopo (Destra) lo studio.

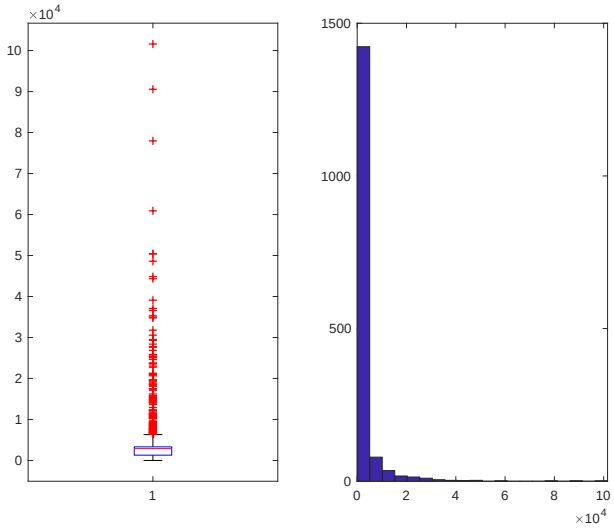

Figura 234. Detrazione RES

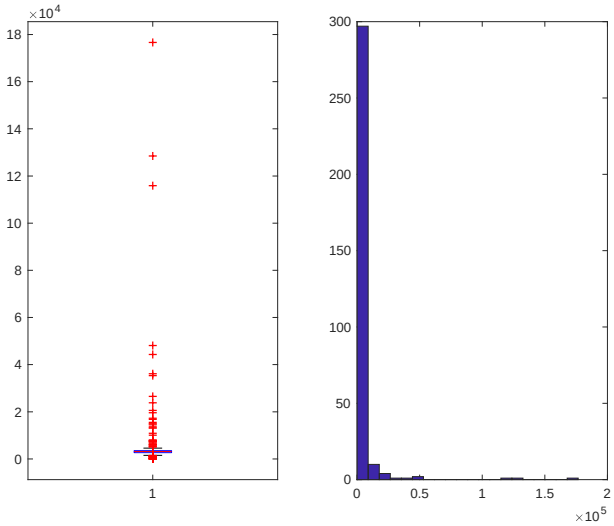

Figura 235. Detrazione NORES

Una rapida considerazione finale arriva dal confronto della somma iniziale del Risparmio e del Costo con la somma delle medesime variabili dopo tutte le imputazioni svolte:

- Risparmio Iniziale: 8733056.61400000 kWh/anno
- Risparmio Finale: 9180349.64502670 kWh/anno
- Costo Iniziale: 18383916.8830000 €
- Costo Finale: 18099515.3648859 €

2 Conclusioni

L'ammontare complessivo per tutti i commi del risparmio energetico annuo e dei costi d'intervento può essere riassunto in queste tabelle.

RIEPILOGO AMMONTARE RISPARMIO E COSTI ANNO 2017

	344	345a	345b	345c	346	347	BA
Risparmio pre (kWh)/anno	189.890.947	602.472.589	39.159.034.906	17.356.981	6.926.223.630	623.568.390	8.733.056
Risparmio Post (kWh)/anno	90.956.886	429.915.065	613.493.437	21.110.915	53.511.507	268.368.649	9.180.349
Costi pre (€)	326.322.805	818.220.880	1.834.856.757	294.600.993	72.237.591	967.413.768	18.383.916
Costi post (€)	351.640.050	876.822.217	1.870.107.289	211.105.235	74.396.691	740.011.961	18.099.515

E' evidente come le correzioni svolte in questo lavoro portino ad un cambiamento consistente di entrambi i totali per ogni comma.

