Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Civile

Tesi di Laurea Magistrale

Analisi dei guasti nella rete acquedottistica di Torino

Relatori:
Prof. Fulvio Boano
Prof. Luca Ridolfi
Ing. Marco Scibetta (SMAT)

Candidato:
Giovanni Saporito

Settembre 2018

Indice

Introduzione

Con il termine acquedotto si indica l'insieme delle opere necessarie al prelievo, trattamento, trasferimento, immagazzinamento e distribuzione della risorsa idrica alle utenze. Tale risorsa è utilizzata in modo quotidiano e continuativo, per mezzo degli impianti acquedottistici, con modalità ed in quantità differenti, a seconda della destinazione di utilizzo. Difatti, possono distinguersi utilizzi quali civile, potabile e non potabile, agricolo e produttivo.

In tale contesto, la distribuzione della risorsa per uso domestico presenta le maggiori problematiche, come la necessità di assicurare con continuità il servizio e la qualità dell'acqua, adattandosi alle richieste fluttuanti delle utenze. Inoltre, molto importante è il controllo dei volumi persi a seguito di guasti che possono verificarsi in ogni componente degli impianti.

Questo studio, nato dalla stretta collaborazione tra il *Politecnico di Torino* e la società *SMAT* (*Società Metropolitana Acque Torino S.p.a.*), ha la finalità di esaminare la problematica relativa alle rotture nel sistema acquedottistico di Torino, una volta definito un periodo di osservazione dei guasti tra gli anni 2006 e 2016. In questo intervallo temporale saranno analizzate le occorrenze delle rotture e le eventuali relazioni con le caratteristiche delle tubazioni della rete acquedottistica. Di primaria importanza sarà l'analisi delle proprietà che, in modo predominante, influenzano la tendenza alla rottura delle condotte e la valutazione della relazione che intercorre tra tali variabili.

La tesi si pone a valle del lavoro svolto in tale ambito dalla collega e *Dottoressa Clara Ghigo*, concluso nel dicembre 2017, e si prefigge di individuare le classi di condotte più vulnerabili alle rotture e la relazione tra i guasti e le differenti proprietà delle condotte. Una corretta previsione delle rotture, infatti, consente un'efficace programmazione degli interventi di manutenzione e rinnovo delle parti danneggiate, prima che avvenga il guasto e la perdita di volume idrico, con conseguente risparmio in termini idrici ed economici.

Tale ricerca si avvarrà di due modelli di regressione statistica: il modello di regressione logistica ed il modello di regressione polinomiale. In particolare, il primo assume una grande importanza in questo lavoro, poiché rappresenta un metodo innovativo, mai applicato nell'ambito della problematica relativa alle rotture delle condotte.

Si riporta di seguito una breve descrizione dei capitoli che comporranno la tesi.

Nel **Capitolo 1** sarà presentata la tematica relativa al problema delle rotture nelle condotte acquedottistiche e i vantaggi derivanti dallo studio e dalla definizione di classi vulnerabili. Inoltre, saranno esposti diversi approcci statistici al problema, riportati in letteratura, e un approccio innovativo che consentirà la valutazione della probabilità

delle condotte di incorrere in un guasto. Una volta scelti i modelli da applicare, questi saranno descritti nei capitoli successivi.

Nel **Capitolo 2** sarà esposto il modello di regressione logistica e le modalità di applicazione e di valutazione della sua bontà di adattamento al campione di dati.

Allo stesso modo, nel **Capitolo 3**, sarà esposto il secondo modello scelto, quello di regressione polinomiale, tratto dalla pubblicazione di *Berardi & al.* del 2008.

Nel **Capitolo 4** sarà analizzata la composizione della rete idrica di Torino e saranno definite le occorrenze delle diverse fasce di lunghezze, diametri, materiali e anni di posa della totalità delle condotte.

Il **Capitolo 5** conterrà una descrizione delle modalità di raccolta dei dati da parte di *SMAT* per la gestione degli interventi sui 292 comuni coordinati. Si descriveranno, inoltre, le tre tabelle fondamentali, contenenti le informazioni sui guasti e, nello specifico, gli interventi di tipo *'fuga condotta*' avvenuti nel comune di Torino tra il 2006 ed il 2016. Dall'unione di queste tabelle e a seguito di filtraggio e integrazione di informazioni (quali l'anno di posa ed il carico massimo) si otterrà la tabella *Wwvcondottetorino*. L'intersezione con il database cartografico (contenente le informazioni riguardo la totalità delle condotte della rete) consentirà di ottenere dapprima la tabella *ReteTorino* e, a seguito di importanti filtrazioni, la tabella *Esatta.* Quest'ultima sarà il punto di partenza per la stima dei modelli di regressione e la successiva fase di *testing* degli stessi.

Il **Capitolo 6** conterrà l'analisi dettagliata delle tabelle appena ricavate e la conseguente valutazione delle classi di condotte più vulnerabili. In alcuni casi, le informazioni fornite dalle tre tabelle ottenute nel **Capitolo 5** non saranno coerenti tra di loro: questo a causa della filtrazione adottata per le stesse e la relativa esclusione di elementi non idonei all'applicazione dei modelli di regressione statistica.

Nel **Capitolo 7** verranno riportati i risultati della regressione logistica. A partire dalla tabella *Esatta* si ricaveranno delle sotto-estrazioni, al fine di valutare la significatività dei singoli parametri. Tali sotto-estrazioni saranno costruite in modo tale da mantenere inalterata la percentuale di rotture della tabella *ReteTorino*. Successivamente, da ogni sotto-estrazione saranno prelevate due ulteriori estrazioni, utili per stimare i parametri di regressione e la capacità del modello di poter prevedere una rottura.

Tra i due modelli analizzati, quello di regressione logistica si presenterà come quello più idoneo per l'applicazione a questo caso studio, poiché sarà capace di stimare delle probabilità di rottura maggiori per condotte effettivamente rotte, comparate con le probabilità ottenute per le tubazioni integre.

Allo stesso modo, nel **Capitolo 8** saranno riportati i risultati dell'applicazione del modello di regressione polinomiale. In accordo con quanto riportato nel **Capitolo 3**, le condotte saranno suddivise in classi di materiali e, successivamente, di diametri ed età equiprobabili, sulle quali saranno stimati differenti modelli polinomiali. Come sarà possibile notare, tale modello mal si presta a rappresentare la relazione tra il numero di rotture e le variabili indipendenti. Infatti, i modelli stimati saranno contraddistinti da una

scarsa capacità di adattamento alle diverse classi, a causa di un numero troppo limitato di dati completi.

Capitolo 1

Le rotture nelle condotte: presentazione e approcci al problema

Nel seguente capitolo sarà descritta la problematica relativa ai guasti delle condotte acquedottistiche e i diversi approcci proposti dalla letteratura per modellare le variabili che influiscono maggiormente sul numero di rotture.
Saranno successivamente scelti due modelli tra quelli esposti e nei successivi capitoli saranno applicati al caso studio del comune di Torino.

1.1 Le rotture delle condotte nei sistemi acquedottistici

I guasti delle condotte nelle reti acquedottistiche hanno luogo a causa di una grande varietà di fattori che possono influire positivamente o negativamente nei riguardi di tale problematica. Ogni sistema mostra negli anni delle condizioni di deterioramento che impattano sull'efficienza delle condotte, fino alla rottura delle stesse.
Un primo importante fattore è l'età della condotta e dei suoi componenti che, in concomitanza con la scarsa disponibilità di fondi da destinare agli interventi di manutenzione ordinaria, porta velocemente a rotture e conseguenti perdite di volumi d'acqua. Inoltre, le cause possono essere ricercate nelle diverse caratteristiche fisiche e geometriche delle condotte come il materiale, il diametro, il carico in esercizio e la lunghezza.
In questo contesto si inserisce la necessità di ricercare la correlazione tra le diverse variabili in gioco, per poter prevedere quali siano le classi di condotte più vulnerabili e, conseguentemente, programmare con largo anticipo operazioni di manutenzione e sostituzione di queste parti della rete, prima che il guasto sopraggiunga.
A tale proposito sono di norma utilizzati modelli statistici *non-fisicamente basati*. Nel prossimo paragrafo verranno brevemente descritti alcuni tra questi modelli individuati in letteratura, rimandando, però, la trattazione approfondita ai capitoli successivi.
Per definire quello che è lo stato della rete, in genere, si fa riferimento al numero di rotture per chilometro di condotta (definito più comunemente *tasso di fallanza*) o, nello stesso modo, al numero di rotture annuali per chilometro di condotta.

1.2 Gli approcci al problema delle rotture nelle reti acquedottistiche

Come precedentemente esposto, i modelli statistici utilizzati per studiare la relazione tra le variabili che caratterizzano le condotte e il numero di rotture possono essere suddivisi in *fisicamente basati* e *non-fisicamente basati.* I primi sono degli approcci che consentono la modellazione su piccola scala del problema e necessitano di numerose e costose misurazioni di parametri fisici che caratterizzano il fenomeno. Vengono utilizzati maggiormente per condotte che trasportano greggio e gas.
I secondi, invece, sono modelli ampiamente utilizzati per studiare l'insorgere di rotture per le reti di approvvigionamento idrico. Se ne riportano di seguito alcuni, tratti dalla letteratura e descritti in maniera sintetica: in particolare, saranno descritti i dati necessari per la stima del modello, l'output e la formulazione dello stesso.

Il modello di *Mailhot* et al. (2000)

Il modello si propone di esaminare il numero di rotture che possono avere luogo per una determinata condotta o una rete di condotte in un intervallo definito di tempo.
Si riporta in **Figura 1.1** lo schema relativo a tale modello: si possono distinguere tre rotture osservate nel periodo che va dall'anno di installazione della tubazione all'inizio del periodo di osservazione T_b e altre tre rotture nel periodo di osservazione (tra T_a e T_b).
Tenendo conto che prima del periodo di osservazione possano esserci informazioni parziali e non complete relative al numero (α) e al tempo di arrivo (t'_i) tra le rotture, la parametrizzazione avviene sulle informazioni disponibili nel periodo di osservazione, cioè il numero di rotture β e il tempo che intercorre tra le rotture, definito come t_i. In particolare, il tempo di arrivo delle rotture e quello di inizio e di fine del periodo di osservazione sono calcolati a partire dall'anno di installazione delle condotte.

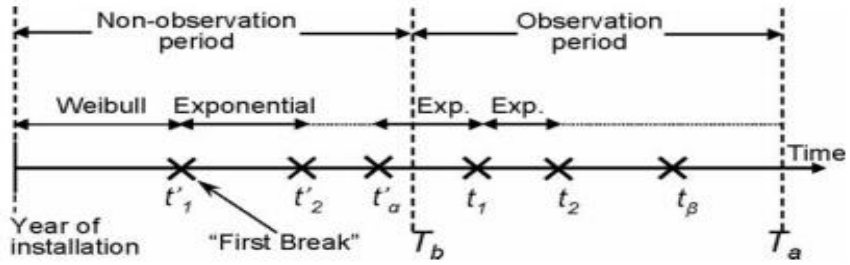


Figura 1.1. Schema temporale e variabili utilizzate nel modello di *Mahilot et al.*

Il modello fa uso di due differenti distribuzioni di probabilità per modellare il tempo che intercorre tra le rotture: la distribuzione a due parametri (γ e κ) di *Weibull* per modellare il tempo di arrivo della prima rottura per una condotta e la distribuzione *esponenziale* a un parametro (λ) per definire il tempo intercorso tra le successive rotture. La stima di questi parametri avviene massimizzando la funzione di massima verosimiglianza per

ogni condotta, tenendo in conto di quelle mai rotte e di quelle che, nel periodo di osservazione, hanno osservato β rotture. Per ogni condotta, i dati necessari per tale stima sono le rotture osservate e il tempo intercorso tra queste.
Una volta stimati i parametri, il numero di rotture previsto m_p per una determinata condotta *p-esima* nel generico intervallo di tempo [T_1, T_2] ammonta a:

$$m_p(T_1, T_2) = [F_1(T_1) - F_1(T_2)\,] + \lambda \left\{ T_2\big[1 - F_1(T_2) - T_1[1 - F_1(T_1)]\big] - \int_{T_1}^{T_2} t \cdot f_1(t)dt \right\} (1)$$

dove $F_1(T_i)$ rappresenta la funzione di sopravvivenza relativa al tempo T_i *e* al modello di *Weibull* e $f_1(t)$ la corrispondente funzione densità di probabilità.
In definitiva, per una rete composta da *p* condotte, il numero di rotture previsto (*m*) nel periodo di tempo [T_1, T_2] è:

$$m(T_1, T_2) = \sum_{p=1}^{n_p} m_p(T_1, T_2) \qquad (2)$$

Il modello di regressione multipla (*Wang_2006*)

Il modello ha lo scopo di determinare il logaritmo del tasso di rottura per le condotte facenti parte di reti acquedottistiche, espresso come il logaritmo del numero di rotture annuali per chilometro di condotta.
L'intera rete è suddivisa in classi di materiale e, per ogni classe, a partire dal campione di osservazioni si definisce per ogni condotta il tasso di fallanza annuale, che si colloca nel modello come variabile dipendente. I parametri quali diametro, anno di posa, lunghezza e carico massimo definiscono invece le variabili indipendenti.
Una volta definito il campione di partenza, per ogni classe vengono generate diverse espressioni polinomiali che possano adattarsi ai dati e si stimano i parametri incogniti della regressione attraverso dei software di modellazione statistica quali *Minitab* e *Gretl*.
Tra le diverse forme, si sceglie quella che meglio si adatta al campione e che è caratterizzata da una giustificata complessità di espressione.
La significatività totale del modello e la bontà di adattamento sono valutati rispettivamente tramite *F-test* e il *coefficiente di determinazione*, mentre la significatività dei singoli parametri attraverso il *t-test*.
Questo modello, a differenza del precedente, non è in grado di predire quanto una rottura avrà luogo, ma solo il numero annuale di rotture stimate per chilometro di condotta. Inoltre, non tiene in conto di precedenti rotture e non effettua una distinzione tra prima rottura osservata e successive.
Infine, il dato riguardante il tasso di fallanza può essere convertito in un numero equivalente di rotture attraverso un'analisi economica e la conseguente programmazione dei fondi da investire.

Il modello logistico

La regressione logistica è un metodo innovativo che permette di studiare la relazione che lega le variabili indipendenti quali materiale, diametro, anno di posa, carico e lunghezza delle condotte con la probabilità di rottura delle stesse. Tale metodo trova la sua prima applicazione in questo lavoro, e si differenzia in modo fondamentale dagli altri metodi. Infatti, i modelli fin qui esposti restituiscono come variabile di interesse il numero di rotture o il tasso di fallanza. Il modello di regressione logistica, invece, necessita che la variabile dipendente di input sia di tipo di tipo dicotomico, cioè capace di assumere i soli valori 0 e 1. Questi valori definiscono, rispettivamente, la non rottura e la rottura di ogni condotta e, assieme alle variabili indipendenti, compongono il campione di dati sul quale tarare le variabili del modello.
Il modello di regressione semplice si presenta nella seguente forma:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} + \varepsilon \tag{3}$$

dove:

- β_0 e β_1 sono i parametri da stimare del modello di regressione logistica;
- ε rappresenta il termine di errore relativo alla previsione della variabile *Y;*
- $\pi(x)$ è la probabilità che un dato soggetto sia caratterizzato dalla presenza di un attributo (*Y*=1)

Nel caso di "*i*" variabili indipendenti, il modello assume la forma:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x + \cdots + \beta_i x_i)}{1 + \exp(\beta_0 + \beta_1 x + \cdots + \beta_i x_i)} + \varepsilon \tag{4}$$

Inoltre, di fondamentale importanza per la definizione del modello e l'interpretazione dei dati è il ruolo degli *Odds, Odds Ratio* e della funzione *logit*.
I parametri del modello sono stimati massimizzando la funzione di massima verosimiglianza e, una volta ottenuti i loro valori, la probabilità stimata che una condotta con determinate caratteristiche vada incontro a rottura è espressa come:

$$\hat{\pi}(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{5}$$

La bontà di adattamento del modello è stimata attraverso il *test rapporto di verosimiglianza* e lo *pseudo* R^2. La significatività dei parametri, invece, è definita dal *test Z* e dalla *statistica di Wald*.

Il modello polinomiale (*Berardi_2008*)

Si riporta il modello *EPR* (*evolutionary polynomial regression)* descritto da *Berardi*, *Giustolisi*, *Zapelan* e *Savic* in una pubblicazione del 2008.
In modo analogo alla regressione multipla, lo scopo della modellazione statistica è quello di determinare un numero stimato di rotture per ogni condotta facente parte della rete acquedottistica in esame.
Il modello polinomiale adottato per definire la dipendenza tra il numero di rotture per una determinata classe di condotte e le caratteristiche delle stesse è contraddistinto dalla seguente formulazione:

$$Y = a_0 + \sum_{j=1}^{m} a_j X_1^{ES(j,1)} \cdot \ldots \cdot (X_k)^{ES(j,k)} \cdot f\left(X_1^{ES(j,k+1)} \cdot \ldots \cdot (X_k)^{ES(j,2k)}\right) \quad (6)$$

dove:

- *Y* è la variabile indipendente, denominata anche *BR,* ed indica il numero di rotture;
- X_k è la generica variabile esplicativa (ad esempio diametro, materiale, anno di posa e lunghezza);
- *ES* è la matrice degli esponenti incogniti;
- a_j è il generico coefficiente polinomiale da stimare;
- *m* è il numero di termini polinomiali addizionali alla costante a_0;
- *f* rappresenta una funzione arbitraria che può incrementare la capacità di previsione delle rotture.

La procedura si compone di due momenti fondamentali:

- in prima istanza si effettua una stratificazione della totalità delle condotte, dapprima secondo il materiale e successivamente secondo classi di diametri ed età rappresentative (**Equazioni 7 e 8**). Gli altri dati di input, una volta definito il numero di rotture della classe come variabile dipendente, sono il diametro, l'età e la lunghezza delle classi. Ogni classe deve essere composta da un numero comparabile di elementi ed è per questo suggerito lo studio preventivo della distribuzione di probabilità cumulata di diametri ed età delle condotte, per poter effettuare una suddivisione coerente in classi;

$$A_{Classe} = \frac{\sum_{Classe}(L_p \cdot A_p)}{L_{Classe}} \quad (7)$$

$$D_{Classe} = \frac{\sum_{classe}(L_p \cdot D_p)}{L_{Classe}} \qquad (8)$$

- successivamente, per ogni materiale, si ricerca un certo numero di espressioni in forma polinomiale, progressivamente più complesse e caratterizzate da più variabili indipendenti ed esponenti incogniti. Tali espressioni devono descrivere la relazione tra il numero di rotture di ogni classe e i parametri indipendenti scelti. I parametri della regressione sono stimati attraverso il software *Gretl*, così come la bontà di adattamento di ogni modello polinomiale, espressa dal *coefficiente di determinazione.* Tale coefficiente permetterà, in un momento successivo, di posizionare tutte le espressioni su una *frontiera di Pareto* per definire quella che meglio si adatta al campione di classi di ogni materiale.

Alla base di questo metodo vi è la suddivisione in classi di condotte per ricavare diametri ed età equivalenti, pesati rispetto alla lunghezza totale di ogni classe. Questo risulta essere uno dei vantaggi dell'utilizzo del modello di regressione polinomiale: infatti, il dato riguardante la lunghezza porta con sé ulteriori informazioni spesso non disponibili quali la variabilità dei carichi stradali e i carichi del terreno.
D'altra parte, anche questo modello non permette di definire il momento in cui una condotta incorrerà in un guasto, ma solo il numero di rotture per una determinata classe, in un intervallo di tempo pari a quello di osservazione.

1.3 Conclusioni

In questo capitolo sono stati descritti quattro approcci al problema delle rotture: il modello di *Mailhot et al._2000*, il modello di regressione multipla (*Wang_2006*), il modello logistico ed il modello polinomiale (*Berardi et al._2008*).
Ognuno di questi è contraddistinto da variabili dipendenti differenti come il numero di rotture, il tasso di fallanza e la probabilità di rottura per una determinata condotta.
Alla luce dei dati a disposizione nel caso studio *SMAT*, è stato scelto di approfondire ed applicare il modello logistico e quello polinomiale. Il primo sarà analizzato ed applicato rispettivamente nel **Capitolo 2** e nel **Capitolo 7**, mentre il secondo sarà descritto nel **Capitolo 3** ed applicato nel **Capitolo 8**.

Capitolo 2

La regressione logistica

L'analisi di regressione logistica è una metodologia impiegata per esaminare la relazione causale che lega una variabile dipendente dicotomica a una o più variabili indipendenti esplicative di tipo quantitativo e qualitativo. Nello specifico, la variabile dipendente dicotomica è codificata come 0-1 e descrive l'esito di un evento aleatorio come, ad esempio, l'insorgere di una rottura di un tratto di condotta in un determinato periodo di osservazione.
Nel caso studio in esame, la variabile dipendente *Y* assume valori pari 0 quando il tratto non presenta una rottura (insuccesso dell'evento) e valori pari a 1 nel caso di presenza dell'attributo di rottura (successo dell'evento) ed è distribuita secondo la distribuzione binomiale.

2.1. Il modello di regressione logistica

Lo studio del modello di regressione logistica ha la finalità di stimare l'influenza di molteplici parametri che differenziano le condotte (ad esempio il diametro, il materiale, l'anno di posa e la pressione) sulla tendenza della condotta stessa alla rottura.
Si riporta di seguito un esempio generale per chiarire quanto riportato nelle righe precedenti.

> **Esempio 1.**
> Si supponga di avere a disposizione un campione di *n* soggetti presi in esame e di voler stimare la probabilità che il singolo individuo possa sviluppare un tumore in relazione all'età dello stesso.
> La variabile risposta *Y* è influenzata da un solo regressore (l'età) e la sua variabilità di tipo binario è contraddistinta per ogni soggetto *i-esimo* (con *i=1,2,…,n)* dal valore:
>
> - Y_i=1 se l'individuo ha sviluppato un tumore;
> - Y_i=0 se l'individuo non ha mai sviluppato un tumore.

Il modello ha lo scopo di definire la probabilità π(x) che un dato soggetto sia caratterizzato dalla presenza di un attributo (*Y*=1) per un determinato valore della variabile indipendente X (*X*=x). Nello specifico, valgono le seguenti relazioni:

$$P(Y = 1|\, X = x) = \pi(x) \quad (1)$$

$$P(Y = 0|\, X = x) = 1 - \pi(x) \quad (2)$$

$$E\big(Y(x)\big) = 1 \cdot P(Y = 1|\, X = x) + 0 \cdot P(Y = 0|\, X = x) = \pi(x) \quad (3)$$

Dove E(Y(x)) indica il valore atteso ed è pari a π(x).
Si potrebbe inizialmente pensare di modellare π(x) utilizzando un modello di regressione lineare del tipo:

$$\pi(x) = P(Y = 1|\, X = x) = \alpha + \beta x + \varepsilon \quad (4)$$

dove:

- α è il parametro che rappresenta l'intercetta e rappresenta il valore di *Y* quando la variabile *x* è nulla;
- β è il coefficiente di regressione e rappresenta l'incremento (o decremento) di *Y* per una variazione unitaria della variabile *x;*
- ε rappresenta il termine di errore relativo alla previsione della variabile *Y.*

Questa tipologia di modello risulta però inadatta nel caso di variabile dicotomica poiché:

- la probabilità π(x) è funzione lineare di *x* e questo comporta, per valori molto grandi o molto piccoli di *x*, possibili valori di π(x) esterni all'intervallo [0,1]. Valori esterni a tale intervallo non hanno però significato poiché la variabile dipendente è una probabilità;
- in questo modello l'errore si distribuisce normalmente con media nulla e varianza costante ma tale assunzione non è valida per una variabile dicotomica perché, in tal caso, l'errore si distribuisce con media nulla e varianza non costante dipendente da *x*. È possibile affermare che, nel caso di variabile dicotomica dipendente, l'utilizzo di un modello lineare implica la violazione dell'ipotesi di normalità dell'errore ε.

Per ovviare a queste problematiche, si sceglie di adottare un modello di regressione logistica che rappresenta in maniera più coerente la dipendenza tra la probabilità π(x) e il regressore *x*. Il modello di regressione semplice si presenta nella seguente forma:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} + \varepsilon \quad (5)$$

dove:

- β_0 e β_1 sono i parametri da stimare del modello di regressione logistica;
- ε rappresenta il termine di errore relativo alla previsione della variabile *Y.*

Nel caso di "*i*" variabili indipendenti, il modello assume la forma:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x + \cdots + \beta_i x_i)}{1 + \exp(\beta_0 + \beta_1 x + \cdots + \beta_i x_i)} + \varepsilon \quad (6)$$

Nel caso di regressione monovariata in cui β_0=0 e β_1=1 si ottiene il grafico riportato in **Figura 2.1**. È possibile notare che la relazione (6) rappresenta una funzione che descrive una curva a forma di "*S*" detta sigmoide. Tale curva risulta limitata superiormente e inferiormente dalle rette y=1 e y=0 che rappresentano gli asintoti orizzontali.
I parametri del modello sono calcolati a partire da una popolazione campionaria di dimensioni finite e la loro stima avviene massimizzando la funzione di verosimiglianza, come descritto nel successivo paragrafo.

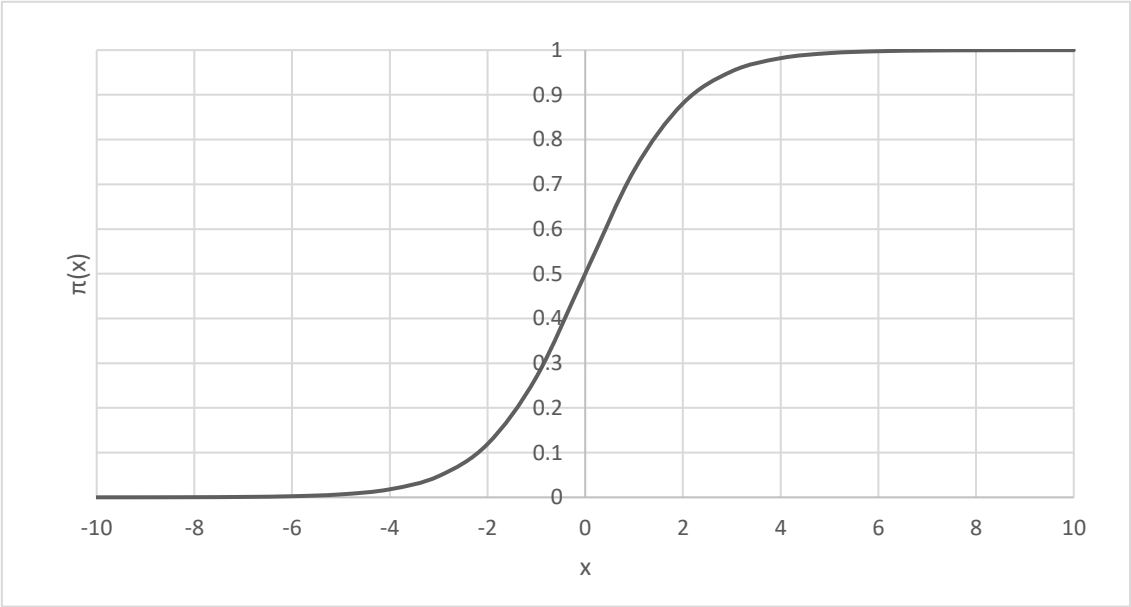


Figura 2.1. Grafico della distribuzione logistica semplice con parametri prefissati

2.2. La stima dei parametri

La stima dei parametri del modello, come accennato in precedenza, è definita a partire da una popolazione campionaria di dimensioni finite. A differenza di quanto accade per la regressione lineare, questa valutazione non può avvenire mediante l'utilizzo del metodo dei minimi quadrati, poiché non vale l'omoschedasticità e gli errori non si distribuiscono con media nulla e varianza costante.
I valori dei coefficienti che meglio adattano le stime del modello al campione si ricavano dall'applicazione dell'algoritmo di massima verosimiglianza che massimizza la funzione di massima verosimiglianza, al variare dei parametri del modello. Nello specifico, la funzione di massima verosimiglianza definisce la probabilità di ottenere il valore atteso di *Y* in funzione dei parametri del modello.

L'algoritmo prevede un processo iterativo, partendo da valori arbitrari dei parametri, e porta a convergenza il processo quando la capacità di miglioramento della stima risulta infinitesima.

Si consideri il modello di regressione logistica semplice descritto dalla relazione (5): la funzione di verosimiglianza della popolazione campionaria $y_1,\dots,y_n$ è descritta dall'equazione (7).

$$L(\beta_0,\beta_1)=\prod_{i=1}^{n} f(y_i,x_i)=\prod_{i=1}^{n}\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \tag{7}$$

La corrispondente funzione di log-verosimiglianza risulta:

$$l(\beta_0,\beta_1)=\sum_{i=1}^{n}\{y_i ln[\pi(x_i)]+(1-y_i)ln[1-\pi(x_i)]\} \tag{8}$$

Una volta annullate le derivate parziali della (8) rispetto ai parametri β_0 e β_1, si ricava il sistema di equazioni che restituisce le stime a convergenza dei parametri, indicate nella (9) e nella (10) con b_0 e b_1:

$$\sum_{i=1}^{n}\left\{y_i-\frac{1}{[1+e^{b_0+b_1x_i}]}e^{b_0+b_1x_i}\right\}=0 \tag{9}$$

$$\sum_{i=1}^{n}\left\{y_i x_i-\frac{1}{[1+e^{b_0+b_1x_i}]}e^{b_0+b_1x_i}x_i\right\}=0 \tag{10}$$

Ricavate le stime dei parametri, la probabilità che la variabile dipendente sia caratterizzata da un determinato attributo (*Y=1*) in relazione alla variabile indipendente è così definita:

$$\hat{\pi}(x)=\frac{\exp(b_0+b_1x)}{1+\exp(b_0+b_1x)} \tag{11}$$

La soluzione del sistema di due equazioni si ricava attraverso metodi iterativi implementati in Software quali *Minitab* e *Gretl*. Per comprendere al meglio la procedura appena descritta, si riporta un esempio tratto dal libro di *Montgomery*, *Peck* e *Vining (2012)*.

Esempio 2.

Sono di seguito illustrati i dati riportati dal giornale *Biometrics* relativi al numero di minatori operanti nel settore minerario che mostrano sintomi di

Pneumoconiosi in relazione al numero di anni di esposizione alle polveri che ne sono la causa (**Tabella 2.1**).

Tabella 2.1 Dati riportati dal *Biometrics* relativi al numero di minatori con sintomi di pneumoconiosi in relazione agli anni di esposizione e relative percentuali

Anni di esposizione	Numero di minatori con sintomi	Numero totale di minatori	Percentuale di malati
5.8	0	98	0
15	1	54	1.9
21.5	3	43	7
27.5	8	48	16.7
33.5	9	51	17.6
39.5	8	38	21.1
46	10	28	35.7
51.5	5	11	45.5

La percentuale di malati per ogni durata di esposizione è calcolata come il rapporto tra il numero di minatori che presentano sintomi e il numero totale di minatori esaminati in corrispondenza di quel dato di esposizione.

Il calcolo dei parametri del modello di regressione logistica avviene, come illustrato in precedenza, attraverso un metodo iterativo ed in questo caso si utilizza il software *Gretl*.

I dati di input del programma sono la variabile dipendente Y=1/Y=0, che rappresenta rispettivamente la presenza e l'assenza di sintomi in un soggetto, e la variabile indipendente rappresentata dagli anni di esposizione. L'output del modello in **Figura 2.2** mostra i parametri stimati attraverso la funzione di massima verosimiglianza e permette di calcolare la probabilità di riscontrare i sintomi di Pneumoconiosi in relazione ai diversi anni di esposizione.

I termini di interesse sono '*const*' e '*Annidiesposizione*' che rappresentano i coefficienti b_o e b_1 dell'equazione (10).

La probabilità stimata di presentare sintomi in relazione agli anni di esposizione è definita come:

$$\hat{\pi}(x) = \frac{\mathrm{e}^{(b_0+b_1x)}}{1+\mathrm{e}^{(b_0+b_1x)}} = \frac{\mathrm{e}^{(-4.69979+0.0916836x)}}{1+\mathrm{e}^{(-4.69979+0.0916836x)}}$$

```
Modello 2: Logit, usando le osservazioni 1-371
Variabile dipendente: Flag
Errori standard basati sull'Hessiana

                    coefficiente  errore std.    z      p-value
  ------------------------------------------------------------------
  const              -4.69979      0.551465    -8.522  1.56e-017 ***
  Annidiesposizione   0.0916836    0.0149215    6.144  8.03e-010 ***

Media var. dipendente  0.123989   SQM var. dipendente    0.330014
R-quadro di McFadden   0.187322   R-quadro corretto      0.172938
Log-verosimiglianza   -113.0032   Criterio di Akaike     230.0063
Criterio di Schwarz    237.8387   Hannan-Quinn           233.1171
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 323 (87.1%)
f(beta'x) nella media delle variabili indipendenti = 0.068
Test del rapporto di verosimiglianza: Chi-quadro(1) = 52.0942 [0.0000]
```

Figura 2.2 Output di *Gretl*

In **Tabella 2.2** sono riportati i valori ottenuti dall'equazione precedente ed il confronto con i valori campionari. A titolo di esempio, si stima la probabilità di presentare sintomi di Pneumoconiosi per un soggetto esposto a 15 anni:

$$\hat{\pi}(x) = \frac{e^{(b_0+b_1x)}}{1+e^{(b_0+b_1x)}} = \frac{e^{(-4.69979+0.0916836\cdot 22)}}{1+e^{(-4.69979+0.0916836\cdot 22)}} = 5.6\%$$

Quanto ottenuto equivale a dire che 5.6 individui su 100 esposti per 22 anni alle polveri presentano i sintomi della malattia.
Tutti i risultati sono illustrati graficamente in **Figura 2.3**: è possibile notare che la curva del modello asseconda in modo adeguato alcuni punti del campione mentre si discosta in maniera marcata in corrispondenza di altri valori di *X*.

Tabella 2.2 Confronto tra la probabilità calcolata nel campione e la probabilità stimata

Anni di esposizione	Probabilità campionaria di presentare sintomi (%)	Probabilità stimata di presentare sintomi (%)
5.8	0	1.4
15	1.9	3.2
21.5	7	5.8
27.5	16.7	9.7
33.5	17.6	15.9
39.5	21.1	24.9
46	35.7	37.8
51.5	45.5	50.4

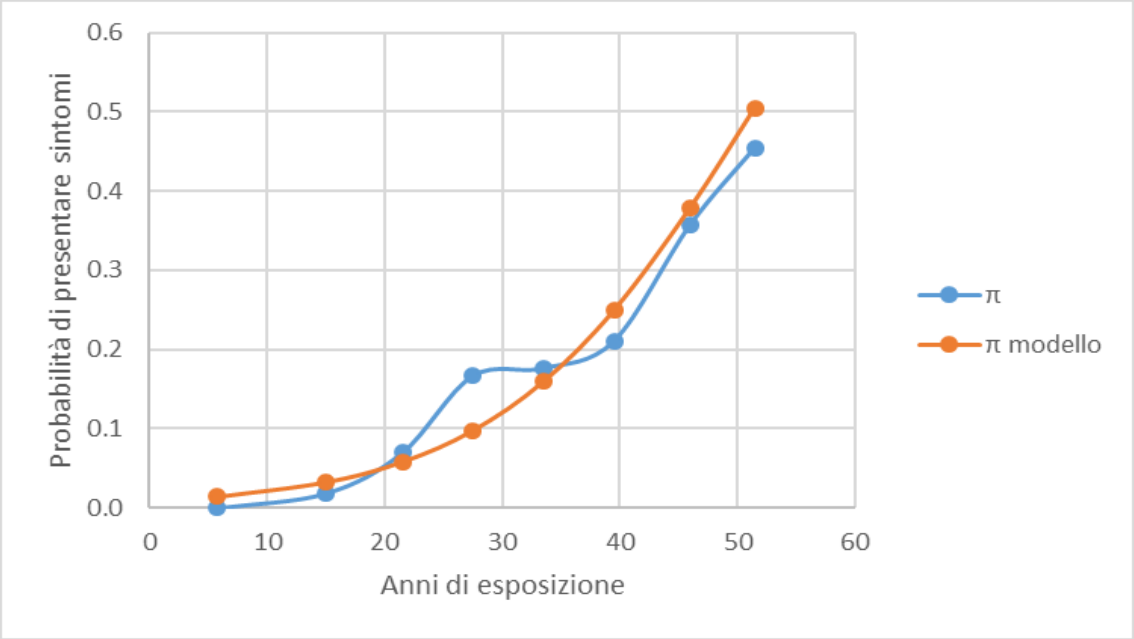


Figura 2.3. Confronto grafico tra la probabilità campionaria e la probabilità stimata del modello

2.3. Il ruolo di Odds, Odds Ratio e funzione Logit

Nel modello di regressione logistica ricoprono un ruolo fondamentale i rapporti Odds, Odds Ratio e la funzione Logit.
L'Odds è definito dal rapporto tra la frequenza di appartenenza ad una categoria e la frequenza di non appartenenza alla stessa, riportate rispettivamente come π(x) e 1-π(x). Fornisce indicazioni riguardo la probabilità di successo ma non è una probabilità: infatti, la probabilità è il rapporto tra la frequenza di appartenenza ad una categoria e la frequenza di appartenenza a tutte le categorie.
L'*Odds* è in relazione con la probabilità come segue:

$$Odds = \frac{\pi(x)}{1 - \pi(x)} \tag{12}$$

In particolare:

- valori di Odds pari a 1 indicano che il successo e l'insuccesso sono equiprobabili;
- valori positivi minori di 1 indicano che il successo è meno probabile dell'insuccesso;
- valori maggiori di 1 indicano che il successo è più probabile dell'insuccesso.

Il campo di variazione appartiene quindi al range [0; +∞]. Si riporta di seguito un esempio per chiarire la definizione di Odds ed il suo significato.

Esempio 3.
Un campione è composto da 42 soggetti così suddivisi: 30 uomini e 12 donne.
La probabilità che il soggetto sia uomo è pari a:

$$P(uomo) = \frac{30}{42} = 0.714$$

Quella di esser donna:

$$P(donna) = \frac{12}{42} = 0.286$$

valore complementare a 1 di 0.714. Si applica l'equazione (12) per calcolare l'Odds relativo agli uomini:

$$Odds(uomo) = \frac{\frac{30}{42}}{\frac{12}{42}} = \frac{0.714}{0.286} = 2.5$$

Il valore ottenuto suggerisce che per ogni donna nel campione ci sono 2.5 uomini. In modo analogo, se si definisce come successo (Y=1) l'attributo uomo, il risultato attesta che la probabilità di successo è 2.5 volte quella di insuccesso.

L'*Odds Ratio*, o rapporto tra gli Odds, rispecchia l'associazione tra due variabili e permette di definire la relazione tra due categorie in funzione di una terza.
Si indica con π_1 la probabilità che si verifichi un evento E_1 e con π_2 la probabilità che si verifichi l'evento E_2; l'Odds Ratio è pari al rapporto tra gli Odds:

$$Odds\ Ratio = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} \tag{13}$$

Si riporta un esempio per chiarire la definizione di questa variabile.

Esempio 4.
Si supponga che il campione di soggetti riportato nell'*Esempio 3* si distribuisca tra i lavori *Ingegneri* e *Insegnanti* come riportato in **Tabella 2.3**.

Tabella 2.3. Distribuzione della variabile sesso e della variabile lavoro per l'Esempio 4

	Sesso		
Lavoro	*Uomini*	*Donne*	*Totale*
Ingegneri	18	2	20
Insegnanti	12	10	22
Totale	30	12	42

Si adotta la stessa formulazione dell'*Esempio 3* per il calcolo delle probabilità e si ricavano i seguenti dati:

- la probabilità π_1 che un ingegnere sia uomo è pari a π_1=18/20=0.9;
- la probabilità π_2 che un insegnante sia uomo è pari a π_2=12/20=0.54;
- le probabilità complementari a 1 (o probabilità di non essere uomo) di π_1 e π_2 sono rispettivamente 0.1 e 0.45;
- l'*odds(uomo)* per gli ingegneri è pari a 9;
- l'*odds(uomo)* per gli insegnanti è pari a 1.2.

L'Odds Ratio è dato dal rapporto tra gli Odds:

$$Odds\ Ratio = \frac{9}{1.2} = 7.5$$

Ciò significa che la popolazione maschile tra gli ingegneri è 7.5 volte più numerosa di quella tra gli insegnanti. Valori dell'indice che si allontanano da 1 indicano una sempre più forte associazione tra le variabili, in questo caso tra sesso e lavoro.

La funzione logit, infine, è definita come il logaritmo naturale dell'Odds secondo la formulazione:

$$logit = ln\frac{\pi(x)}{1-\pi(x)} \tag{14}$$

Dove:

- π(x) è la probabilità di successo;
- 1-π(x) è la probabilità di insuccesso.