3 Riferimenti bibliografici

1) Achtert, E., Kriegel, H.-P., Reichert, L., Schubert, E., Wojdanowski, R., Zimek, A. 2010. Visual Evaluation of Outlier Detection Models. In Proc. International Conference on Database Systems for Advanced Applications (DASFAA), Tsukuba, Japan.
2) Aggarwal, C.C. and Yu, P.S. 2000. Outlier detection for high dimensional data. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Dallas, TX.
3) Angiulli, F. and Pizzuti, C. 2002. Fast outlier detection in high dimensional spaces. In Proc. European Conf. on Principles of Knowledge Discovery and Data Mining, Helsinki, Finland.
4) Arning, A., Agrawal, R., and Raghavan, P. 1996. A linear method for deviation detection in large databases. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Portland, OR.
5) Barnett, V. 1978. The study of outliers: purpose and model. Applied Statistics, 27(3), 242–250.
6) Bay, S.D. and Schwabacher, M. 2003. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Washington, DC.
7) Breunig, M.M., Kriegel, H.-P., Ng, R.T., and Sander, J. 1999. OPTICS-OF: identifying local outliers. In Proc. European Conf. on Principles of Data Mining and Knowledge Discovery (PKDD), Prague, Czech Republic.
8) Breunig, M.M., Kriegel, H.-P., Ng, R.T., and Sander, J. 2000. LOF: identifying density-based local outliers. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Dallas, TX.
9) Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Portland, OR.
10) Fan, H., Zaïane, O., Foss, A., and Wu, J. 2006. A nonparametric outlier detection for efficiently discovering top-n outliers from engineering data. In Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Singapore.
11) Ghoting, A., Parthasarathy, S., and Otey, M. 2006. Fast mining of distance-based outliers in high dimensional spaces. In Proc. SIAM Int. Conf. on Data Mining (SDM), Bethesda, ML.
12) Hautamaki, V., Karkkainen, I., and Franti, P. 2004. Outlier detection using k-nearest neighbour graph. In Proc. IEEE Int. Conf. on Pattern Recognition (ICPR), Cambridge, UK.
13) Hawkins, D. 1980. Identification of Outliers. Chapman and Hall.
14) Jin, W., Tung, A., and Han, J. 2001. Mining top-n local outliers in large databases. In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), San Francisco, CA.
15) Jin, W., Tung, A., Han, J., and Wang, W. 2006. Ranking outliers using symmetric neighborhood relationship. In Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Singapore.
16) Johnson, T., Kwok, I., and Ng, R.T. 1998. Fast computation of 2-dimensional depth contours. In Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), New York, NY.
17) Knorr, E.M. and Ng, R.T. 1997. A unified approach for mining outliers. In Proc. Conf. of the Centre for Advanced Studies on Collaborative Research (CASCON), Toronto, Canada.
18) Knorr, E.M. and NG, R.T. 1998. Algorithms for mining distance-based outliers in large datasets. In Proc. Int. Conf. on Very Large Data Bases (VLDB), New York, NY.
19) Knorr, E.M. and Ng, R.T. 1999. Finding intensional knowledge of distance-based outliers. In Proc. Int. Conf. on Very Large Data Bases (VLDB), Edinburgh, Scotland.

20) Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. 2009. Outlier detection in axis-parallel subspaces of high dimensional data. In Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand.
21) Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. 2009a. LoOP: Local Outlier Probabilities. In Proc. ACM Conference on Information and Knowledge Management (CIKM), Hong Kong, China.
22) Kriegel, H.-P., Schubert, M., and Zimek, A. 2008. Angle-based outlier detection, In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), Las Vegas, NV.
23) McCallum, A., Nigam, K., and Ungar, L.H. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), Boston, MA.
24) Papadimitriou, S., Kitagawa, H., Gibbons, P., and Faloutsos, C. 2003. LOCI: Fast outlier detection using the local correlation integral. In Proc. IEEE Int. Conf. on Data Engineering (ICDE), Hong Kong, China.
25) Pei, Y., Zaiane, O., and Gao, Y. 2006. An efficient reference-based approach to outlier detection in large datasets. In Proc. 6th Int. Conf. on Data Mining (ICDM), Hong Kong, China.
26) Preparata, F. and Shamos, M. 1988. Computational Geometry: an Introduction. Springer Verlag.
27) Ramaswamy, S. Rastogi, R. and Shim, K. 2000. Efficient algorithms for mining outliers from large data sets. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Dallas, TX.
28) Rousseeuw, P.J. and Leroy, A.M. 1987. Robust Regression and Outlier Detection. John Wiley.
29) Ruts, I. and Rousseeuw, P.J. 1996. Computing depth contours of bivariate point clouds. Computational Statistics and Data Analysis, 23, 153–168.
30) Tao Y., Xiao, X. and Zhou, S. 2006. Mining distance-based outliers from large databases in any metric space. In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), New York, NY.
31) Tan, P.-N., Steinbach, M., and Kumar, V. 2006. Introduction to Data Mining. Addison Wesley.
32) Tang, J., Chen, Z., Fu, A.W.-C., and Cheung, D.W. 2002. Enhancing effectiveness of outlier detections for low density patterns. In Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Taipei, Taiwan.
33) Tukey, J. 1977. Exploratory Data Analysis. Addison-Wesley.
34) Zhang, T., Ramakrishnan, R., Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Montreal, Canada.
35) Aguinis, H. (2004). Regression analysis for categorical moderators. New York, NY: Guilford. Aguinis, H., Forcum, L. E., & Joo, H. (in press). Using market basket analysis in management research. Journal of Management. doi:10.1177/0149206312466147
36) Aguinis, H., Pierce, C. A., Bosco, F. A., & Muslin, I. S. (2009). First decade of Organizational Research Methods: Trends in design, measurement, and data-analysis topics. Organizational Research Methods, 12, 69-112.
37) Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric sci- ence: Reporting significant research results with rigor, relevance, and practical impact in mind. Organizational Research Methods, 13, 515-539.