Il suo intervallo di variazione, in relazione alla (14), è descritto in **Figura 2.4**: è possibile notare che nel dominio di variazione di π(x), tra 0 e 1, la funzione logit assume valori che comprendono tutto l'insieme dei numeri reali. In modo analogo, all'aumentare del valore assoluto del logit, la probabilità di successo si avvicina ai valori estremi dell'intervallo senza mai raggiungerli.
Un'altra importante proprietà, in aggiunta alla simmetria del dominio della funzione, è la relazione di linearità che la lega alle variabili indipendenti esplicative. Nel caso di una singola variabile indipendente *X* vale:

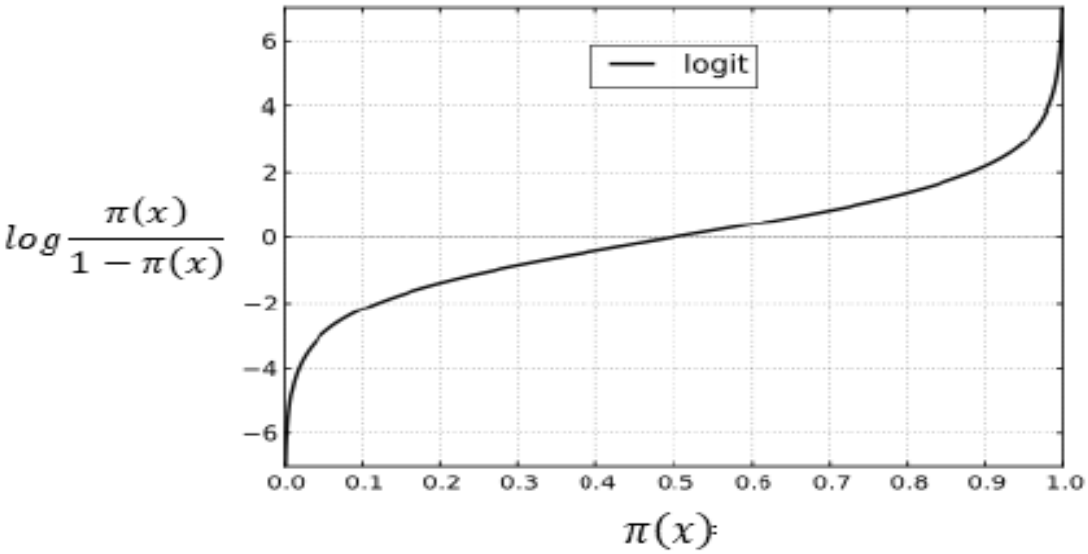
$$logit = ln\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x \tag{15}$$

La relazione (14) definisce la possibilità di applicare un modello lineare alla funzione che lega il logit ai regressori, con tutte le semplificazioni che tale schema comporta.
Inoltre, per le proprietà dei logaritmi:

$$\pi(x) = \frac{e^{logit}}{1 + e^{logit}} = \frac{\frac{\pi}{1-\pi(x)}}{1 + \frac{\pi(x)}{1-\pi(x)}} \tag{16}$$

Infine, sostituendo l'equazione (14) nell'equazione (16) si ricava l'equazione (5).

Figura 2.4. La funzione logit nel caso di unica variabile esplicativa



Nel caso in cui il modello presenti *i* variabili esplicative, la funzione logit assume la seguente forma:

$$logit = ln\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x + \cdots + \beta_i x_i \tag{17}$$

Per comprendere al meglio le differenze tra probabilità, odds e logit e i relativi range di variazione si riporta di seguito un esempio qualitativo.

Esempio 5.

Sono stati osservati 7 set di campioni (0-6), ognuno contraddistinto da 10 misurazioni (misurazione 1-10) di successo (Y=1) o fallimento (Y=0).

I set riportati in **Tabella 2.4** presentano differenti probabilità di successo e fallimento e, di conseguenza differenti valori di odds e logit.

Tabella 2.4. Dati del campione

	Set di campioni						
	0	1	2	3	4	5	6
Mis. 1	1	1	1	1	1	1	1
Mis. 2	1	1	1	1	1	1	1
Mis. 3	0	1	0	1	1	1	1
Mis. 4	0	0	0	1	1	1	1
Mis. 5	0	0	0	1	1	1	1
Mis. 6	0	0	0	0	0	1	1
Mis. 7	0	0	0	0	0	0	1
Mis. 8	0	0	0	0	0	0	0
Mis. 9	0	0	0	0	0	0	0
Mis. 10	0	0	0	0	0	0	0

Una volta condensati i dati in maniera compatta, si calcolano gli odds e i logit di ogni set di misurazioni, secondo le equazioni (12) e (13) e si riportano i risultati in **Tabella 2.5**.

Tabella 2.5. Percentuale di successi, insuccessi, odds e logit dei diversi set

Set	0	1	2	3	4	5	6
Y=1	2	3	2	5	5	6	7
% RS	0.2	0.3	0.2	0.5	0.5	0.6	0.7
Y=0	8	7	8	5	5	4	3
% RF	0.8	0.7	0.8	0.5	0.5	0.4	0.3
Odds	0.25	0.43	0.25	1	1	1.5	2.33
Logit	-1.39	-0.85	-1.39	0	0	0.41	0.85

Nello specifico, per ogni set è stata calcolata la percentuale di successi (% RS), quella di fallimenti (% RF), l'odds del successo e il logit del rispettivo odds.
I corrispondenti valori variano:

- tra 0 e 1 per le percentuali di successo e fallimento;
- tra 0 e più infinito per gli odds;
- su tutto l'insieme R per i logit.

Sono di fondamentale importanza le seguenti osservazioni:

- quando il numero di fallimenti (Y=0) supera il numero di successi (Y=1), l'odds assume valore positivo minore di 1 ed il rispettivo logit valore negativo;
- quando il numero di fallimenti è pari al numero di successi, l'odds assume valore unitario ed il logit valore nullo;

- quando il numero di successi è superiore al numero di fallimenti, l'odds assume valori compresi tra 1 e più infinito ed il logit valori positivi.

2.4. Valutazione della bontà del modello e significatività dei parametri

Dopo aver stimato un modello che possa rappresentare la relazione che lega la variabile indipendente ai suoi regressori, è di fondamentale importanza valutare la sua efficacia di adattamento ai dati campionari. Di analogo interesse è la valutazione del contributo specifico che ogni variabile indipendente apporta sulla stima della variabile dipendente, al fine di valutare il grado di informazione che le stesse introducono nelle stime del modello.
Di seguito verranno elencati e descritti i test di valutazione della bontà di adattamento del modello e i test di significatività di ogni singolo parametro. Fanno parte del primo gruppo di test:

- il test rapporto di verosimiglianza;
- lo pseudo-R^2;
- devianza, Chi-quadro di Pearson e test di Hosomer-Lemeshow.

La stima della significatività di ogni singolo parametro ha luogo, tra gli altri, mediante:

- il test Z;
- la statistica di Wald.

2.4.1 Il test rapporto di verosimiglianza

Si definisce la variabile *G* come il rapporto tra due verosimiglianze:

$$G = -2log\frac{L(0)}{L(\beta)} \tag{18}$$

Nello specifico:

- *L(0)* è definita come la funzione di massima verosimiglianza in corrispondenza del modello caratterizzato dalla sola intercetta;
- *L(β)* è la funzione di massima verosimiglianza in corrispondenza del modello completo.

La statistica *G* segue una distribuzione del χ^2 con un numero di gradi di libertà pari alla differenza tra il numero di parametri del modello completo e il numero di parametri del modello ridotto.

Attraverso questo test si verifica che le variabili del modello aggiungano molte informazioni rispetto al modello caratterizzato dalla sola intercetta: infatti, se questo avviene, la quantità *L(β)* risulta essere più grande di *L(0)* e di conseguenza il rapporto tra le verosimiglianze raggiunge valori molto piccoli prossimi allo 0.
Si sottopongono a verifica le ipotesi H_0 e H_1 che implicano:

$$H_0: \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_1: almeno\ un\ parametro \beta_j \neq 0, j = 1, \dots, p$$

L'ipotesi H_0, definita ipotesi nulla, implica che nessuna delle *p* variabili indipendenti apporti un contributo significativo alle stime del modello, mentre l'ipotesi H_1 afferma che almeno una variabile indipendente sia diversa da zero e che quindi sia significativa.
L'ipotesi nulla può essere rifiutata se sussistono due condizioni:

- il valore di *G* è maggiore del valore tabellato del $\chi 2_{\frac{\alpha}{2}}$, funzione di α e del numero di gradi di libertà.

$$G > \chi^2_{\frac{\alpha}{2}} \tag{19}$$

  dove α è la significatività scelta per il test.

- il corrispondente valore del *p-value* è inferiore al livello di significatività scelto.

Se le condizioni risultano valide, è possibile rifiutare l'ipotesi nulla e questo implica che la potenza predittiva del modello viene migliorata dalla presenza delle variabili indipendenti.
A titolo di esempio, si faccia riferimento all'output riportato in **Figura 2.2**: il valore della statistica *G (Test del rapporto di verosimiglianza)* è pari a 52.0942, maggiore del valore del $\chi 2_{\frac{\alpha}{2}}$ con un grado di libertà e significatività pari a 0.05.

$$G > \chi^2_{\frac{\alpha}{2}} \rightarrow 52.0942 > 5.02$$

In aggiunta, nella stessa figura sono riportati i valori del *p-value* corrispondenti all'intercetta e al parametro del modello: è possibile notare che i numeri restituiti nell'ultima colonna sono sensibilmente inferiori al livello di significatività scelto e si rifiuta quindi l'ipotesi nulla.

2.4.2 Lo pseudo-R^2

In modo analogo, è possibile stimare la bontà di adattamento del modello ai dati attraverso un'ulteriore coefficiente, denominato *pseudo-R^2*, definito dalla relazione:

$$Pseudo\ R_g^2 = 1 - 2log\left[\frac{L(0)}{L(\beta)}\right]^{\frac{2}{n}} \tag{20}$$

Dove:

- lo $Pseudo\ R_g^2$ è il coefficiente di determinazione per modelli non lineari;
- *L(0)* è definita come la funzione di massima verosimiglianza in corrispondenza del modello caratterizzato dalla sola intercetta;
- *L(β)* è la funzione di massima verosimiglianza in corrispondenza del modello completo.
- *n* individua il numero totale di osservazioni.

Il range di variazione del coefficiente di determinazione appartiene all'insieme dei numeri tra 0 e 1. Nello specifico, valori prossimi allo 0 indicano che il modello completo non aggiunge informazioni rispetto al modello con la sola intercetta, mentre valori prossimi all'unità implicano un ottimale adattamento del modello ai dati.
Si può fare riferimento al coefficiente di determinazione scalato $\bar{R}_g^2$, secondo la formulazione:

$$\bar{R}_g^2 = \frac{R_g^2}{R_{g,max}^2} \tag{21}$$

dove il coefficiente di determinazione è scalato rispetto al valore massimo $R_{g,max}^2$ che è definito come:

$$R_{g,max}^2 = 1 - [L(0)]^{\frac{2}{n}} \tag{22}$$

È importante sottolineare che questa statistica non tiene in conto del numero di gradi di libertà del modello e non è sottoposta a test di significatività.

2.4.3 Devianza, Chi-quadro di Pearson e test di Hosomer-Lemeshow

Si illustrano di seguito ulteriori test di bontà di adattamento forniti in aggiunta dal Software *Minitab* (del quale si riporta l'output relativo all'**Esempio 2** in **Tabella 2.6**) che consentono di valutare la capacità del modello di descrivere la popolazione campionaria. Si definisce la *devianza* come una statistica che segue una distribuzione del χ^2 con *n-p* gradi di libertà, una volta definite *n* osservazioni e *p* parametri del modello. Affinché il modello fornisca un'adeguata stima dei dati, il valore di devianza deve essere contenuto e, nello specifico, il rapporto tra quest'ultima e il numero di gradi di libertà non deve superare l'unità.

Tabella 2.6. Output del programma *Minitab* relativo all'Esempio 2

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	6	6,05	0,418
Pearson	6	5,03	0,540
Hosmer-Lemeshow	5	5,00	0,415

Nel caso studio riportato nell'**Esempio 2** i gradi di libertà sono 6, determinati dalla differenza tra le 8 osservazioni e i 2 parametri. La devianza assume valore pari a 6.05 e, una volta divisa per il numero di gradi di libertà, restituisce un valore prossimo all'unità. Si definisce successivamente il *Chi-quadro di Pearson*, caratterizzato da una distribuzione del χ^2 con *n-p* gradi di libertà. Affinché il modello sia adatto a descrivere i dati, devono valere le stesse considerazioni fatte per la statistica devianza. Si riprenda l'**Esempio 2** caratterizzato da 6 gradi di libertà e un valore del *Chi-quadro di Pearson* pari a 5.03: se si divide il valore della statistica per il numero di gradi di libertà, si ottiene un numero prossimo all'unità.
Si riporta infine il test di *Hosomer-Lemeshow* che prevede di classificare i dati in gruppi caratterizzati da prefissate probabilità di successo stimate. Se il campione di dati è sufficientemente grande, la statistica segue una distribuzione del χ^2 con *g-p* gradi di libertà, dove *g* è il numero di classi nelle quali è stato suddiviso il campione. Nel caso dell'**Esempio 2**, una volta suddivisi i dati in 7 gruppi e definiti i 2 parametri del modello, si ricava un valore di *HL* pari a 5. Dividendo tale statistica per i gradi di libertà, si ottiene un valore prossimo all'unità.
Nel complesso, i 3 test confermano che il modello risulta adeguato.

2.4.4 Significatività dei parametri del modello

Una volta verificata la bontà di adattamento del modello e la sua capacità di previsione della variabile dipendente, risulta necessario valutare il contributo di ogni singola variabile indipendente attraverso dei test di significatività. Nel caso in esame si fa riferimento al *test Z* e alla statistica di *Wald*.
Come nella trattazione dei test di bontà di adattamento, si consideri il sistema di ipotesi H_0 e H_1:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

Dove:

- β_j è il *j-esimo* coefficiente di regressione;

- H_0 è l'ipotesi nulla.

Si consideri inoltre:

- b_j la stima del parametro β_j;
- $s(b_j)$ l'errore standard relativo alla stima b_j.

La variabile *test Z* si distribuisce secondo una distribuzione normale standardizzata e risulta:

$$Z = \frac{b_j}{s(b_j)} \tag{23}$$

Affinché sia respinta l'ipotesi nulla H_0 è necessario che la variabile test soddisfi la relazione:

$$|Z| > z_{\frac{\alpha}{2}} \tag{24}$$

La variabile $z_{\alpha/2}$ segue una distribuzione normale standard e assume valori tabellati dipendenti dal livello di significatività scelto.
Se è soddisfatta l'equazione (24), si respinge l'ipotesi nulla, si conclude che il parametro è diverso da 0 e che la variabile esplicativa corrispondente influisce in modo concreto sulla variabile risposta.
La statistica di *Wald* è definita come il quadrato di *Z*, secondo l'equazione (25):

$$W^2 = \left[\frac{b_j}{s(b_j)}\right]^2 \tag{25}$$

e segue una distribuzione del tipo χ^2 ad un grado di libertà.
Al fine di chiarire complessivamente quanto descritto in questo capitolo, si riporta un ulteriore esempio tratto dalla letteratura.

Esempio 6.
Si prendano in esame i dati riportati in **Tabella 2.7** che riportano alcuni verdetti giudiziari riguardo la condanna a morte a seguito di omicidio e si determini la probabilità di ricevere tale condanna in funzione della razza dell'imputato e della vittima.

Tabella 2.7. Distribuzione dei verdetti di pena di morte al variare della natura dell'imputato e della vittima

Razza imputato	Razza vittima	Pena di Morte Sì	Pena di Morte No	% Sì
Bianca	Bianca	53	414	11.3
	Nera	0	16	0.0
Nera	Bianca	11	37	22.9
	Nera	4	139	2.8

Nell'esempio in esame, si definisce la variabile dipendente/risposta *Y* con distribuzione dicotomiale:

- *Y*=1 se è stata assegnata la condanna a morte;
- *Y*=0 se non è stata assegnata la condanna a morte a seguito di omicidio.

Le variabili indipendenti del modello sono 2, la razza dell'imputato e la razza della vittima, e sono identificate attraverso dei valori numerici poiché la loro natura è di tipo qualitativa e non quantitativa.
Nello specifico, i valori assegnati per i due campi sono:

- 1 se la razza dell'imputato/vittima è bianca;
- 0 se la razza dell'imputato/vittima è nera.

La tabella sopra riportata dispone di informazioni riguardo la razza dell'imputato, la razza della vittima, il numero di casi in cui è stata assegnata la condanna a morte, il numero di casi in cui non è stata assegnata e la conseguente percentuale di condannati per le 4 differenti combinazioni.
L'analisi di questi dati porta a delle osservazioni preliminari:

- nel caso di imputato e vittima di razza bianca la condanna a morte è assegnata nell'11.3% dei casi;
- nel caso di imputato di razza bianca e vittima di razza nera, la condanna non è stata assegnata in alcun caso;
- nel caso di imputato di razza nera, la condanna a morte è assegnata nel 22.9% dei casi di vittima bianca e nel 2.8% dei casi di vittima nera.

Data la grande variabilità delle percentuali riscontrate nelle diverse combinazioni, si vuole determinare l'eventuale dipendenza tra la probabilità di essere condannati a morte e la razza dell'imputato e della vittima. A tal fine si costruisce un modello di regressione logistica, descritto dalle variabili:

- *Y*, pena di morte (sì=1/no=0);
- X_1, razza dell'imputato (bianca =1/nera=0);

- X_2, razza della vittima (bianca =1/nera=0);

e definito secondo l'equazione (16):

$$logit[P(Y=1)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

in cui β_0 è la costante del modello, β_1 e β_2 rappresentano rispettivamente l'effetto della razza dell'imputato e della vittima.
Si importano le 674 osservazioni totali in *GretI* e l'output del programma fornisce le stime dei parametri (riquadro rosso), gli errori sulle stime, la significatività di ogni variabile indipendente e i valori dei test di bontà di adattamento del modello (**Figura 2.5**).

```
Modello 1: Logit, usando le osservazioni 1-674
Variabile dipendente: penadimorte
Errori standard basati sull'Hessiana

                coefficiente   errore std.      z       p-value
  ------------------------------------------------------------
  const           -3,59610      0,506914    -7,094   1,30e-012 ***
  razzavittima     2,40444      0,600616     4,003   6,25e-05  ***
  razzaimputato   -0,867797     0,367074    -2,364   0,0181    **

Media var. dipendente  0,100890   SQM var. dipendente    0,301407
R-quadro di McFadden   0,049646   R-quadro corretto      0,036036
Log-verosimiglianza   -209,4783   Criterio di Akaike     424,9565
Criterio di Schwarz    438,4962   Hannan-Quinn           430,1995
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 606 (89,9%)
f(beta'x) nella media delle variabili indipendenti = 0,077
Test del rapporto di verosimiglianza: Chi-quadro(2) = 21,8861 [0,0000]

            Previsto
              0     1
Effettivo 0  606     0
          1   68     0
```

Figura 2.5. Output del software *GretI* relativo all'Esempio 6

Di conseguenza, il modello assume la forma:

$$logit[P(Y=1)] = -3.596 - 0.868x_1 + 2.404x_2$$

I segni dei parametri forniscono importanti indicazioni:

- il segno negativo di β_1 indica che la razza bianca ha una probabilità inferiore di condanna rispetto alla razza nera;
- il segno positivo di β_2 indica che nel caso di vittima bianca, la probabilità di condanna è superiore rispetto al caso di vittima nera.

A partire dall'espressione del *logit* è possibile passare alla definizione della probabilità stimata estendendo l'equazione (11) al caso di regressione logistica multivariata:

$$\hat{\pi}(x) = \frac{\exp(-3.596 - 0.868x_1 + 2.404x_2)}{1 + \exp(-3.596 - 0.868x_1 + 2.404x_2)}$$

Per esempio, nel caso di imputato di razza nera (x_1=0) e vittima bianca (x_2=1) si ottiene una probabilità stimata di condanna a morte pari a:

$$\hat{\pi}(x) = \frac{\exp(-3.596 - 0.868 \cdot 0 + 2.404 \cdot 1)}{1 + \exp(-3.596 - 0.868 \cdot 0 + 2.404 \cdot 1)} = 23.3\%$$

a fronte della probabilità del 22.9% calcolata nel campione di dati.
Il valore dell'odds ratio tra la variabile dipendente e la variabile 'razza dell'imputato' è calcolato come:

$$OR_1 = e^{\beta_1} = e^{-0.868} = 0.42$$

Il risultato indica che l'odds della condanna a morte nel caso di imputato bianco è circa la metà dell'odds del caso di imputato nero, cioè la tendenza alla condanna nel primo caso è circa la metà rispetto al secondo.
Il valore dell'odds ratio tra la variabile dipendente e la variabile 'razza della vittima' è calcolato come:

$$OR_2 = e^{\beta_2} = e^{2.404} = 11.1$$

Il risultato indica che l'odds della condanna a morte nel caso di vittima bianca è 11.1 volte il valore del caso di vittima nera, cioè la propensione alla condanna nel primo caso è circa 11 volte maggiore rispetto al secondo caso.
Si esamina adesso la significatività di ogni singolo parametro del modello attraverso la statistica di *Wald*. Con riferimento all'equazione (23), si esegue il test Z per la variabile indipendente X_1 nei confronti dell'ipotesi nulla H_0 (β_1=0):

$$Z = \frac{b_j}{s(b_j)} = -\frac{0.868}{0.367} = 2.36$$

Dove i valori di $s(b_j)$ sono tratti dalla colonna *'errore std.'* in **Figura 2.5**.
Nel caso di α=0.05 si ottiene:

$$|Z| = 2.364 > z_{\frac{\alpha}{2}} = 1.96$$

La corrispondente statistica di *Wald* è:

$$W^2 = \left[\frac{b_j}{s(b_j)}\right]^2 = 5.59$$

ed il *p-value* è pari a 0.018 < 0.05 (livello di significatività scelto).