38) Amiot, C. E., Terry, D. J., Jimmieson, N. L., & Callan, V. J. (2006). A longitudinal investigation of coping processes during a merger: Implications for job satisfaction and organizational identification. Journal of Management, 32, 552-574.
39) Aytug, Z. G., Rothstein, H. R., Zhou, W., & Kern, M. C. (2012). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. Organizational Research Methods, 15, 103-133.
40) Barnett, V., & Lewis, T. (1994). Outliers in statistical data (3rd ed.). New York, NY: John Wiley.
41) Becker, C., & Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. Journal of the American Statistical Association, 94, 947-955.
42) Belsley, D. A., Kuh, E., & Welsh, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York, NY: John Wiley.
43) Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009a). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. Journal of Applied Psychology, 94, 567-582.
44) Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P. E. (2009b). Transparency should trump trust: Rejoinder to McConnell and Leibold (2009) and Ziegert and Hanges (2009). Journal of Applied Psychology, 94, 598-603.
45) Bollen, K. A., & Jackman, R. W. (1990). Regression diagnostics: An expository treatment of outliers and influential cases. In J. Fox & J. S. Long (Eds.), Modern methods of data analysis (pp. 257-291). Newbury Park, CA: Sage.
46) Brown, D. J., Cober, R. T., Kane, K., Levy, P. E., & Shalhoop, J. (2006). Proactive personality and the successful job search: A field investigation with college graduates. Journal of Applied Psychology, 91, 717-726.
47) Brutus, S., Aguinis, H., & Wassmer, U. (2013). Self-reported limitations and future directions in scholarly reports: Analysis and recommendations. Journal of Management, 39, 48-75. doi:10.1177/0149206312455245
48) Byrne, B. M. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. International Journal of Testing, 1, 55-86.
49) Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
50) Cortina, J. M. (2002). Big things have small beginnings: An assortment of ‘‘minor’’ methodological misunder- standings. Journal of Management, 28, 339-362.
51) Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. American Psychologist, 55, 34-43.
52) Edwards, J. R., & Cable, D. M. (2009). The value of value congruence. Journal of Applied Psychology, 94, 654-677.
53) Edwards, J. R., Cable, D. M., Williamson, I. O., Lambert, L. S., & Shipp, A. J. (2006). The phenomenology of fit: Linking the person and environment to the subjective experience of person-environment fit. Journal of Applied Psychology, 91, 802-827.
54) Fidell, L. S., & Tabachnick, B. G. (2003). Preparatory data analysis. In J. A. Schinka & W. F. Velicer (Eds.), Handbook of psychology: Research methods in psychology (Vol. 2, pp. 115-141). New York, NY: John Wiley.
55) Gladwell, M. (2008). Outliers: The story of success. New York, NY: Little, Brown. Godfrey, P. C., Merrill, C. B., & Hansen, J. M. (2009). The relationship between corporate social responsibility and shareholder value: An empirical test of the risk management hypothesis. Strategic Management Journal, 30, 425-445.