Si applicano le stesse formulazioni per la variabile indipendente X_2 e si ricava:

$$Z = \frac{b_j}{s(b_j)} = \frac{2.404}{0.601} = 4.003$$

$$|Z| = 4.003 > z_{\frac{\alpha}{2}} = 1.96$$

$$W^2 = \left[\frac{b_j}{s(b_j)}\right]^2 = 16.03$$

ed il *p-value* è pari a 0.000006 < 0.05.
I calcoli appena effettuati restituiscono i valori della variabile Z riportati in **Figura 2.5** e indicano che i parametri β_0, β_1 e β_2 del modello sono significativi.
La significatività dei parametri è espressa nell'output di *Gretl* dagli asterischi che contraddistinguono la colonna alla destra dei *p-value*: in particolare, questo segnalatore visivo indica che la significatività di un parametro aumenta all'aumentare del numero di asterischi, fino a un massimo di 3.
Si testa infine la bontà di adattamento del modello attraverso il test *rapporto di verosimiglianza*.
Si definisce la variabile *G* come il rapporto tra due verosimiglianze:

$$G = -2log\frac{L(0)}{L(\beta)} = 21.886$$

Nello specifico:

- *L(0)* è definita come la funzione di massima verosimiglianza in corrispondenza del modello caratterizzato dalla sola intercetta;
- *L(β)* è la funzione di massima verosimiglianza in corrispondenza del modello completo.

Tale valore è riportato nell'output in **Figura 2.5** come *'Chi-quadro (2)'* e attesta che, nel caso di 2 gradi di libertà, l'effetto delle variabili indipendenti è significativo poiché:

$$G = 21.886 > \chi^2_{\frac{\alpha}{2}} = 7.38$$

Il corrispondente *p-value* è minore di 0.05 e, di conseguenza, l'ipotesi nulla può essere rifiutata.

Capitolo 3

La regressione polinomiale

L'analisi di regressione polinomiale è una metodologia impiegata per esaminare la relazione causale che lega una variabile dipendente a una o più variabili indipendenti esplicative. Nello specifico, la variabile dipendente è di tipo quantitativo e, nel caso studio in esame, indicherà il numero di rotture previste per una determinata classe di condotte. Tale variabile sarà indicata dapprima come *Y* e successivamente come *NR* (*numero di rotture/*B*urst Rate)*.

3.1 Il modello di regressione polinomiale (EPR)

Lo studio del modello di regressione polinomiale ha la finalità di stimare l'influenza di molteplici parametri che differenziano le condotte (ad esempio il diametro, il materiale, l'anno di posa, la pressione e la lunghezza) sul numero di rotture previste per determinate classi di aggregazione.
Il modello polinomiale qui esposto, denominato anche *Evolutionary Polynomial Regression*, fa riferimento a quanto riportato dall'autore *Berardi* nella pubblicazione *"Development of pipe deterioration models for water distribution systems using EPR", Berardi et al. (2008).*
Il modello polinomiale adottato per definire la dipendenza tra il numero di rotture per una determinata classe di condotte e le caratteristiche delle stesse è contraddistinto dalla seguente formulazione:

$$Y = a_0 + \sum_{j=1}^{m} a_j X_1^{ES(j,1)} \cdot\ldots\cdot (X_k)^{ES(j,k)} \cdot f\left(X_1^{ES(j,k+1)} \cdot\ldots\cdot (X_k)^{ES(j,2k)}\right) \quad (1)$$

dove:

- *Y* è la variabile indipendente, denominata anche *NR,* ed indica il numero di rotture;
- X_k è la generica variabile esplicativa (ad esempio diametro, materiale, anno di posa e lunghezza);
- *ES* è la matrice degli esponenti incogniti;
- a_j è il generico coefficiente polinomiale da stimare;
- *m* è il numero di termini polinomiali addizionali alla costante a_0;

- f rappresenta una funzione arbitraria che può incrementare la capacità di previsione delle rotture.

I parametri del modello appena definiti sono calcolati a partire da una popolazione campionaria di dimensioni finite e la loro stima avviene attraverso il software statistico *Gretl.* Nello specifico, il numero di coefficienti polinomiali e di esponenti incogniti è imposto e variato arbitrariamente dall'utente, affinché il modello si adatti nel miglior modo possibile al campione.

3.2 La scelta delle classi

Un importante e preliminare aspetto è la scelta delle classi di condotte: ogni classe contiene un determinato numero di condotte ed è contraddistinta da un'età equivalente A_{Classe} e da un diametro equivalente D_{Classe}, pesati rispetto alla lunghezza totale delle condotte di ogni classe e definiti secondo le relazioni:

$$A_{Classe} = \frac{\sum_{Classe}(L_p \cdot A_p)}{L_{Classe}} \tag{2}$$

$$D_{Classe} = \frac{\sum_{Classe}(L_p \cdot D_p)}{L_{Classe}} \tag{3}$$

- L_p è la lunghezza di ogni condotta facente parte di una determinata classe;
- A_p e D_p rappresentano rispettivamente l'età e il diametro di ogni condotta facente parte di una determinata classe;
- L_{Classe} indica la lunghezza totale delle condotte facenti parte di una determinata classe.

Pesare i dati rispetto alla lunghezza ha una rilevanza statistica poiché questa indicazione include informazioni non disponibili e correlate alla stessa quali: variabilità dei carichi stradali, variabilità dei valori di carico idraulici e carichi del terreno.
Si riporta di seguito un esempio per chiarire la procedura di aggregazione dei dati in classi.

Esempio 1.

Si supponga di dover esaminare una rete composta da 9 condotte, contraddistinte da un codice identificatore, numero di rotture, età, lunghezza e diametro (**Tabella 3.1)**.

Tabella 3.1. Composizione della rete esempio

idPipe	BR	A_p (anni)	L_p (m)	D_p (mm)
1	1	30	10	63
2	3	30	55	63
3	0	30	35	63
4	0	40	10	75
5	5	40	55	75
6	2	40	5	90
7	2	25	10	100
8	3	25	35	100
9	4	25	15	100

Definito con *N* il numero totale di condotte per una determinata classe (3 per ogni classe) e applicando le equazioni **2** e **3**, si ottengono i risultati in **Tabella 3.2**. Nello specifico, si osservano 3 classi, formate dallo stesso numero di condotte, caratterizzate da un'età e un diametro equivalenti e, per i campi relativi alle rotture e alla lunghezza, dalla somma dei valori riscontrati per ogni singola condotta.

Tabella 3.2. Condotte raggruppate in classi di diametri ed età equivalenti

Classe	BR	A_p (anni)	L_p (m)	D_p (mm)	N
1	4	30	100	63	3
2	7	40	70	76	3
3	9	25	60	100	3

La suddivisione in classi per reti reali è ben più complessa di quanto riportato nell'esempio. Infatti, affinché la classificazione sia ragionevole, a seguito di una suddivisione per i differenti materiali, è necessario che ogni classe risulti composta da un numero comparabile di elementi. In questo contesto si inserisce l'importanza del calcolo delle distribuzioni di probabilità per i diversi diametri ed età. Questo procedimento verrà ampiamente descritto nel **Capitolo 9**, riguardante i risultati del modello polinomiale.

3.3 La stima dei parametri

In problemi di regressione non lineare, il metodo dei minimi quadrati generalmente non ammette soluzioni in forma chiusa, per cui si deve fare ricorso a metodi numerici per ricavare le stime dei parametri. Si applica, in genere, il metodo dei *minimi quadrati non lineari*.

Si prenda in esame un campione di *n* elementi caratterizzato dalla variabile dipendente *Y*, le variabili indipendenti (o regressori) x_i, con $i=1, 2, \ldots, n$ e ϑ parametri. L'**Equazione 1** può essere riscritta come:

$$y_i = f(x_i, \theta) + \varepsilon_i, \quad i = 1,2, \dots, n \tag{4}$$

La funzione dei minimi quadrati è definita come:

$$S(\vartheta) = \sum_{i=1}^{n} [y_i - f(x_i, \theta)]^2 \tag{5}$$

Per ricavare le stime dei parametri è necessario derivare l'**Equazione 5** rispetto a ogni parametro ϑ. Avendo a disposizione *p* parametri, si otterrà un sistema di *p* equazioni, definite come:

$$\sum_{i=1}^{n} [y_i - f(x_i, \theta)]^2 \left[\frac{\partial f(x_i, \theta)}{\partial \theta_j}\right] \quad per\ j = 1, 2, \dots, p \tag{6}$$

Le derivate in parentesi quadre sono funzione dei parametri incogniti, inoltre la funzione *f* è non lineare.
Per risolvere tale problema si utilizzano metodologie quali la massimizzazione della *funzione di massima verosimiglianza* (esposta nel **Capitolo 2**) o la linearizzazione delle funzioni non lineari mediante il metodo iterativo di *Gauss-Newton* che porta a convergenza il processo di stima dei parametri per step successivi.

Esempio 2.
Si supponga di voler stimare i parametri della regressione polinomiale con riferimento ai dati della **Tabella 3.3**. Nello specifico, è stato suddiviso il campione iniziale in 9 classi (classe 1-1, 1-2, …,3-3). Si ipotizzi, inoltre, che le condotte siano caratterizzate dallo stesso materiale. Le variabili indipendenti sono il diametro, l'età e la lunghezza, mentre la variabile dipendente è il numero di rotture (*NR*).

Tabella 3.10. Classi di condotte

Classe		D_{classe} (mm)	$Età_{classe}$ (anni)	L_{tot} (m)	NR
1	1	269.21	12.57	9580	1
	2	263.59	23.41	4690	2
	3	161.1	53.03	1870	2
2	1	492.14	13.36	1590	0
	2	463.43	23.18	5220	0
	3	456.47	115.79	680	0
3	1	781.19	12.93	6790	0
	2	749.21	25.34	14000	0
	3	769.43	47.46	4360	0

Si cerca un'equazione nella forma:

$$NR = \alpha \cdot Diametro^{\beta} \cdot Età^{\gamma} \cdot Lunghezza^{\delta}$$

Si riporta l'output di *Gretl* in **Figura 3.1**.
Una volta stimati i parametri secondo i minimi quadrati non lineari, il modello assume la forma:

$$NR = 5668.98 \cdot Diametro^{(-2.16)} \cdot Età^{(0.06)} \cdot Lunghezza^{0.41}$$

Applicando tale formulazione ai dati iniziali, si ottiene la **Tabella 3.2**.

```
Tolleranza = 1.81899e-012
Convergenza raggiunta dopo 533 iterazioni

Modello 1: NLS, usando le osservazioni 1-9
BR = alpha*D^beta*EtA^gamma*L^delta

                stima        errore std.    rapporto t    p-value
  ------------------------------------------------------------------
  alpha       5668.98          94678.1          0.05988     0.9546
  beta         -2.16331            1.06481     -2.032       0.0979   *
  gamma         0.0161396          1.52892      0.01056     0.9920
  delta         0.405219           1.33457      0.3036      0.7736

Media var. dipendente   0.555556   SQM var. dipendente      0.881917
Somma quadr. residui    1.234415   E.S. della regressione   0.496873
R-quadro non centrato   0.801612   R-quadro centrato       -0.033520
Log-verosimiglianza    -3.830623   Criterio di Akaike       15.66125
Criterio di Schwarz     16.45015   Hannan-Quinn             13.95881
Note: SQM = scarto quadratico medio; E.S. = errore standard

GNR: R-squared = 7.0176e-012, max |t| = 5.84172e-006
La convergenza sembra ragionevolmente completa
```

Figura 3.1. Output di *Gretl* con riferimento all'esempio 2

Tabella 3.2. Confronto tra le rotture osservate e quelle stimate

Classe		NR	NR stimate
1	1	1	1.60
	2	2	1.29
	3	2	2.70
2	1	0	0.21
	2	0	0.40
	3	0	0.20
3	1	0	0.14
	2	0	0.21
	3	0	0.13

L'output afferma la significatività, seppur minima, del singolo parametro β, nel caso di significatività pari al 10%. Inoltre, è possibile osservare che il numero di rotture aumenta all'aumentare dell'età e della lunghezza, mentre diminuisce all'aumentare del diametro. I valori dei *p-value* indicano che, probabilmente, la

forma conferita inizialmente all'equazione in via arbitraria non è adatta a descrivere in maniera ottimale il campione, poiché i parametri risultano non significativi.
Si riporta in **Figura 3.2** il confronto tra il numero di rotture osservato e quello calcolato per ogni classe di condotte, una volta numerate da 1 a 9 le classi.

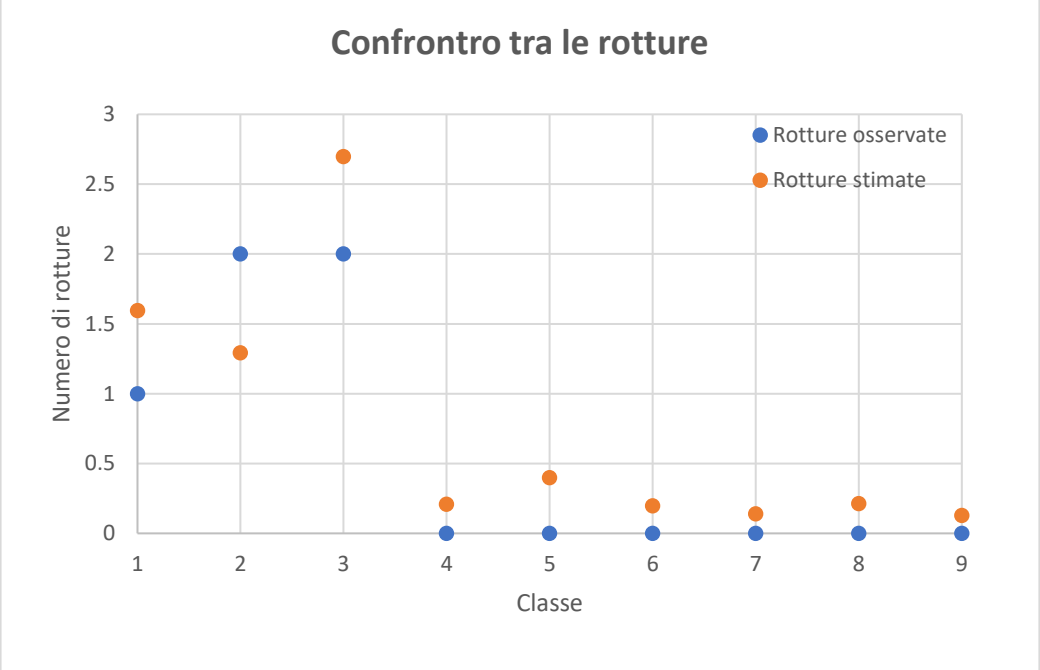


Figura 3.2. Confronto tra le rotture osservate e stimate per le 9 classi del campione

3.4. Il *Coefficiente di Determinazione (CoD)*

Ogni modello stimato secondo l'**Equazione 1** è caratterizzato da un determinato livello di bontà di adattamento, definito dal *Coefficiente di Determinazione (CoD)* secondo la formulazione:

$$CoD = 1 - \frac{\sum_n \left(\hat{y} - y_{exp}\right)^2}{\sum_n \left(y_{exp} - avg\left(y_{exp}\right)\right)^2} \qquad (7)$$

dove:

- *n* è il numero di elementi del campione;
- $\hat{y}$ è il numero di rotture previste dal modello per una determinata classe;
- y_{exp} indica il numero di rotture osservate per una determinata classe;
- $avg(y_{exp})$ indica il valore medio delle rotture per una determinata classe.

La definizione di questo parametro è di fondamentale importanza poiché permette di posizionare ogni modello stimato sulla frontiera di *Pareto* e sarà indispensabile per la

scelta del modello migliore per ogni classe di condotte. Infatti, con riferimento a ciascuna classe di condotte, verranno stimati più modelli, caratterizzati da un numero sempre più alto di coefficienti polinomiali, differenti esponenti incogniti e conseguenti valori del coefficiente di determinazione.
Il range di variazione del *CoD* è nell'intervallo tra 0 e 1: nello specifico, migliore è la bontà di adattamento del modello, più alto è il valore del coefficiente.
In generale, un numero più elevato di coefficienti polinomiali e variabili indipendenti porta a un valore più elevato del coefficiente di determinazione. È altresì vero che, superata una determinata soglia, un ulteriore incremento nel numero di variabili e coefficienti polinomiali può portare a incrementi trascurabili della bontà di adattamento del modello.
La scelta della relazione più idonea a rappresentare una classe sarà affidata a considerazioni di tipo ingegneristico con riferimento alle *frontiere di Pareto* per le coppie a_j*-(1-CoD) e* X_K*-(1-CoD)*.

> **Esempio 3.**
> Con riferimento **all'Esempio 2** e all'output riportato in **Figura 3.1**, il coefficiente di determinazione del modello è stato stimato dal software ed è riportato come *R-quadro non centrato*. Per il modello così composto, ammonta a 0.801612 e indica che il numero di rotture stimate per ogni classe approssima in maniera piuttosto adeguata il numero di rotture osservato, ma il modello può essere migliorato, cambiando il numero di polinomi ed esponenti incogniti. Come esposto in precedenza, infine, i parametri risultano non significativi.

3.2 La procedura di applicazione del modello polinomiale

Di seguito vengono riassunti i passi appena descritti per l'applicazione del modello di regressione polinomiale:

1. A partire dal database contenente le condotte della rete in esame, complete di informazioni riguardanti il diametro, materiale, anno di posa e lunghezza delle stesse, è necessario estrarre delle sotto-tabelle riguardanti i singoli materiali. Nel caso in esame, il campione di partenza è la tabella *Esatta* e i materiali analizzati sono la ghisa grigia, ghisa sferoidale, eternit ed acciaio. Le conseguenti sotto-tabelle saranno denominate come *EsattaGhisaGrigia*, *EsattaGhisaSferoidale*, *EsattaEternit* ed *EsattaAcciaio.*
2. Per tutte le sotto-tabelle è effettuata una suddivisione in classi: a seguito di tale stratificazione, ogni classe sarà caratterizzata da un diametro e da un'età equivalenti pesati rispetto alla lunghezza. La suddivisione deve essere effettuata in modo tale che ogni classe contenga un numero comparabile di condotte, per poter assicurare una valenza statistica. Affinché questo sia rispettato, nel caso di

reti complesse, è necessario studiare la distribuzione di probabilità di diametri ed età delle condotte facenti parte delle sotto-tabelle.

3. Nel caso studio in esame, sono state scelte 3 classi di suddivisione per diametri ed età. Ne seguono 9 classi di condotte per ogni materiale, contraddistinte da diametro ed età equivalenti e dalla somma delle lunghezze e del numero di rotture di ogni condotta.
4. Una volta noti questi dati, tramite il software *Gretl* si formulano differenti forme dell'**Equazione 1**, si stimano i parametri del modello e i rispettivi coefficienti di determinazione. Ogni forma analizzata sarà contraddistinta da un certo numero di coefficienti polinomiali ed esponenti incogniti che, insieme al *CoD* si posizioneranno in un punto sulla *frontiera di Pareto*. L'analisi dell'*ottimo paretiano* porterà alla scelta dell'equazione ottimale per ogni materiale.
5. Infine, si valuta la significatività dei singoli parametri, come visto nel **Capitolo 2**, nel caso della regressione logistica.

Si riporta un esempio che chiarisca in maniera integrale i 5 passi appena descritti.

Esempio 4.

Si prenda in considerazione un campione di condotte e si estraggano tutte quelle caratterizzate dal materiale Ghisa grigia. Si decide di suddividere i dati in 3 classi di diametri e 3 classi di età.

Le corrispondenti distribuzioni di probabilità e le classi di variazione di diametri ed età sono calcolate come segue:

1. Si calcolano le occorrenze per tutti i valori di diametri ed età.
2. Per ogni valore si calcola la funzione densità di probabilità, come il rapporto tra il numero di occorrenze di un valore di diametro/età e le occorrenze totali di tutti i valori di diametri/età. A titolo di esempio, se ad un determinato diametro corrispondono 204 occorrenze e il numero totale di elementi è 2775, la corrispettiva funzione densità di probabilità *f(x)* sarà (204·100)/2775=7.46%.
3. Per ogni valore, si calcola la funzione probabilità cumulata *F(x)* come la sommatoria progressiva delle densità di probabilità *f(x)* di tutti gli elementi precedenti a quello in esame. Si riportano in **Figura 3.3** e **3.4** le funzioni appena descritte per tutti i valori di diametri ed età.
4. Scelto il numero di classi, si suddivide la *F(x)* in 3 classi, cercando di ottenere degli intervalli di larghezza pari al 33.33%. Si ricavano i corrispondenti intervalli di variazione di diametri ed età mediante i quali suddividere i dati del campione.

   Applicando tale procedura al campione di condotte, si ottengono le classi riportate in **Tabella 3.3**.

Su tali classi sono stati stimati parametri di differenti modelli polinomiali, diversi per numero di termini ed esponenti incogniti.
Le possibili combinazioni delle variabili indipendenti, i corrispettivi modelli e i coefficienti di determinazione stimati sono riportati in **Tabella 3.4.**

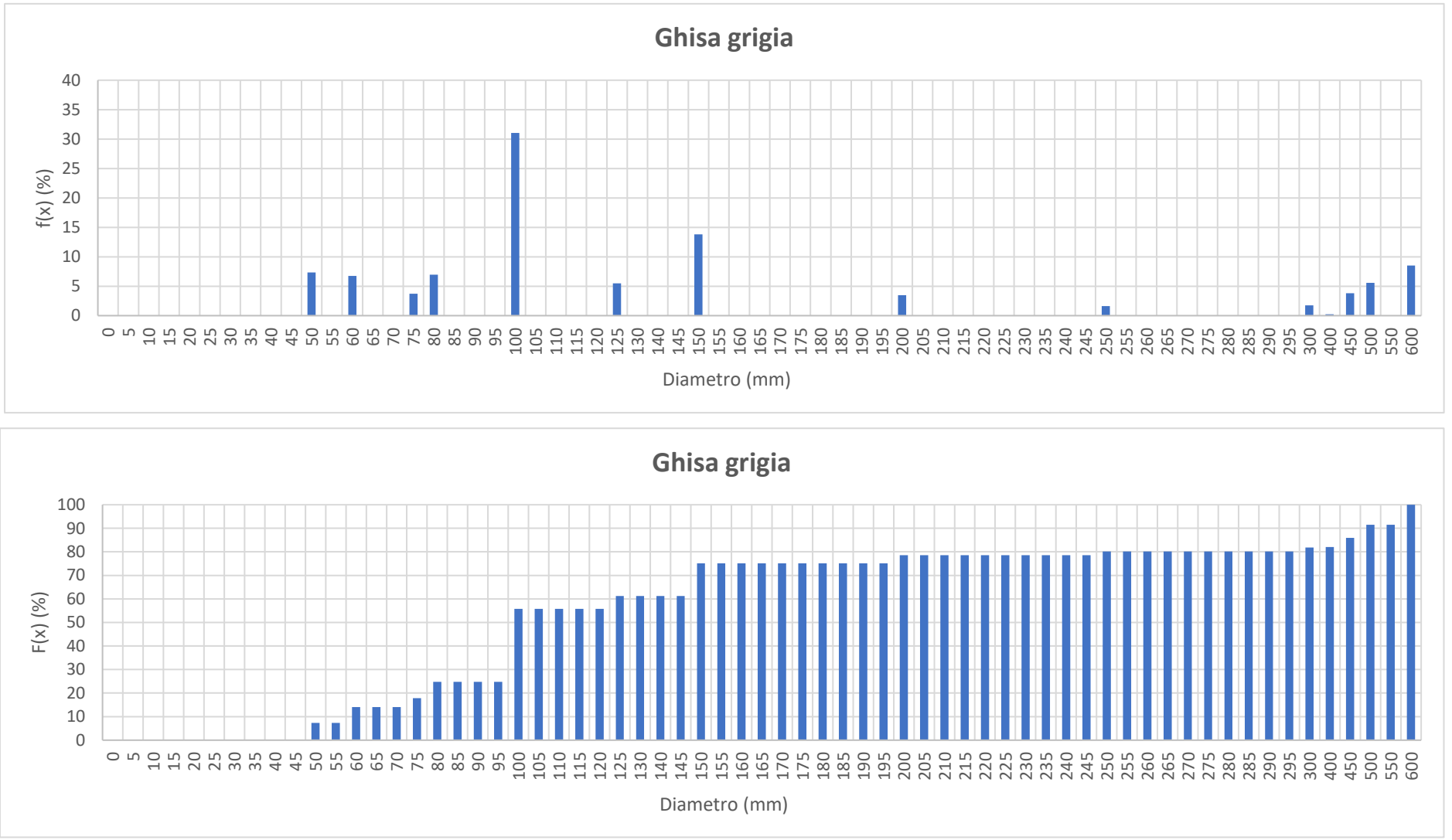


Figura 3.3. Funzione densità di probabilità e funzione di probabilità cumulata per i diametri

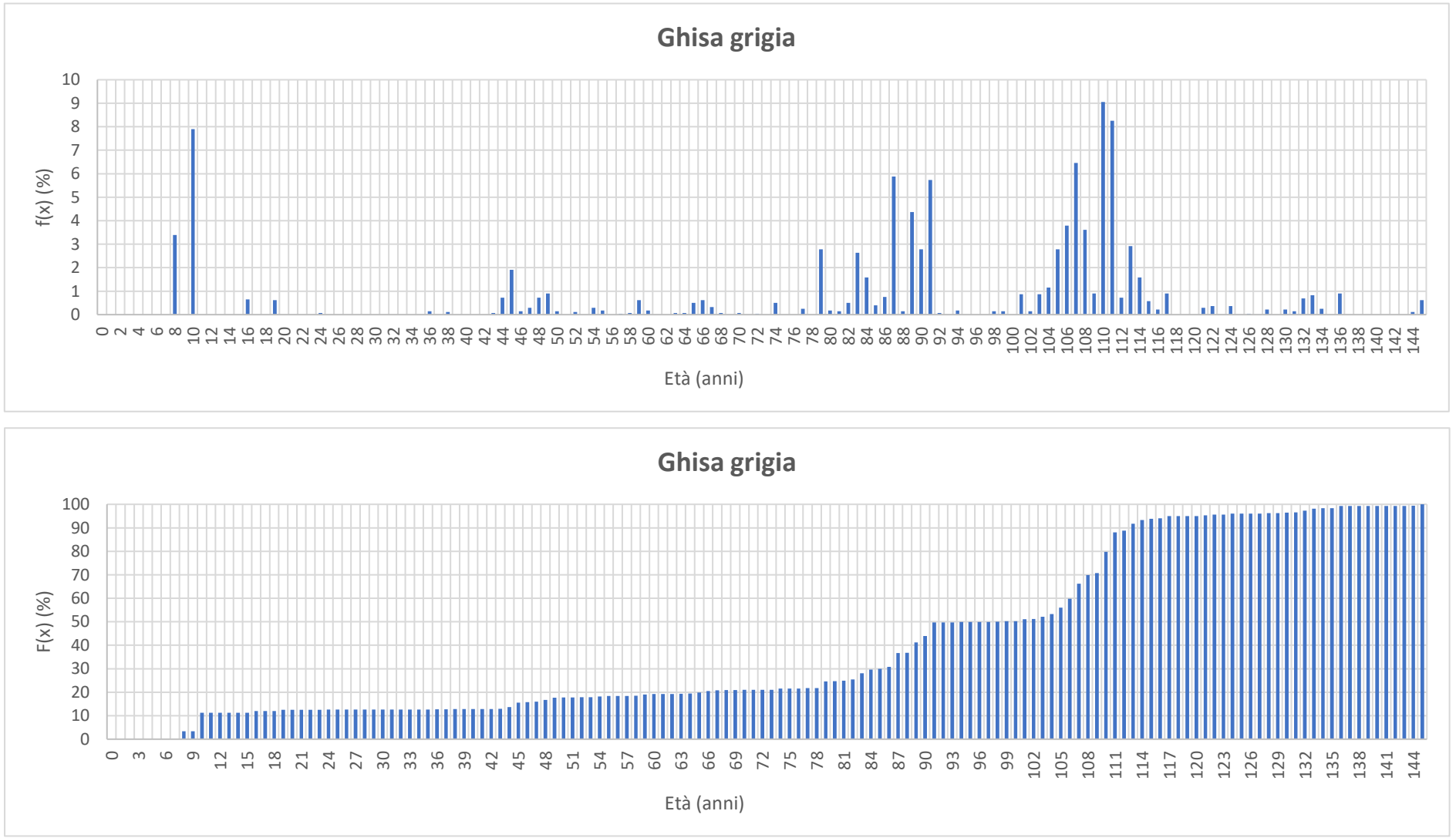


Figura 3.4. Funzione densità di probabilità e funzione di probabilità cumulata relative all'età delle condotte

Tabella 3.3. Campione di condotte suddiviso in classi

Ghisa Grigia					
Classe		D_{classe} (mm)	$Età_{classe}$ (anni)	L_{tot} (m)	NR
1	1	60.69	35.98	1660	9
	2	58.96	95.88	1780	3
	3	57.84	122.23	1480	1
2	1	97.74	45.7	4070	19
	2	94.26	100.35	2230	8
	3	96.08	111.47	4250	10
3	1	315.18	47.08	2800	1
	2	336.51	90.45	4040	1
	3	306.27	111.7	5440	7

Tabella 3.4. Formulazioni utilizzate per stimare il numero di rotture e relativi *CoD*

Ghisa grigia				
Formula	CoD	No. X_h	No.A_j	1-CoD
$NR = 20.47 \cdot D^{-0.24}$	0.05	1	1	0.95
$NR = 0.0015 \cdot L^{1.71} \cdot D^{-1.11}$	0.63	2	1	0.37
$NR = 1.92 \cdot 10^{-8} \cdot \ln(L^{13.42}) \cdot D^{-1.05} \cdot A^{-0.76}$	0.87	3	1	0.13
$NR = 0.0068 \cdot (L^{1.62}) \cdot D^{-1.04} \cdot A^{-0.84} \cdot \ln(L \cdot A)$	0.87	5	1	0.13
$NR = 6.14 \cdot 10^{-8} \cdot \ln(L^{13.51}) \cdot D^{-1.04} \cdot lnA^{-3.20}$	0.87	3	1	0.13
$NR = 0.81 \cdot L^{1.12} \cdot D^{-0.89} \cdot A^{-0.49} - 41.73 \cdot D^{0.49}$	0.88	4	2	0.12

Diagrammando le sei coppie di punti *(1-Cod)-No.X_h* (numero di variabili indipendenti) si ottiene un grafico raffigurante una frontiera di Pareto (**Figura 3.5**).

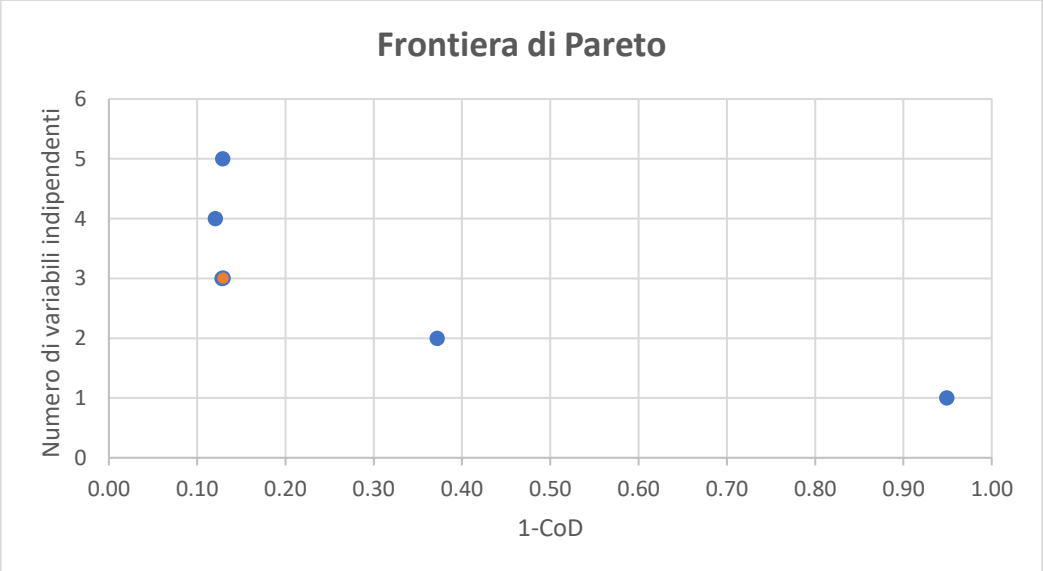


Figura 3.5. Coppie di punti relative ai valori (1-CoD)-No.X_h di ogni modello stimato

In arancione si riporta il punto ottimale: infatti, un numero di variabili maggiore di 3 non porta un incremento elevato nella bontà di adattamento del modello. In definitiva, il modello ottimale assume la forma:

$$NR = 1.92 \cdot 10^{-8} \cdot \ln(L^{13.42}) \cdot D^{-1.05} \cdot A^{-0.76}$$

Il confronto tra le rotture osservate e previste è riportato in **Tabella 3.5**.

Tabella 3.5. Confronto tra le rotture osservate e quelle stimate dal modello ottimale

Classe		NR	NR stimate
1	1	9	8.03
	2	3	4.46
	3	1	2.71
2	4	19	18.80
	5	8	3.92
	6	10	10.42
3	7	1	2.90
	8	1	3.02
	9	7	4.55

Successivamente, numerate le classi da 1 a 9, è possibile riportare tali risultati nel grafico in **Figura 3.6**.

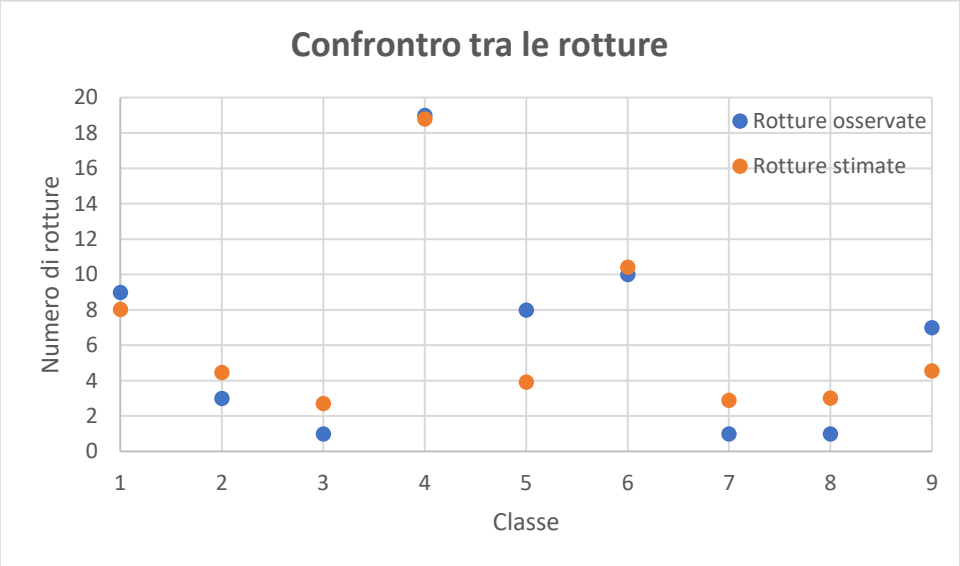


Figura 3.6. Confronto grafico tra le rotture osservate e quelle stimate per ogni classe

Per via grafica è più facile notare che il modello si avvicina al valore reale di rotture in corrispondenza di determinate classi, ma si allontana da una stima coerente col campione in corrispondenza di altre.

3.5 Valutazione della bontà del modello e significatività dei parametri

Dopo aver stimato un modello che possa rappresentare la relazione che lega la variabile indipendente ai suoi regressori, è di fondamentale importanza valutare la sua efficacia di adattamento ai dati campionari. Di analogo interesse è la valutazione del contributo specifico che ogni variabile indipendente apporta sulla stima della variabile dipendente, al fine di valutare il grado di informazione che le stesse introducono nelle stime del modello.
Di seguito verranno elencati e descritti i test di valutazione della bontà di adattamento del modello e i test di significatività di ogni singolo parametro. Fanno parte del primo gruppo di test:

- Il coefficiente di determinazione;
- il test rapporto di verosimiglianza *G*.

Il primo è stato ampiamente descritto nel **Paragrafo 3.4**, mentre il secondo test è stato analizzato nel paragrafo **2.4.1**. Si riporta qui di seguito quanto esposto nei riguardi del test di verosimiglianza.
Si definisce la variabile *G* come il rapporto tra due verosimiglianze:

$$G = -2log\frac{L(0)}{L(\beta)} \tag{8}$$

Nello specifico:

- *L(0)* è definita come la funzione di massima verosimiglianza in corrispondenza del modello caratterizzato dalla sola intercetta;
- *L(β)* è la funzione di massima verosimiglianza in corrispondenza del modello completo.

La statistica *G* segue una distribuzione del χ^2 con un numero di gradi di libertà pari alla differenza tra il numero di parametri del modello completo e il numero di parametri del modello ridotto.
Attraverso questo test si verifica che le variabili del modello aggiungano molte informazioni rispetto al modello caratterizzato dalla sola intercetta: infatti, se questo avviene, la quantità *L(β)* risulta essere più grande di *L(0)* e di conseguenza il rapporto tra le verosimiglianze raggiunge valori molto piccoli prossimi allo 0.
Si sottopongono a verifica le ipotesi H_0 e H_1 che implicano:

$$H_0: \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_1: almeno\ un\ parametro \beta_j \neq 0, j = 1, \dots, p$$

L'ipotesi *H_0*, definita ipotesi nulla, implica che nessuna delle *p* variabili indipendenti apporti un contributo significativo alle stime del modello, mentre l'ipotesi *H_1* afferma che almeno una variabile indipendente sia diversa da zero e che quindi sia significativa.
L'ipotesi nulla può essere rifiutata se sussistono due condizioni:

- il valore di *G* è maggiore del valore tabellato del $\chi 2_{\frac{\alpha}{2}}$, funzione di α e del numero di gradi di libertà.

$$G > \chi^2_{\frac{\alpha}{2}} \tag{9}$$

  dove α è la significatività scelta per il test.
- il corrispondente valore del *p-value* è inferiore al livello di significatività scelto.

Se le condizioni risultano valide, è possibile rifiutare l'ipotesi nulla e questo implica che la potenza predittiva del modello viene migliorata dalla presenza delle variabili indipendenti.
A titolo di esempio, si faccia riferimento all'output riportato in **Figura 3.1**: il valore della statistica *G (Test del rapporto di verosimiglianza)* è pari a 3.83, minore del valore del $\chi 2_{\frac{\alpha}{2}}$ con tre gradi di libertà e significatività pari a 0.05.

$$G < \chi^2_{\frac{\alpha}{2}} \rightarrow 3.83 < 9.35$$

Il risultato afferma che il modello non è in grado di prevedere delle rotture, in quanto non adatto a descrivere il campione di dati e, di conseguenza, non è possibile rifiutare l'ipotesi nulla.
Inoltre, il *p-value* consente di valutare la significatività dei singoli parametri: affinché un parametro sia significativo, il valore del *p-value* deve essere minore del livello di significatività scelto. In questo caso, con un livello di significatività pari al 5%, nessun parametro risulta essere significativo.

Capitolo 4

La composizione della rete idrica di Torino

Nel seguente capitolo sarà descritta la composizione della rete acquedottistica gestita dal gruppo *SMAT* (*Società Metropolitana Acque Torino*) con particolare riferimento a quella del solo comune di Torino, oggetto di studio del presente lavoro.
La rete di condotte che caratterizza il comune è rappresentata in **Figura 4.1**: sono riportate in colore magenta le condotte presenti sul territorio torinese, quest'ultimo contraddistinto da colorazione grigia.

Figura 4.1. Rappresentazione della rete di Torino

In particolare, la rete di condotte si sviluppa per 1640 chilometri ed è composta da 154725 tubazioni che compongono la rete di adduzione e distribuzione di Torino.
Si riporta di seguito una breve analisi dello stato della rete. I dati a cui si fa riferimento sono disponibili nel database cartografico della società *SMAT* e, in particolare, si pone attenzione alle caratteristiche delle condotte quali:

- collocazione geografica;
- lunghezza;

- materiale;
- diametro;
- anno di posa.

Per ogni tipologia di dato sono state effettuate delle suddivisioni in classi di valori. Successivamente sono state calcolate le relative percentuali di appartenenza alle classi rispetto al totale ed il numero di chilometri di condotte facenti parte di ogni classe. Queste informazioni sono riportate in maniera tabellare e mediante grafici a torta.

Lunghezza delle condotte

Si individuano 5 classi di lunghezza delle condotte nel range 0-600 metri. Si riportano i risultati in **Tabella 4.1**.

Tabella 4.1. Suddivisione delle condotte in classi di lunghezza

Lunghezza		N° condotte	%	Chilometri	%
CLASSE 1:	**0 m - 10 m**	129046	83.4	134.2	8.2
CLASSE 2:	**10 m - 50 m**	14558	9.4	385.62	23.6
CLASSE 3:	**50 m - 100 m**	7536	4.9	533.57	32.7
CLASSE 4:	**100 m - 300 m**	3365	2.2	496	30.4
CLASSE 5:	**300 m - 600 m**	220	0.1	84.2	5.2
	TOT	154725	100.0	1633.59	100.0

La tabella mostra chiaramente che la maggior percentuale di condotte è caratterizzata da lunghezze inferiori ai 10 metri che, però, contraddistinguono solo l'8.2% della lunghezza totale della rete. Infatti, il 63.1% della rete è caratterizzato da condotte comprese tra i 50 e i 300 metri.
Le stesse informazioni sono riportate graficamente in **Figura 4.2**.