56) Goerzen, A., & Beamish, P. W. (2005). The effect of alliance network diversity on multinational enterprise performance. Strategic Management Journal, 26, 333-354.
57) Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. Technometrics, 11, 1-21.
58) Hawawini, G., Subramanian, V., & Verdin, P. (2003). Is performance driven by industry- or firm-specific factors? A new look at the evidence. Strategic Management Journal, 24, 1-16.
59) Hitt, M. A., Harrison, J. S., Ireland, R. D., & Best, A. (1998). Attributes of successful and unsuccessful acquisitions of US firms. British Journal of Management, 9, 91-114.
60) Hollenbeck, J. R., DeRue, D. S., & Mannor, M. (2006). Statistical power and parameter stability when subjects are few and tests are many: Comment on Peterson, Smith, Martorana, and Owens (2003). Journal of Applied Psychology, 91, 1-5. Hox, J. J. (2010). Multilevel analysis: Techniques and applications (2nd ed.). New York, NY: Routledge.
61) Huffman, M. L., Cohen, P. N., & Pearlman, J. (2010). Engendering change: Organizational dynamics and workplace gender desegregation, 1975–2005. Administrative Science Quarterly, 55, 255-277.
62) Hunter, J. E., & Schmidt, F. L. (2004). Methods of meta-analysis: Correcting error and bias in research findings (2nd ed.). Thousand Oaks, CA: Sage. Jasso, G. (1985). Marital coital frequency and the passage of time: Estimating the separate effects of spouses' ages and marital duration, birth and marriage cohorts, and period influences. American Sociological Review, 50, 224-241.
63) Kahn, J. R., & Udry, J. R. (1986). Marital coital frequency: Unnoticed outliers and unspecified interactions lead to erroneous conclusions. American Sociological Review, 51, 734-737.
64) Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. Organizational Research Methods, 15, 722-752.
65) Kulich, C., Trojanowski, G., Ryan, M. K., Haslam, S. A., & Renneboog, L. D. R. (2011). Who gets the carrot and who gets the stick? Evidence of gender disparities in executive remuneration. Strategic Management Journal, 32, 301-321.
66) Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). Applied linear statistical models (5th ed.). Boston: McGraw-Hill/Irwin. Langford, I. H., & Lewis, T. (1998). Outliers in multilevel data. Journal of the Royal Statistical Society, Series A, 161, 121-160.
67) Leung, K. (2011). Presenting post hoc hypotheses as a priori: Ethical and theoretical issues. Management and Locke, E. A. (2007). The case for inductive theory building. Journal of Management, 33, 867-890.
68) Martin, M. A., & Roberts, S. (2010). Jackknife-after-bootstrap regression influence diagnostics. Journal of Nonparametric Statistics, 22, 257-269. Martin, N., & Pardo, L. (2009). On the asymptotic distribution of Cook's distance in logistic regression models. Journal of Applied Statistics, 36, 1119-1146.
69) Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. Journal of Applied Psychology, 97, 951-966. McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. Journal of Experimental Social Psychology, 37, 435-442.
70) McConnell, A. R., & Leibold, J. M. (2009). Weak criticisms and selective evidence: Reply to Blanton et al. (2009). Journal of Applied Psychology, 94, 583-589.
71) McNamara, G., Aime, F., & Vaaler, P. M. (2005). Is performance driven by industry- or firm-specific factors? A response to Hawawini, Subramanian, and Verdin. Strategic

Management Journal, 26, 1075-1081. Mohrman, S. A., & Lawler, E. E., III. (2012). Generating knowledge that drives change. Academy of Management Perspectives, 26, 41-51.
72) O'Boyle, E., Jr., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. Personnel Psychology, 65, 79-119.
73) Orr, J. M., Sackett, P. R., & DuBois, C. L. Z. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. Personnel Psychology, 44, 473-486.
74) Pek, J., & MacCallum, R. C. (2011). Sensitivity analysis in structural equation models: Cases and their influence. Multivariate Behavioral Research, 46, 202-228.
75) Peterson, R. S., Smith, D. B., Martorana, P. V., & Owens, P. D. (2003). The impact of chief executive officer personality on top management team dynamics: One mechanism by which leadership affects organizational performance. Journal of Applied Psychology, 88, 795-808.
76) Pierce, J. R., & Aguinis, H. (in press). The too-much-of-a-good-thing effect in management. Journal of Management. doi:10.1177/0149206311410060
77) Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., Jr., & Du Toit, M. (2004). HLM 6: Hierarchical linear and nonlinear modeling. Lincolnwood, IL: Scientific Software International.
78) Reuer, J. J., & Arin˜o, A. (2002). Contractual renegotiations in strategic alliances. Journal of Management, 28, 47-68.
79) Rousseeuw, P. J., & Leroy, A. M. (2003). Robust regression and outlier detection. Hoboken, NJ: John Wiley.
80) Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. American Psychologist, 55, 5-14.
81) Shepherd, D. A., & Sutcliffe, K. M. (2011). Inductive top-down theorizing: A source of new theories of organization. Academy of Management Review, 36, 361-380.
82) Shi, L., & Chen, G. (2008). Detection of outliers in multilevel models. Journal of Statistical Planning and Inference, 138, 3189-3199.
83) Smillie, L. D., Yeo, G. B., Furnham, A. F., & Jackson, C. J. (2006). Benefits of all work and no play: The relationship between neuroticism and performance as a function of resource allocation. Journal of Applied Psychology, 91, 139-155.
84) Snijders, T. A. B., & Berkhof, J. (2008). Diagnostic checks for multilevel models. In J. de Leeuw & E. Meijer (Eds.), Handbook of multilevel analysis (pp. 141-176). New York, NY: Springer.
85) Snijders, T. A. B., & Bosker, R. J. (2012). Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.).
86) Thousand Oaks, CA: Sage. St. John, C. H., & Harrison, J. S. (1999). Manufacturing-based relatedness, synergy, and coordination. Strategic Management Journal, 20, 129-145.
87) Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (5th ed.). Boston, MA: Pearson.
88) Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconcep- tions. Annual Review of Clinical Psychology, 1, 31-65.
89) Van Dick, R., Van Knippenberg, D., Kerschreiter, R., Hertel, G., & Wieseke, J. (2008). Interactive effects of work group and organizational identification on job satisfaction and extra-role behavior. Journal of Vocational Behavior, 72, 388-399.

90) Wanberg, C. R., Glomb, T. M., Song, Z., & Sorenson, S. (2005). Job-search persistence during unemployment: A 10-wave longitudinal study. Journal of Applied Psychology, 90, 411-430.
91) Wiggins, R. R., & Ruefli, T. W. (2005). Schumpeter's ghost: Is hypercompetition making the best of times shorter? Strategic Management Journal, 26, 887-911.
92) Worren, N., Moore, K., & Cardona, P. (2002). Modularity, strategic flexibility, and firm performance: A study of the home appliance industry. Strategic Management Journal, 23, 1123-1140.
93) Yuan, K.-H., & Bentler, P. M. (1998). Structural equation modeling with robust covariances. Sociological Methodology, 28, 363-396.
94) Yuan, K.-H., & Zhong, X. (2008). Outliers, leverage observations, and influential cases in factor analysis: Using robust procedures to minimize their effect. Sociological Methodology, 38, 329-368.
95) Zhang, Z., & Yuan, K.-H. (2012). semdiag: Structural equation modeling diagnostics (R Package Version 0.1.2) [Computer software manual]. Retrieved from http://cran.at.r-project.org/web/packages/semdiag/semdiag.pdf
96) Zhong, X., & Yuan, K.-H. (2011). Bias and efficiency in structural equation modeling: Maximum likelihood versus robust methods. Multivariate Behavioral Research, 46, 229-265.
97) Ziegert, J. C., & Hanges, P. J. (2009). Strong rebuttal for weak criticisms: Reply to Blanton et al. (2009).

4 Appendice

4.1 Curriculum scientifico del gruppo di lavoro

MAURIZIO VICHI
maurizio.vichi@uniroma1.it
Born: September 13, 1959;
https://scholar.google.it/ Maurizio Vichi
RESEARCHGATE: https://www.researchgate.net/profile/Maurizio_Vichi

- JOB POSITION
Full Professor of Statistics, Sapienza University of Rome
Facoltà di Ingegneria dell'Informazione, Informatica e Statistica

- WORK EXPERIENCE
** From 2000 – up today: Full Professor of Statistics*
Department of Statistical Sciences, Sapienza University of Rome, P.le A. Moro 5, 00185 Roma
Deputy-director of Department of Statistical Sciences (from 2013), Scientific Sector Statistics
From 1992 – to 1999: Associate Professor of Statistics
Department of Quantitative Methods and Economic Theory, University of Chieti "Gabriele D'Annunzio", Scientific Sector Statistics
** From 1990 – to 1991: University Researcher*
Department of Statistics Probability and Applied Statistics, Faculty of Statistics, Sapienza University of Rome, Scientific Sector Statistics
** 1986 Research Fellow,*
Rutgers University of New Jersey, USA
** 1985 Research Fellow*
St. Andrews University, United Kingdom,