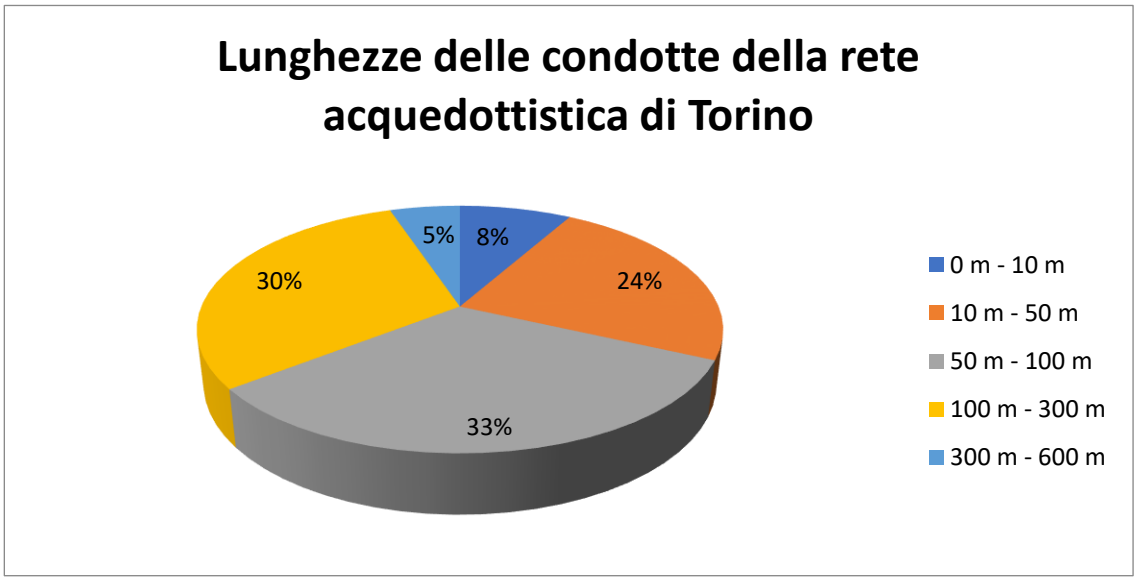


Figura 4.2. Suddivisione delle condotte in classi di lunghezza e relative percentuali

Materiale

Il materiale è uno dei campi che maggiormente influenza il comportamento delle condotte nei confronti delle rotture: infatti, ogni materiale presenta specifiche proprietà fisiche, chimiche e meccaniche che ne diversificano il comportamento.
Quelli più utilizzati nelle reti acquedottistiche risultano:

- di tipo metallico, come acciaio, ghisa, ferro zincato-nudo e piombo;
- di tipo plastico, come il cloruro di polivinile (PVC), polietilene a bassa densità (PEBD), polietilene ad alta densità (PEAD) e vetroresina;
- di tipo cementizio, come il calcestruzzo o il cemento amianto.

La scelta del materiale da utilizzare per una condotta dipende da diversi fattori di tipo tecnico ed economico. In particolare, è influenzata dalle caratteristiche delle acque convogliate e dei terreni in cui le opere sono collocate e dalle condizioni di esercizio cui le tubazioni sono sottoposte (carichi massimi, medi e minimi). In generale, la progettazione e il controllo delle tubazioni segue gli standard imposti dalla "*Normativa tecnica sule tubazioni*" contenuta nel *Decreto del Ministero dei lavori Pubblici del 12/12/1985*. Si riporta in **Tabella 4.2** la suddivisione delle condotte per classi di materiale, le relative percentuali di appartenenza e i chilometri di condotte facenti parte di ogni classe.

Tabella 4.2. Suddivisione delle condotte in classi di lunghezza

Materiale	N° Condotte	%	Chilometri	%
Ghisa grigia	79545	51.4	950.8	58.1
Ghisa sferoidale	31334	20.3	417.2	25.5
Acciaio	15409	10.0	180.9	11.1
Eternit	2775	1.8	53.9	3.3
MATERIALI INDEFINITI	24394	15.8	11.9	0.7
PEAD	754	0.5	8.9	0.5
Chameroy	32	0.0	4.9	0.3
PVC	248	0.2	4.1	0.2
Ferro	200	0.1	1.8	0.1
Cemento armato	7	0.0	0.7	0.0
Piombo	24	0.0	0.5	0.0
Pebd	6	0.0	0.1	0.0
TOT	154728	100.0	1635.6	100.0

È possibile notare che, tra i materiali metallici, la ghisa grigia è quella maggiormente utilizzata: infatti, caratterizza oltre la metà della lunghezza totale delle condotte. Presenta un'elevata resistenza alla corrosione ed altrettanto elevata fragilità. A partire dalla seconda metà dello scorso secolo, è stata gradualmente sostituita dalla ghisa sferoidale che presenta migliori caratteristiche meccaniche e a fatica.

A causa della suscettività dei materiali metallici alla corrosione, questi sono largamente impiegati dei rivestimenti protettivi.
Le tubazioni di tipo plastico hanno trovato spazio solo in tempi recenti, quindi è ancora in fase di studio il loro comportamento sul lungo periodo. Sono dei materiali economici e altamente deformabili.
Il calcestruzzo riveste un ruolo marginale, a causa della sua fragilità e alla sua incompleta impermeabilità.
Il cemento amianto è un materiale non più utilizzato a partire dal 1992, a causa della sua pericolosità all'inalazione.
Si riportano in **Tabella 4.3** le caratteristiche appena elencate e gli ambiti di applicazione dei materiali, con particolare riferimento ai vantaggi nell'utilizzo dell'uno o dell'altro.

Tabella 4.3. Utilizzi, vantaggi e svantaggi dei materiali presenti nella rete di Torino

Materiale	Utilizzo	Vantaggi	Svantaggi
Materiali metallici (acciaio e ghisa)	Per tutti i diametri e tutte le pressioni	Robustezza e resistenza a pressioni elevate	Necessità di trattamenti contro la corrosione, elevato peso e costo
Materiali plastici (PEAD)	Per piccoli diametri e pressioni fino a 25bar	Impermeabilità, alta levigatezza, basso costo e leggerezza	Bassa resistenza, elevata deformabilità
Calcestruzzo	Per grandi diametri e pressioni medie	Robustezza	Elevato peso e costo
Eternit	Vietato per legge (ancora presente in rete)	Leggerezza e basso costo	Nocivo a seguito di inalazione

Si riporta in **Figura 4.3** il grafico relativo alle informazioni della **Tabella 4.2**.

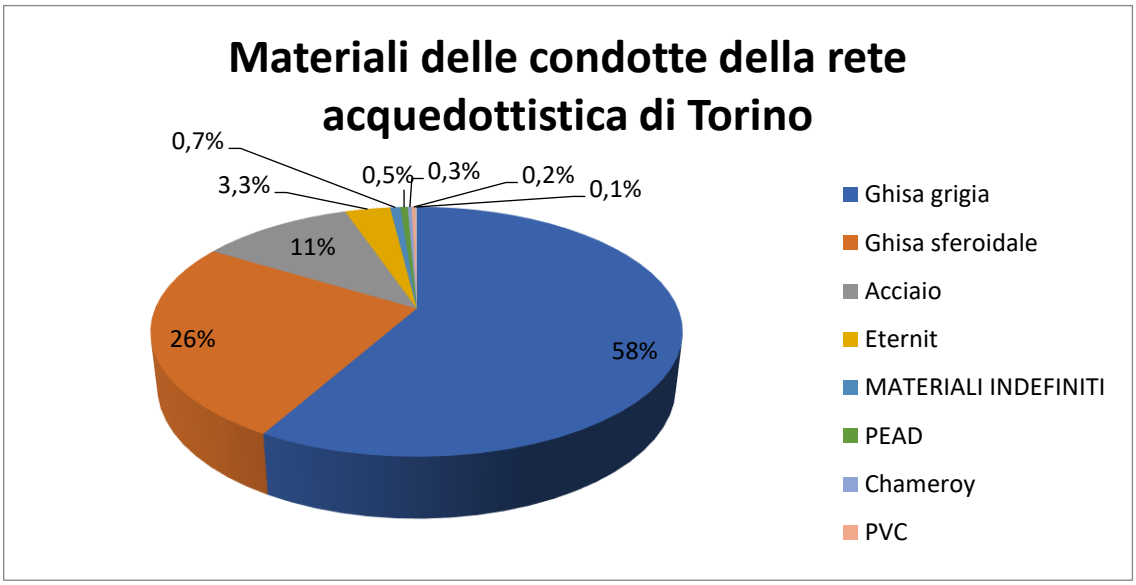


Figura 4.3. Suddivisione delle condotte in materiali e relative percentuali

Anche in questo caso è possibile notare che il materiale dominante è la ghisa grigia, in termini di condotte e di lunghezza ricoperta, seguita dalla ghisa sferoidale.

Diametro

Anche il diametro è una caratteristica molto importante che incide sul comportamento delle condotte. Nel database cartografico si ha a disposizione l'informazione riguardante il diametro nominale (DN) che risulta un'informazione convenzionale che fa riferimento a grandezze differenti. In generale, nel caso di materiali metallici e cementizi, esso coincide con il diametro interno, mentre per i materiali plastici fa riferimento al diametro esterno (comprensivo quindi dello spessore della tubazione).
Si riporta in **Tabella 4.4** la suddivisione delle condotte in classi di diametri.

Tabella 4.4. Suddivisione delle condotte in classi di diametro

Diametro		N° condotte	%	Chilometri	%
CLASSE 1:	**D ≤ 100 mm**	105375	68.1	888.5	54.3
CLASSE 2:	**100mm ≤ D ≤ 200 mm**	36383	23.5	411.2	25.1
CLASSE 3:	**D ≥ 200 mm**	12964	8.4	335.9	20.5
	TOT	154722	100.0	1635.6	100.0

La tabella mostra che oltre la metà dei diametri presenta dimensione inferiore ai 100 millimetri. Un'altra informazione riguarda la lunghezza delle condotte contraddistinte da questi diametri: la lunghezza media delle condotte con diametro inferiore ai 100 millimetri è più piccola della lunghezza media che caratterizza i diametri più grandi. A maggiori diametri, corrispondono maggiori lunghezze. Infatti, le condotte facenti parte della prima classe presentano lunghezza media di 8.4 metri, quelle di classe 2 di 11.3 metri e quelle di classe 3 di 25.9 metri.
Si riporta in **Figura 5.4** il grafico relativo alle informazioni della **Tabella 4.4**.

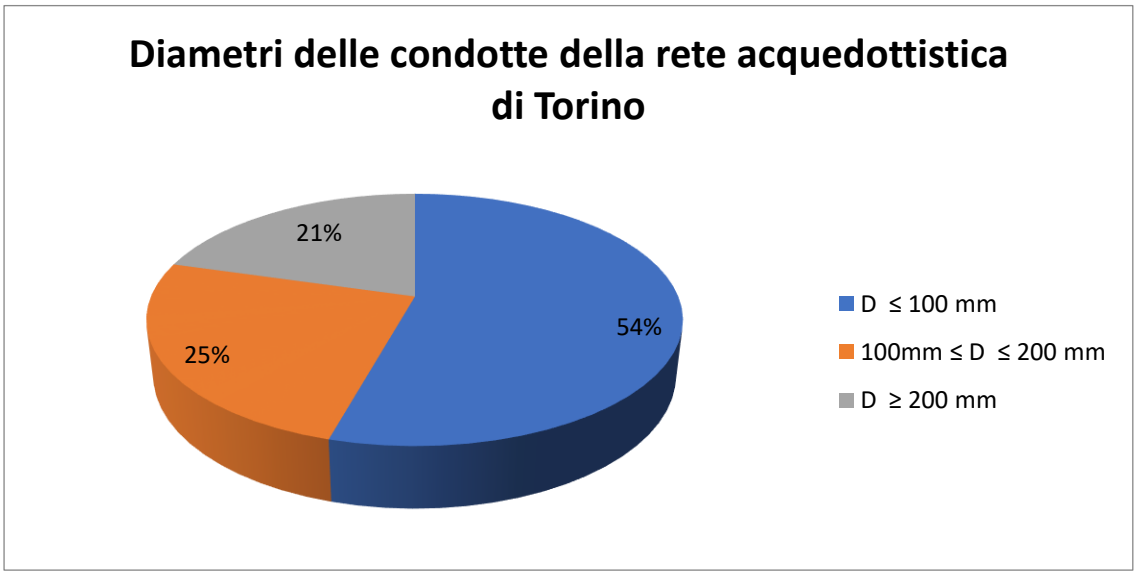


Figura 5.4. Suddivisione delle condotte in classi di diametro e relative percentuali

Anno di posa

L'anno di posa corrisponde all'informazione riguardante l'anno in cui la condotta è stata posata in trincea. Sfortunatamente, solo il 20% delle condotte possiede questa informazione (31903 condotte su 154725). Si riporta in **Tabella 4.5** la suddivisione in classi dei dati disponibili e in **Figura 4.5** il corrispettivo grafico a torta. Queste informazioni mostrano che la quasi totalità delle condotte è stata posata tra il 1990 e il 2010. Con molta probabilità, solo in quel ventennio si è cominciato a riportare tale dato in maniera più continuativa. Tale affermazione sarà riconfermata successivamente durante l'analisi dei singoli materiali. Infatti, si vedrà che i materiali di recente utilizzo, come la ghisa sferoidale e il polietilene, possiedono una copertura maggiore di tale dati rispetto ai materiali meno recenti.

Tabella 4.5. Suddivisione delle condotte in classi di età e relative percentuali

Anno di posa		N°condotte	%	Chilometri	%
CLASSE 1:	**1870 - 1890**	114	0.4	1.3	0.3
CLASSE 2:	**1890 - 1910**	936	2.9	14.4	3.6
CLASSE 3:	**1910 - 1930**	539	1.7	10.5	2.6
CLASSE 4:	**1930 - 1950**	258	0.8	4.1	1.0
CLASSE 5:	**1950 - 1970**	456	1.4	6.5	1.6
CLASSE 6:	**1970 - 1990**	2692	8.4	39.3	9.8
CLASSE 7:	**1990 - 2010**	24968	78.3	301	75.3
CLASSE 8:	**2010 →**	1940	6.1	22.6	5.7
	TOT	31903	100.0	399.7	100.0

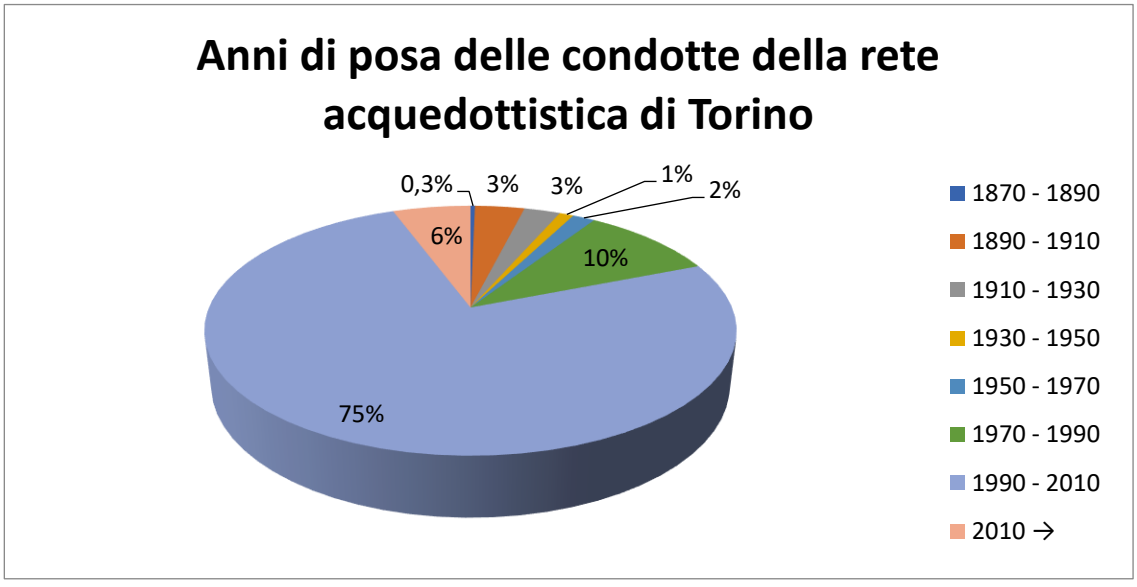


Figura 4.5. Suddivisione delle condotte in classi di età e relative percentuali

4.1 Conclusioni

In questo capitolo è stata analizzata la rete acquedottistica che fa parte del solo comune di Torino, prendendo in esame i dati a disposizione dal database cartografico riguardanti lunghezza, materiale, diametro e anno di posa delle condotte.
La rete, nel complesso, si presenta per oltre il 60% con condotte comprese tra i 50 e i 300 metri.
I materiali predominanti sono la ghisa grigia e sferoidale (contraddistinguono oltre il 70% delle condotte).
Il 68.1% delle condotte è caratterizzato da un diametro inferiore ai 100 millimetri e, infine, oltre il 70% delle condotte per le quali è disponibile l'informazione sull'età è stato posato nel ventennio 1990-2010.

Capitolo 5

Fonti dei dati

Nel seguente capitolo verranno brevemente descritte le modalità attraverso cui i dati alla base di questo lavoro sono stati raccolti e gestiti negli anni dal personale della Società Metropolitana Acque Torino (SMAT). Verrà fornita in aggiunta una descrizione delle tabelle fondamentali utilizzate per estrapolare le informazioni riguardo gli interventi del tipo '*fuga condotta*' che hanno avuto luogo nell'arco temporale tra l'anno 2006 e il 2016 nella rete di Torino: nello specifico, le tabelle *Woserviceaddress*, *Workorder*, e *Vie*.

Tutti gli interventi sul campo sono documentati dagli operatori che eseguono le riparazioni, assieme alle informazioni riguardanti le condotte coinvolte, la tipologia di operazione, il costo dell'intervento e la sua localizzazione. I dati sono stati raccolti in un database (la piattaforma *Maximo*) che agevola la lettura delle informazioni, la pianificazione degli interventi e la verifica delle condizioni della rete attraverso un approccio strettamente modulare che permettere di seguire le operazioni dalla segnalazione del guasto fino alla conclusione dell'intervento, immagazzinando infine l'intervento in memoria.

Come appena descritto, la compilazione dei documenti a seguito della segnalazione è effettuata su tablet dai tecnici preposti all'intervento ed in molti casi sono state omesse le voci relative ad alcuni campi: questo porterà a filtrare i dati affinché il modello di previsione possa disporre di un campione di dati quanto più completo possibile relativo alle rotture delle condotte.

A partire dal database contenente tutti gli interventi effettuati nel sistema acquedottistico e fognario da SMAT (su circa 292 comuni), sono state estrapolate le tre tabelle fondamentali citate in precedenza, sotto forma di foglio elettronico, contraddistinte da righe che rappresentano un intervento e da colonne che apportano un grande numero di informazioni riguardo la rottura. Si riporta di seguito una descrizione delle stesse, con particolare attenzione ai campi descritti da ogni colonna. È importante sottolineare che sono state apportate alcune correzioni tramite il software *Python*, in quanto sono stati riscontrati errori di trascrizione e di posizionamento di alcune informazioni.

5.1 La tabella *Workorder*

È la tabella principale alla quale saranno connesse le successive, è composta da 382865 righe e 67 colonne e contiene i campi riportati qui di seguito:

- *wonum* che identifica mediante un codice numerico il guasto;
- *workorderid*, codice identificatore;
- *description*, descrizione dell'ordine di lavoro;
- *comune*, comune in cui si è verificato il guasto;
- *indirizzo*, indirizzo in cui si è verificato il guasto;
- *numeroangolo*, civico al quale si è verificato il guasto;
- *segnalazione*, codice che identifica la tipologia di segnalazione;
- *status*, descrive lo stato dell'ordine (in corso, completato);
- *reportdate* riporta la data e l'ora dell'ordine di lavoro;
- *costiimpresa* identifica i costi sostenuti ad un'impresa esterna per l'intervento;
- *pericolo* indica la presenza o meno di un pericolo;
- *reparto* indica il centro di servizio acqua;
- *datacontratto* indica la data e l'ora della stipulazione del contratto;
- *historyflag*, cella che segnala la presenza di un record storico;
- *wopriority*, numero tra 0 e 999 indicante l'importanza dell'ordine di lavoro;
- *origrecordid* identifica il record di origine;
- *ripristino* indica se è stata ripristinata la condotta;
- *esitoverifica* descrive l'esito della verifica;
- *dataverifica* descrive la data e l'ora della verifica;
- *failurecode*, classe di guasto di primo livello;
- *problemcode*, classe di guasto di secondo livello;
- *numero mappa*, numero sopralluogo;
- *actstart*, data e ora in cui è iniziato il lavoro;
- *wol4*, data e ora di fine ripristino stradale;
- *actfinish*, data e ora in cui è stato completato il lavoro effettivo;
- *numeropresa,* numero della presa;
- *lunghezza*, lunghezze in considerazione;
- *targstartdate*, data in cui è previsto l'inizio dell'ordine di lavoro;
- *targcompdate*, data per la quale si prevede la fine dell'ordine di lavoro;
- *diametro*, diametro della presa o della condotta;
- *codicemateriale*, codice rappresentativo del materiale;
- *manovra* descrive la presenza o meno di manovre;
- *orestop*, ore di interruzione;
- *numeromanovre*, numero di manovre effettuate;
- *preavviso* indica se c'è stato un preavviso di 24 ore;
- *gelatura* indica se c'è stata una gelatura della condotta;
- *actlabcost*, costo risorse;
- *acttoolcost*, costo attrezzature;
- *actmatcost*, costo materiali;
- *zanomaliapos*, indica la presenza di un'anomalia nel posizionamento dell'asset;

- *znotepos*, eventuali note sull'anomalia.