- SCIENTIFIC QUALIFICATIONS
** From 2012 – up-today*
President of the European Federation of National Statistical Societies (FENStatS)
During the Statistical Italian Presidency he has worked at the foundation of FENStatS (2011) with the 10 major European Societies. He has been elected first President of FENStatS in 2012, and after 2 years of his presidency, FENStatS has doubled the associations. Now, 20 national associations are affiliated to FENStatS. They almost cover the totality of European countries with a statistics association. For additional information: www.fenstats.eu
** From 2013 – up-today*
Deputy-chair of the European Statistics Advisory Committee (ESAC) of EU
ESAC is one of the Three Committees of the European Union that deals with statistics (European Statistical System) and in particular gives advises on the European Statistical Programme.
In ESAC he represents the Council of the European Union, and the scientific community of statisticians. He has organized the first European Conference of Statistics Stakeholders (24-25 November 2014), and will organize CESS 2016.
Info: http://epp.eurostat.ec.europa.eu/portal/page/portal/esac/composition/eleven_members
** From 2014 – to 2015*
President of International Federation of Classification Societies (IFCS)
This is the largest international association of multivariate statistics and data analysis methodologies that strongly contributes to the modernization of statistics.
Info: http://ifcs.boku.ac.at/cms/tiki-index.php
** From 2012 – to 2013*

President-elect of International Federation of Classification Societies (IFCS)
Info: http://ifcs.boku.ac.at/cms/tiki-index.php
** From 2008 – to 2012*
President of the Italian Statistical Society (SIS)
As President of the SIS, he worked with the Ministry of Education, to promote the inclusion of statistics in the programs of the school, since 2009. Now, statistics is an integral part of the mathematics courses in schools of all levels. In 2011 he organized the first Italian Day of Statistics, under the high patronage of the President of the Republic. This is now inserted in the annual events of national importance. He also promoted the Olympic games of Statistics at school, and many other cultural activities to promote the discipline. He currently participates to the Commission on users CUIS Istat. Info: www.sis-statistica.it
** From 2005 – to 2007*
President of the Section of SIS Classification and Data Analysis (CLADAG), First Coordinator and Founder of the group SIS CLADAG
** From 1998 – to 2002*
Secretary General of the Italian Statistical Society (SIS)
Unique Secretary-General to be elected also President of SIS, up to now.
As Secretary General, he founded the group CLADAG of classification and data analysis, in 1997, now, Section of the SIS, and he founded and directed the School of the Italian Statistical Society.
** From 2001 – to 2005*
Member of the Council International Association of Statistical Computing (IASC)
From 1996– to 2004 Member of the Council of International Federation of Classification Societies (IFCS)

- EDITOR OF JOURNALS and INTERNATIONAL SERIES
** From 2007 – up-today*
Editor of Journal: Advances in Data Analysis and Classification
International Journal of Springer with impact factor 2012, 1.38
http://www.springer.com/statistics/statistical+theory+and+methods/journal/11634
** From 2001 – to 2006*
Editor of Journal: Statistical Methods and Applications
International Journal of Springer with impact factor 2012, 0.35
http://www.springer.com/statistics/journal/10260
** From 2009 – up-today Editor of the Series: Studies in Theoretical and Applied Statistics*
International Series of Springer
From 2003 – up-today
Editor of the Series :Studies in Classification, Data Analysis and Knowledge Organization
International Series of Springer

- EDUCATION AND TRAINING
** 1979 – 1983 Laurea in Scienze Statistiche e Demografiche, 110 e laude, Faculty of Statistical Sciences, Sapienza University of Rome*
** 1984 Scholarship CNR (203.10.20 del 24.5.83)*
** 1986 Scholarship CNR CNR (203.10.21 del 5.6.1985)*
**1985–1991 Scholarship LUISS anni 1985/86, 1986/87, 1987/88, 1988/89, 1990/91*

- ORGANISATIONAL/ MANAGERIAL SKILLS
He has good organizational and management skills also in reference to the organisation and management of the Italian Statistical Society that has seen a good improvement of the activities organised and careful management of the finances of the association. In addition, he has been and he is responsible of several research projects with national leadership.