5.2 La tabella *Vie*

È la tabella contenente, tra le altre informazioni, l'indirizzo del guasto da associare mediante un codice identificatore all'intervento da operare ed è composta da 22779 righe e 9 colonne. Ogni riga di questa tabella non contraddistingue univocamente un guasto, poiché in alcuni casi più ordini di lavoro fanno riferimento alla stessa riga: è per questo motivo che non c'è corrispondenza con il numero di righe della tabella *Workorder*. Si riportano i campi presenti:

- *comune*, comune nel quale il guasto ha avuto luogo;
- *description,* riportato nella forma "comune*via/strada/piazza", è un campo fondamentale per unire le due tabelle descritte fino ad ora;
- *centrofg*, centro di servizio fognature,
- *centroh20,* centro di servizio acqua;
- *circoscrizione;*
- *indirizzo,* indirizzo dove ha avuto luogo il guasto, spesso coincidente con l'indirizzo dal quale è partita la segnalazione;
- *vieid*, codice identificatore.

5.3 La tabella *Woserviceaddress*

Questa tabella contiene le informazioni strettamente connesse alla localizzazione del guasto, dove ogni riga è associata in modo univoco ad un intervento: presenta, infatti, lo stesso numero di righe della *Tabella Workorder* e 32 colonne. Due dei campi più importanti sono quelli contenenti la latitudine e la longitudine, i quali risultano però mancanti nella maggior parte dei casi per tutti gli anni, eccezion fatta per il 2016 che presenta una percentuale di copertura elevata.
Come per le altre tabelle, si riportano i campi contenuti in ogni colonna:

- *wonum*, numero che identifica il guasto;
- *siteid*, identificatore della sede;
- *streetaddress*, indirizzo;
- *formataddress*, indirizzo riportato in modo formattato;
- *city*, città dove è stato segnalato il guasto;
- *regiondistrict*, distretto;
- *postacode*, codice postale;
- *latitudey*, latitudine del guasto (y);

- *longitudex*, longitudine del guasto (x);
- *referencepoint*, punto di riferimento utile per il tecnico a localizzare l'intervento;
- *woserviceaddressid*, codice identificatore.

5.4 Estrazione dei dati di interesse e calcolo delle occorrenze

A partire dalle tabelle precedentemente descritte sono stati estratti i soli campi di interesse per una valutazione statistica. Lo scopo di questo lavoro è valutare l'eventuale relazione tra il numero di rotture delle condotte e le caratteristiche delle stesse. Per tale motivo si farà riferimento esclusivamente alla tipologia di guasto *"fuga condotta*" e verranno eliminati i campi superflui (contenenti ad esempio codici identificatori, campi a compilazione libera e riguardanti lo stato dell'intervento) e caratterizzati da scarse percentuali di dati.
Gli interventi descritti nelle tabelle originarie non fanno esclusivamente riferimento a rotture su condotte. Le tipologie più ricorrenti eseguite da SMAT o da operatori esterni sono:

- *fuga condotta*, circa il 10 % dei casi;
- *ripristino definitivo*, circa il 10%;
- *fuga presa*, il 6.3%;
- *spurgo* (FG), 4.2%;
- *disotturazione* (FG), 3.7%;
- *cambio contatore*, 3.5%;
- *nuova presa contatore*, 3.5%;
- *rifacimento presa*, 2.7%;
- *posizionamento nuovo contatore*, 2.1%;
- *fuga privata*, 2.1%.

I campi selezionati per ogni tabella sono riportati in **Figura 5.1** e le relative occorrenze sono descritte nelle **Tabelle 5.1**, **5.2** e **5.3**.
È possibile notare che molti campi, specie nel caso di c*odicemateriale*, *latitudey* e *longitudex*, mancano di informazioni complete e non si ha la copertura totale delle informazioni per ogni guasto. D'altra parte, il database così composto permette di identificare gli interventi mediante il *wonum* e *l'indirizzo* nella quasi totalità dei casi, ma solo il 17% degli interventi dispone del campo *latitudey* e *longitudex*.

Tabella 5.1. Occorrenze dei campi di interesse della tabella *Workorder*

Workorder					
Campo	Wonum	Comune	Indirizzo	Reportdate	Costiimpresa
%	100	100	100	99.9	99.9
Campo	Pericolo	Diametro	Codicemateriale	Orestop	Problemcode
%	100	35.4	26.5	9.4	83

Tabella 5.2. Occorrenze dei campi di interesse della tabella *Vie*

Vie				
CAMPO	Comune	Indirizzo	Description	Circoscrizione
%	100	99.9	100	98.6

Tabella 6.3. Occorrenze dei campi di interesse della tabella *Woserviceaddress*

Woserviceaddress			
CAMPO	Wonum	Latitudey	Longitudex
%	100	17	17

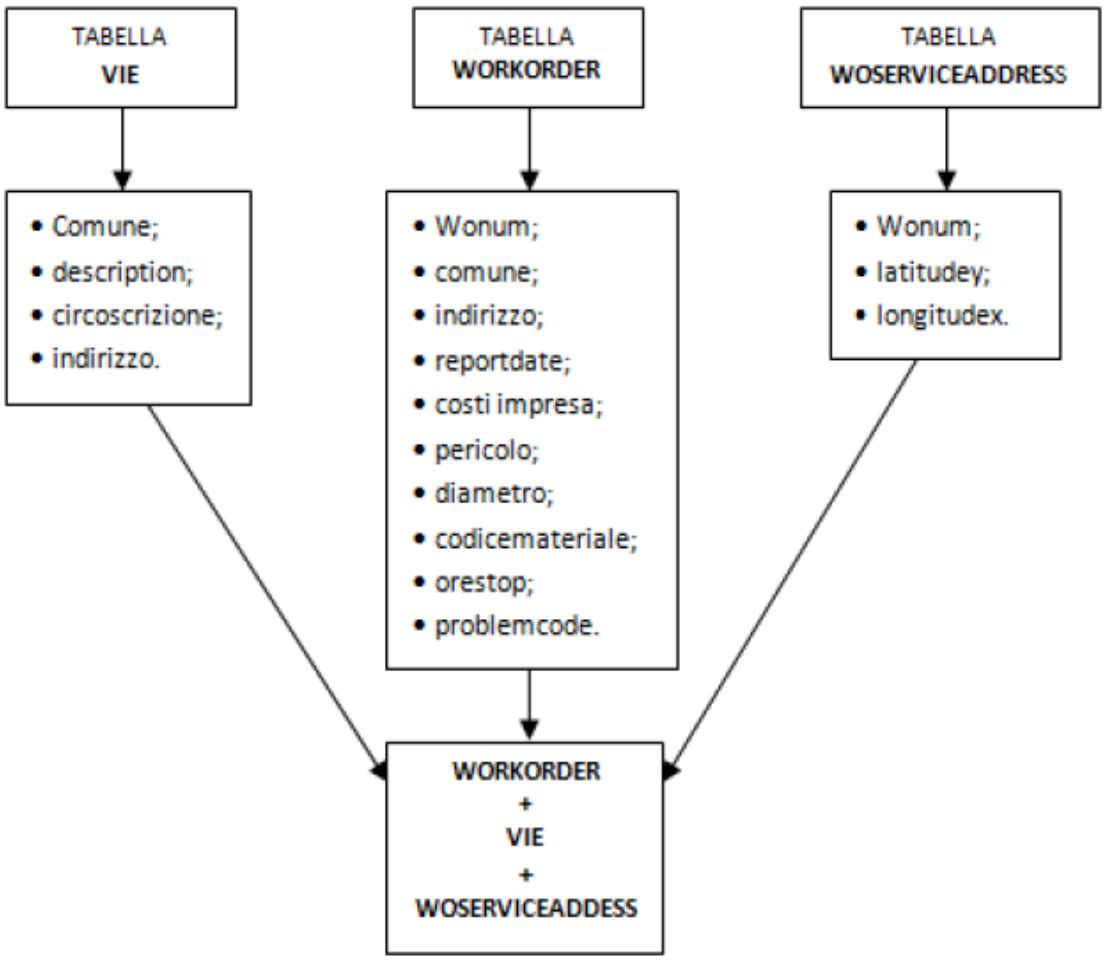


Figura 5.1. Estrazione dei dati e unione delle tabelle

Una volta filtrati i dati, è possibile unire le tre tabelle attraverso dei record identificatori contenuti all'interno delle colonne: il *Wonum*, per le tabelle *Workorder* e *Woserviceaddress*, ed il campo *Indirizzo* per la tabella *Vie*.

5.5 La tabella *Wwvcondottetorino*

La tabella ottenuta dall'unione delle tre tabelle fondamentali è caratterizzata da percentuali eterogenee di dati disponibili in ogni campo. Infatti, è possibile notare che la maggior parte dei campi presenti una copertura totale delle informazioni relative ai guasti, mentre le indicazioni riguardanti la latitudine e la longitudine si attestano intorno al 12%.

Tabella 5.5. Occorrenze della tabella finale inerente gli interventi *Fuga Condotta* a Torino

Unione tabelle *Workorder+ Vie+Woserviceaddress*				
CAMPO	Wonum	Comune	Indirizzo	Reportdate
%	100	100	100	100
CAMPO	Costimpresa	Pericolo	Diametro	Codicemateriale
%	100	100	90.7	87.8
CAMPO	Orestop	Circoscrizione	Latitudey	Longitudex
%	76.5	99.8	12	12

Questi campi sono di fondamentale importanza perché permettono di localizzare in modo preciso il guasto, in modo da poter calcolare, attraverso un modello idraulico della rete sviluppato in ambiente *Epanet*, la pressione in corrispondenza delle condotte dove ha avuto luogo l'intervento di riparazione.
Per aumentare la percentuale di informazioni riguardanti la collocazione geografica, è stato scritto un codice in *Python* che, attraverso una chiave di *Google Maps*, ha fornito la latitudine e la longitudine degli interventi caratterizzati da indirizzo completo di numero civico. Una volta effettuata la georeferenziazione, circa il 95% dei guasti è stato caratterizzato dalle informazioni *latitudey* e *longitudex* ed è stato quindi possibile riportare le rotture sulla rete di Torino tramite l'applicazione *QGIS* (**Figura 5.2**).
In aggiunta alla pressione, è stata integrata l'informazione relativa all'anno di posa delle condotte, attraverso l'utilizzo del database cartografico in possesso dell'ente gestore, contenente tutte le caratteristiche delle condotte.

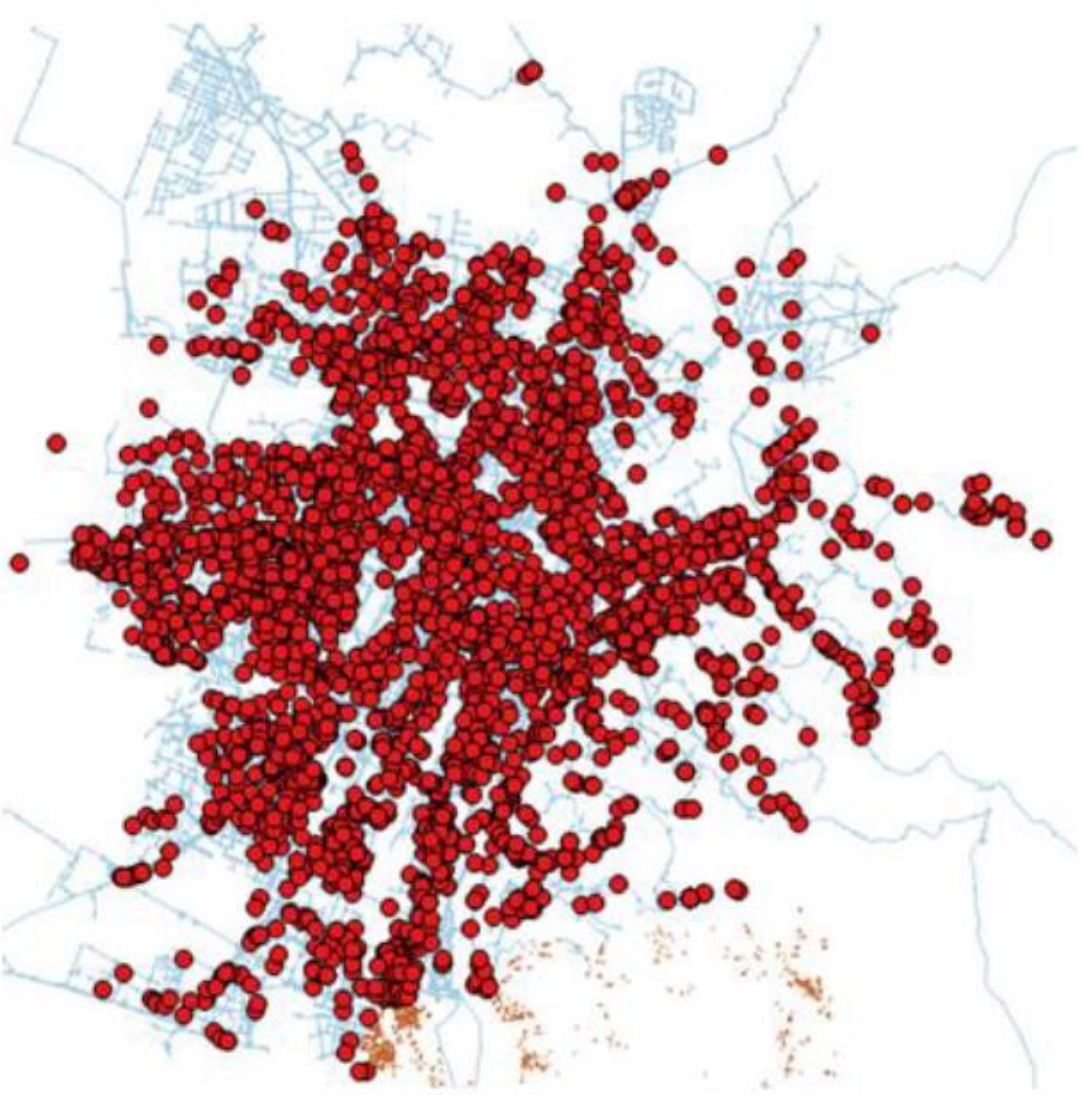


Figura 5.2. Distribuzione delle rotture di tipo *Fuga Condotta* all'interno della rete di Torino

A partire quindi dalla georeferenziazione del guasto, il modello in *Epanet* ed il database cartografico hanno fornito rispettivamente le informazioni sulla pressione e l'anno di posa delle condotte. Questi ultimi non sono contenuti all'interno delle tabelle originarie ma sono contemplati nella versione finale della tabella *Wwvcondottetorino*, che costituisce il punto di partenza per la pre-analisi dei dati utili ai fini statistici.
Questa tabella, nella sua versione finale, presenta gli interventi di tipo *'fuga condotta'* che hanno avuto luogo nel comune di Torino tra il 2006 e il 2016 ed è caratterizzata da 4042 righe e 12 colonne che fanno rispettivamente riferimento ai guasti e ai campi.
Si riporta di seguito un esempio di guasto descritto nella tabella *Wwvcondottetorino*.

Tabella Wwvcondottetorino		
WONUM	COMUNE	INDIRIZZO
33139	TORINO	Corso Regina Margherita 304
PERICOLO	REPORTDATE	DIAMETRO

0	08/08/2006 12:12	150
CODICEMATERIALE	ORESTOP	COSTIMPRESA
GHISA SFEROIDALE	1	468.76
CIRCOSCRIZIONE	LATITUDEY	LONGITUDEX
7	42.08718	7.65295

Tabella 5.6. Esempio di guasto nella tabella *Wwvcondottetorino*

Per una maggiore chiarezza, si riporta di seguito in **Figura 5.3** il diagramma di flusso che illustra le operazioni effettuate in questo capitolo per ottenere la tabella *Wwvcondottetorino*.

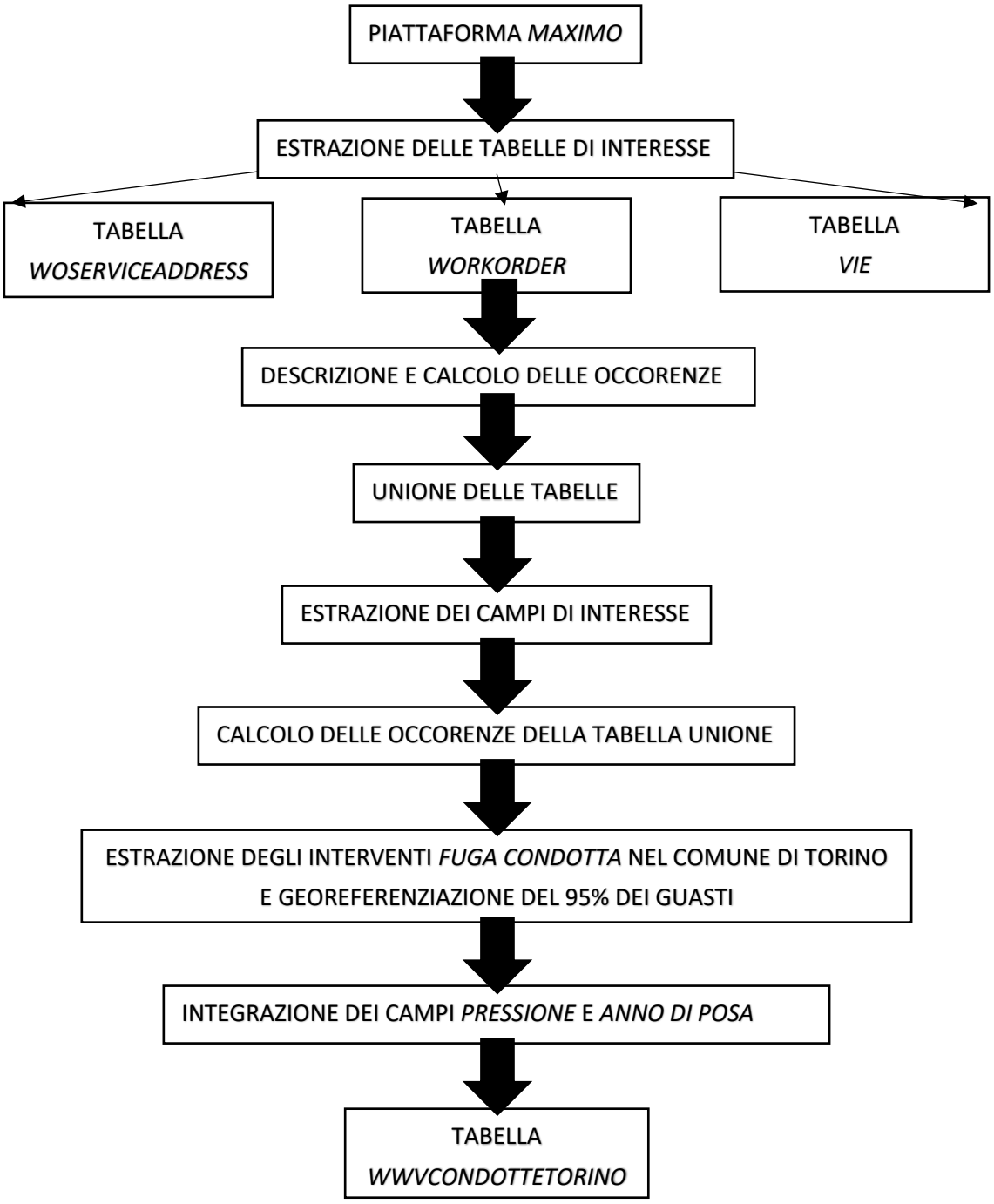


Figura 5.3. Diagramma di flusso raffigurante le operazioni eseguite

5.6 La tabella *ReteTorino*

Il fine di questo lavoro è quello di applicare un modello statistico di regressione ai dati a disposizione per poter prevedere quali condotte siano a rischio di rottura. Affinché ciò avvenga, è necessario che i modelli siano tarati a partire da un'unica tabella contenente tutte le caratteristiche delle condotte, comprensive dell'informazione di rottura o non rottura.
Per comprendere la procedura applicata in seguito per ottenere la tabella in esame, si consideri un'ipotetica rete caratterizzata da 5 condotte, le cui caratteristiche sono riportate in **Tabella 5.7**.