- MAIN RESEARCH INTERESTS
Multivariate Statistics, Classification, Clustering, Dimensionality Reduction, Three-way Data Analysis, Structural Equation Modeling

- SCIENTIFIC PROFILE
He has good professional and scientific skills. This is confirmed by the following results: he is author of 110 publications, with 731 citations and an H-index equal to 15; of these, 40 have a high impact in the field of international multivariate statistics and in particular in the classification and data analysis, which represent the most innovative methodologies to statistically analyse data, and are useful for modernizing statistics. In this field of research, he has gained a significant international reputation . He has published in several Journal with impact factor: ADAC, CSDA, , Hournal of Applied Statistics, Journal of Classification, Metron, Psychomethika, Statistics and Computing, Statistical Modelling.
This is confirmed also by the score of 22.28 Research GATE (https://www.researchgate.net), which is higher than 72.5% of researchers around the world. Another result is his election as President of IFCS and also his position of: editor of ADAC, with the impact factor 1.4, which is at the top statistics journals in the field of classification and clustering; editor of Springer, Studies in Classification Data Analysis and Knowledge Organization and Studies in Theoretical and Applied Statistics.

- FINANCED NATIONAL AND INTERNATION RESEARCH PROJECTS
He has a large experience in the coordination of scientific projects as reported below:
** 1996-1997 Identification of Standards of Living and Poverty in South Africa, World Bank*
** 1997–1998 Coordinator of the project "New methodologies for repeated surveys: applications in Social-Economic and Demographic fields";*
** 1998–1999 Cross-national Social Capital and Poverty Survey in Uganda, PRMPO, World Bank.*
** 1998–2000 National Coordinator of CNR project "Multivariate models for analysis of data with complex structure in socio-economic-demographic fields".2001–2002 Coordinator of the Project Miur: "Methods and models for the analysis of data with complex structure". 2003–2004 (PRIN) National Coordinator of the Project (5 Universities) New statistical methods of classification and dimensional reduction for the valuation and customer satisfaction in services of public utility.*
** 2005-2006 (PRIN) National Coordinator of the Project (5 Universities) Advanced multivariate statistical methods for quality assessment in public utility services: effectiveness-efficiency, risk of the provider, customer satisfaction. This project proposes the organization of an evaluation system evidence based that includes effectiveness, efficiency, satisfaction , and risk.*
** 2008 – 2010 (PRIN) National Coordinator of the Project (5 Universities) New Multivariate Methods for Statistical Quality Assessment and Risk Analysis of Services.*
** 2010– 2013 (PRIN) Coordinator of Project Multivariate Methods for Risk Analysis*
- ORGANISED CONFERENCES
CLADAG, 1997, IFCS 1998, CLADAG, 1999, SIS 1999, 2000, 2001, 2002, COMPSTAT 2006, SIS 2008, Conference of European Statistics Stakeholders (CESS 2014), CESS 2016

- INVITED KEYNOTE SPEAKER or SPEAKER
French Classification Society 2005, Royal Statistical Society 2009, Portuguese Statistical Society 2010, German Statistical Society 2012, Portuguese Classification Society 2013, German Classification Society 2013, Hungarian Statistical Association 2013, Multivariate Group of the Spanish Statistical Association 2014, German Classification Society 2015, Hungarian Academic of Science 2015, South Africa Statistical Society 2015, International Statistical Institute 2015.

CARLO CAVICCHIA

carlo.cavicchia@uniroma1.it
Educational Back-Ground
November 2016 – October 2019 (Expected)
Universita' di Roma La Sapienza
Ph.D in Statistics Science "Scuola di Scienze Statistiche – Statistica Metodologica"

September 2013 – January 2016
Universita' di Roma La Sapienza
Master of Science Degree in Statistics (Scienze Statistiche e Decisionali)
Final Grade*: 110/110 cum laude*
Thesis*: UN INDICATORE COMPOSITO GERARCHICO PER URBES*
SAS Certification: *SAS Certified Predictive Modeler Using SAS Enterprise Miner 13 (at SAS Institute)*

September 2009 – July 2013
Universita' di Roma La Sapienza
Bachelor's Degree in Statistics (Statistica Gestionale)
Final Grade*: 105/110*
Thesis*: COSTRUZIONE DI UN INDICATORE COMPOSITO PER BES*
Erasmus Studies Program: *one semester at University College of Dublin (January 2012- May 2012)*

September 2004 – July 2009
Liceo Scientifico Statale Ignazio Vian di Bracciano
High School Diploma Scientific Studies (P.N.I. – Piano Nazionale Informatico)
Final Grade*: 83/100*

Professional Back-Ground
May 2016 – December 2016
Research Project (as person in charge)
Title*:* "Costruzione e programmazione MATLAB dell'indicatore composito, gerarchico, non-negativo, disgiunto di un insieme di variabili quantitative"