Tabella 5.7. Caratteristiche delle 5 condotte che formano la rete

Condotta	Materiale	Diametro	Anno di posa	Latitudine	Longitudine
1	Pead	50	1990	45.078	7.693
2	Ghisa grigia	100	1980	45.071	7.712
3	Ghisa sferoidale	150	1996	45.076	7.635
4	Eternit	200	1974	45.078	7.661
5	Acciaio	250	1983	45.076	7.621

Si immagini di avere a disposizione inoltre la tabella contenente le rotture di tale rete, riportata in **Tabella 5.8**.

Tabella 5.8. Tabella delle rotture della nella rete

Materiale	Diametro	Anno di posa	Latitudine	Longitudine	Rottura
Pead	50	1990	45.078	7.693	a
Eternit	200	1974	45.078	7.661	b

Per poter applicare dei modelli di regressione sul campione di dati, è necessario unire le due tabelle in esame ed ottenere un'unica tabella contenente le caratteristiche delle condotte e l'informazione riguardante la rottura o la non rottura della stessa. Questo processo associativo è possibile mediante l'utilizzo delle coordinate geografiche latitudine e longitudine associate ad ogni condotta.
Nell'esempio appena descritto, la tabella finale assume la forma riportata in **Tabella 5.9**, nella quale l'ultima colonna definisce una rottura nel caso di valore unitario e una non rottura nel caso di valore nullo.
Si noti che il sistema acquedottistico di Torino risulta, ovviamente, ben più complesso dato il grande numero di condotte che lo compongono.

Tabella 5.9. Tabella contenente tutte le informazioni riguardanti l'intera rete, comprensive dei dati di rottura e non rottura

Condotta	Materiale	Diametro	Anno di posa	Latitudine	Longitudine	Rottura
1	Pead	50	1990	45.078	7.69	1
2	Ghisa grigia	100	1980	450071	7.712	0
3	Ghisa sferoidale	150	1996	450076	7.635	0
4	Eternit	200	1974	45.078	7.661	1
5	Acciaio	250	1983	45.076	7.621	0

Per ottenere la tabella contenente tutte le caratteristiche delle stesse, denominata in seguito tabella *ReteTorino*, sono stati seguiti i seguenti passi:

1. È stato esaminato il database cartografico a disposizione di *SMAT*, contenente le informazioni riguardo le condotte della rete quali diametro, posizione geografica, materiale e anno di posa. Assieme a questo, è stato elaborato in ambiente *Epanet* un modello della rete in grado di fornire i valori di pressione in ogni punto. È stato così disponibile conoscere la totale composizione della rete, integrata alle informazioni riguardanti i carichi nelle condotte.
2. Le condotte con lunghezza inferiore ai 10 metri sono state escluse dall'analisi, in quanto scarsamente significative ai fini della stima del modello di regressione. A seguito di questa scelta, gran parte delle condotte in PEAD sono state escluse dalla tabella *ReteTorino*, poiché caratterizzate da lunghezze molto contenute. Successivamente, la lunghezza di ogni condotta è stata per semplicità arrotondata al multiplo di 10 più vicino.
3. Alle condotte del database cartografico sono state associate quelle della tabella *Wwvcondottetorino* (che contiene esclusivamente informazioni riguardo condotte rotte nel periodo di riferimento 2006-2016) attraverso le coordinate geografiche. Ciò ha permesso di aggiungere l'informazione riguardo la rottura o la non rottura delle condotte.

   È importante sottolineare che non tutti gli interventi contenuti nella tabella *Wwvcondottetorino* dispongono delle informazioni geografiche ed in alcuni casi tali informazioni non combaciano con quelle riportate nel database cartografico. Questo ha portato all'impossibilità di associare il guasto ad alcune condotte ed è probabilmente causato da errori di geolocalizzazione o di trascrizione dell'operatore che ha eseguito l'intervento. Quindi, dopo aver inserito tutti gli interventi contenuti nella tabella *Wwvcondottetorino* sulla mappa della rete, i guasti sono stati associati alle condotte solo nel caso in cui questi distassero meno di 500 metri dalla condotta cui facevano riferimento.
4. Per ogni condotta, nel caso di coincidenza tra le informazioni geografiche riportate nella tabella *Wwvcondottetorino* e il database cartografico, è stata riportata l'informazione di rottura contraddistinta dal flag pari a 1.

5. Si è ottenuta così la tabella *ReteTorino*, nella quale ogni riga caratterizza una condotta.

La tabella appena ottenuta si compone di tutte le condotte facenti parte della rete di Torino e delle informazioni riguardo la loro rottura o non rottura nel periodo di osservazione. Si presenta nella seguente forma:

- 25887 righe che fanno riferimento a condotte prese singolarmente. Se una condotta non ha mai presentato una rottura, è stata riportata in tabella un'unica volta con flag pari a 0. Se la condotta ha presentato un'unica rottura, è stata riportata in tabella una sola volta con flag unitario. Se, invece, la condotta ha presentato più rotture nel periodo di osservazione, questa occupa un numero di righe pari al numero di guasti avvenuti;
- 10 colonne che riportano le caratteristiche di interesse delle condotte. Sono state escluse le righe che non apportano informazioni utili per l'applicazione della regressione;
- di queste 25887 righe, 3881 fanno riferimento a dei guasti.

Si riportano in Tabella **5.10** due righe della tabella *ReteTorino* dove la prima descrive i campi che caratterizzano le condotte e la seconda i valori specifici di una condotta.

Tabella 5.10. Esempio di condotta appartenente alla tabella *ReteTorino*

Tabella *ReteTorino*					
Campo	Condotta	Materiale	Anno di posa	Diametro (mm)	Carico massimo (m)
Condotta	1663	Ghisa Sferoidale	1996	100	46.95
Campo	Lunghezza (m)	Carico medio (m)	Latitudine	Longitudine	Flag
Condotta	10	41.23	49.96	13.94	1

La tabella *ReteTorino*, nella sua versione finale, è caratterizzata da un numero di righe pari a 26998 e da una percentuale di condotte rotte pari al 14.4% (3881 guasti nel periodo di osservazione 2006-2016). L'analisi dei dati contenuti sarà di fondamentale importanza per la validazione dei risultati conseguiti con l'applicazione dei modelli di regressione. Tale analisi sarà presentata nel capitolo successivo.
Nonostante questa tabella rappresenti tutta la rete di Torino, nella sua forma non risulta adeguata all'applicazione dei modelli di regressione: infatti, tali modelli, per una stima dei parametri, hanno bisogno della totale copertura dei campi che riportano le informazioni sul materiale, diametro, anno di posa, carico massimo e lunghezza di ogni condotta.
La tabella, invece, è così caratterizzata:

- tutte le condotte dispongono dell'informazione riguardante il materiale;

- tutte le condotte dispongono dell'informazione riguardante il diametro;
- solo 6150 condotte su 26998 riportano l'informazione relativa all'anno di posa. La copertura di questo campo è pari al 22.78%;
- 25800 condotte su 26998 presentano indicazioni sul carico massimo. Tra queste si presentano valori di carico pari a 0 o negativi. Questa situazione è inattuabile e tali informazioni non potrebbero essere utilizzate. In definitiva, la copertura di questo campo ammonta al 95.56% delle condotte presenti in tabella;
- tutte le condotte presentano l'informazione riguardante la lunghezza.

Come già detto in precedenza, i modelli di regressione devono fare affidamento su dati completi, privi di errori e con totale copertura dei campi che fanno riferimento alle variabili indipendenti che caratterizzeranno la regressione. Per tale ragione, è stato necessario elaborare una nuova tabella, denominata tabella *Esatta* e presentata nel prossimo paragrafo.

5.7 La tabella *Esatta*

A partire dalla tabella *ReteTorino* sono state eliminate le condotte con informazioni mancanti nei campi materiale, diametro, anno di posa, carico massimo e lunghezza. Infatti, i modelli statistici di regressione polinomiale e logistico utilizzeranno i campi appena menzionati per la stima dei parametri e, di conseguenza, non è possibile prendere in considerazione eventuali campi con informazioni parziali.
Successivamente, ad ogni materiale presente nella tabella è stato assegnato un numero, come esposto precedentemente:

- Acciaio, numero 1;
- Cemento armato, numero 2;
- Eternit, numero 4;
- Ferro, numero 5;
- Ghisa grigia, numero 6;
- Ghisa sferoidale, numero 7;
- PEAD, numero 8;
- Piombo, numero 10;
- PVC, numero 11.

In aggiunta, mediante un controllo incrociato tra la tabella *ReteTorino* e la tabella delle rotture *Wwvcondottetorino*, sono stati riscontrati dei problemi di non coincidenza tra le caratteristiche delle condotte (quali diametro e materiale) presenti nella tabella *ReteTorino* e quelle presenti nella tabella *Wwvcondottetorino*.
Come esposto in precedenza, la geolocalizzazione delle condotte è stata effettuata facendo affidamento sugli indirizzi riportati dagli operatori addetti all'intervento. In

molti casi, tali indirizzi fanno riferimento all'indirizzo più vicino alla condotta e non alla posizione della stessa, con conseguente scarsa precisione. Un' ulteriore causa scatenante potrebbe essere la presenza di errori nel database *Maximo*: alcuni interventi di tipo *fuga presa* potrebbero essere stati erroneamente classificati come interventi *fuga condotta.*

Per tale motivo, per localizzare la rottura, è stato imposto che il guasto si posizionasse sulla condotta più vicina con lo stesso diametro e materiale. In questo modo sono state incrociate la tabella *Wwvcondottetorino* e il database cartografico a disposizione dell'ente gestore per ottenere la tabella *ReteTorino*.

Incrociando, quindi, i medesimi dati di rottura contenuti nelle tabelle *ReteTorino* e *Wwvcondottetorino*, si sarebbe dovuta ottenere una perfetta coincidenza tra il numero assegnato ad ogni rottura e le caratteristiche di ogni condotta quali diametro e materiale. In molti casi, ciò non è avvenuto e, per tale motivo, sono stati eliminati tutti quei guasti aventi caratteristiche non coincidenti con quelle delle condotte associate nella tabella *Wwvcondottetorino.*

In definitiva, a partire dalla tabella *ReteTorino*, una volta eliminate le condotte con campi parzialmente completi e i guasti che presentavano una non perfetta coincidenza con le caratteristiche delle condotte associate nella tabella *Wwvcondottetorino*, è stata ottenuta la tabella *Esatta*. Questa rappresenta la base sulla quale i modelli di regressione saranno tarati ed è così composta:

- 5855 righe che fanno riferimento alle condotte filtrate secondo le modalità appena descritte;
- di queste 5855 righe, 106 fanno riferimento a condotte con flag pari a 1 e 5749 a condotte non guaste (1.8% di rotture);

Inoltre, 5749 condotte sono caratterizzate da:

- la totale disponibilità di informazioni riguardanti materiale, diametro, anno di posa, carico massimo, lunghezza;
- flag pari a 0.

Per concludere, è necessario fornire ulteriori informazioni riguardo un importante aspetto della tabella appena ricavata. Come detto in precedenza, la tabella *ReteTorino* nella sua versione iniziale presenta il 14.4% di rotture, mentre la tabella *Esatta* solo l'1.8%. Questo calo percentuale è dovuto all'eliminazione delle condotte con incomplete informazioni e, soprattutto, è attribuibile all'eliminazione delle condotte che non presentavano coincidenza di caratteristiche tra la tabella *ReteTorino* e la tabella *Wwvcondottetorino*. Infatti, quest'ultimo processo di selezione viene effettuato esclusivamente sulle condotte con flag pari a 1 ed ha abbattuto notevolmente il numero di rotture presenti nella tabella *ReteTorino*.

5.8 Conclusioni

In questo capitolo sono state descritte le fonti dei dati dalle quali è stato attinto: in particolare, si è fatto affidamento alle tre tabelle fondamentali *Vie*, *Workorder* e *Woserviceaddress* fornite dal responsabile SMAT in formato elettronico e contenute nella piattaforma *Maximo.*

A seguito dell'analisi di questi dati, sono stati estrapolati dalle tabelle sopracitate solo i campi di interesse ed è stata generata un'unica tabella sulla quale è stato effettuato il calcolo delle occorrenze.

A partire da questa tabella sono stati estratti i soli interventi di tipo *Fuga Condotta* che hanno avuto luogo nel comune di Torino tra il 2006 e il 2016, ottenendo così la tabell*a Wwvcondottetorino*. Attraverso un codice in *Python* è stato possibile aggiungere le informazioni di latitudine e longitudine, fino a raggiungere una copertura del 95% totale dei guasti in questi campi. Infine, mediante la georeferenziazione, l'utilizzo del database cartografico e del modello idraulico in *Epanet*, sono state aggiunte le informazioni relative agli anni di posa e alla pressione delle condotte che hanno riscontrato dei guasti. Successivamente, attraverso la geolocalizzazione è stata aggiunta alle condotte del database cartografico l'informazione riguardante la rottura o non rottura. L'intersezione tra la tabella *Wwvcondottetorino* e il database cartografico ha dato così luce alla tabella *ReteTorino*, caratterizzata da 26998 righe e 10 colonne.

La tabella *ReteTorino* è però contraddistinta da numerose informazioni mancanti e da alcuni errori di localizzazione delle condotte. Di conseguenza, sono state prese in considerazione solo le condotte caratterizzate da totale copertura dei dati per i campi materiale, anno di posa, diametro, carico massimo e lunghezza. Inoltre, sono stati eliminati i guasti caratterizzati da materiale e diametro non coincidenti con le caratteristiche riportate nella tabella *Wwvcondottetorino*. La tabella così ottenuta è stata denominata tabella *Esatta* e presenta 5855 righe e lo stesso numero di colonne della tabella *ReteTorino*. Il passaggio da questa tabella alla tabella *Esatta* ha portato un decremento della percentuale di rotture dal 14.4% all'1.8%.

Nel successivo capitolo verranno analizzate le tabelle fondamentali appena ricavate: la tabella *Wwvcondottetorino*, la tabella *ReteTorino* e la tabella *Esatta*. Sarà inoltre presentata la definizione di tasso di fallanza, uno strumento di primaria importanza nel definire la vulnerabilità di classi di condotte.

Si riporta, infine, in **Figura 5.4** il diagramma relativo alle operazioni intermedie effettuate nel passaggio dalla tabella *ReteTorino* alla tabella *Esatta.*

Figura 5.4. Diagramma di flusso delle operazioni eseguite per passare dalla tabella *ReteTorino* alla tabella *Esatta*

Capitolo 6

Analisi delle tabelle fondamentali

Nel capitolo precedente sono state descritte le fonti dei dati e le modalità con le quali sono state ottenute le tabelle *Wwvcondottetorino*, *ReteTorino* ed *Esatta*.
Nello specifico, la tabella *Wwvcondottetorino* è contraddistinta da 4042 righe e 12 colonne, caratterizzante i guasti del tipo '*fuga condotta*' avvenuti nel solo comune di Torino negli anni che vanno dal 2006 al 2016.
I dati al suo interno forniscono informazioni riguardo le condotte in esame quali lunghezze, diametri, materiali, anni di posa, pressioni e altre caratteristiche che possono influire sulla probabilità di accadimento di una rottura.
La tabella *ReteTorino* è formata da 26998 righe e 10 colonne e contiene tutte le condotte della rete di Torino con l'integrazione dell'informazione riguardante la rottura o non rottura delle stesse.
Infine, la tabella *Esatta* è ottenuta da una filtrazione dei dati contenuti nella tabella *ReteTorino* ed è formata da 5855 righe e 10 colonne.
Nel seguente capitolo verrà analizzata la composizione di queste tabelle al fine di determinare eventuali correlazioni tra il numero di rotture, le caratteristiche riportate nei campi della tabella e le relazioni che possono sussistere tra i campi stessi.

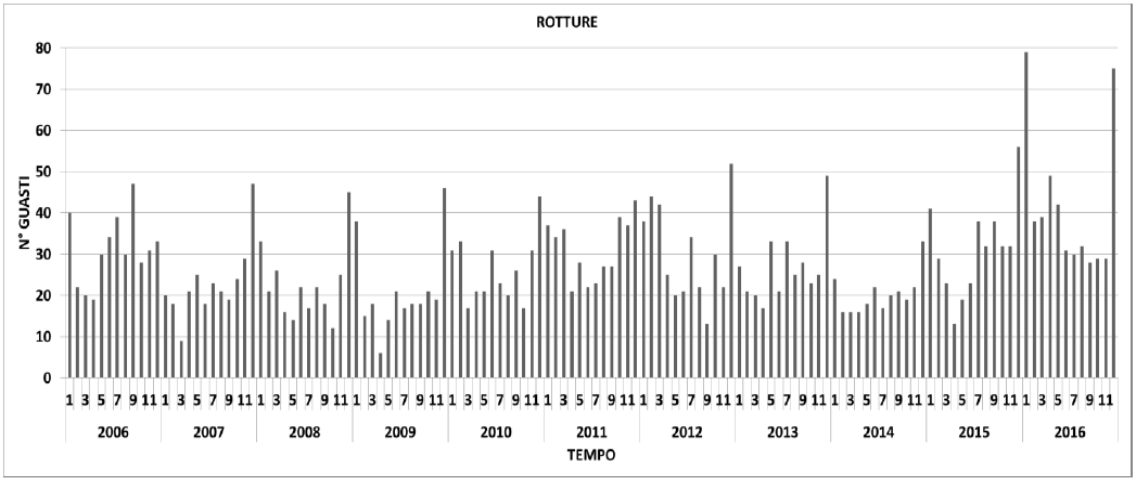
6.1 Calcolo delle occorrenze della tabella *Wwvcondottetorino*

Il calcolo delle occorrenze ha lo scopo di determinare il numero di volte che un determinato valore si presenta in un intervallo di valori. Nello specifico, è stato effettuato questo calcolo per ogni valore presente nei campi della tabella *Wwvcondottetorino*. È importante precisare che il numero di rotture per un determinato campo non può essere un indice affidabile della vulnerabilità di una determinata classe e non si può prescindere dalla lunghezza delle condotte in esame: per questo motivo, nei capitoli successivi verrà presentato il tasso di fallanza, cioè il numero di rotture normalizzato rispetto alla lunghezza delle condotte.
Di seguito verranno riportate in formato tabellare le occorrenze riguardanti i campi *Repordate*, *Diametro*, *Materiale*, O*restop*, *Pericolo*, *Costiimpresa*, *Pressioni* e *Anno di posa* delle condotte.

Campo *Reportdate*

Questo campo fornisce indicazioni riguardo la data e l'ora in cui ha avuto luogo la segnalazione e caratterizza tutti gli interventi riportati nella tabella *Wwvcondottetorino* (numero totale di dati pari a 4042).
La distribuzione delle rotture nei vari mesi degli anni dal 2006 al 2016 è riportata in **Figura 6.1**: è possibile notare che il numero più elevato di guasti si presenta nei mesi più freddi e severi, quali dicembre e gennaio, e l'andamento complessivo delle rotture rispetta un trend in maniera quasi ciclica durante l'anno, con valori comparabili per lo stesso mese nei diversi anni.

Figura 6.1. Distribuzione mensile delle rotture avvenute negli anni dal 2006 al 2016



Allo stesso modo, è possibile prendere in esame il numero totale di rotture avvenute nei diversi anni, come riportato in **Tabella 7.1**. Il numero medio annuale di rotture si attesta intorno a 333, ma l'anno 2016 ha visto superare di gran lunga tale valore.

Tabella 6.1. Distribuzione delle occorrenze totali negli anni

Anno	2006	2007	2008	2009	2010	2011
Occorrenze	373	274	271	251	315	374
Percentuale (%)	10.2	7.5	7.4	6.9	8.6	10.2

Anno	2012	2013	2014	2015	2016
Occorrenze	363	322	244	376	501
Percentuale (%)	9.9	8.8	6.7	10.3	13.7

Le occorrenze delle rotture per ogni mese dell'anno sono invece riportate in **Tabella 6.2** e graficamente in **Figura 6.2**. Queste confermano quanto visto in **Figura 6.1**: il 14% delle rotture totali è avvenuto nel mese di dicembre a fronte del 5.54% del mese di aprile. È possibile affermare che i mesi invernali presentano un numero più elevato di rotture rispetto a quelli estivi, ma, come detto in precedenza, questo dato ha un ruolo marginale se non integrato con i valori riscontrati per i singoli materiali e le rispettive lunghezze. In

seguito si mostrerà come alcuni materiali siano fortemente sensibili alle basse temperature, mentre altri risultano vulnerabili alle alte temperature o totalmente insensibili alle condizioni termiche esterne.

Tabella 6.2. Occorrenze delle rotture nei diversi mesi dell'anno

Mese	1	2	3	4	5	6
Occorrenze	485	334	296	224	268	286
Percentuale (%)	12	8.26	7.32	5.54	6.63	7.08

Mese	7	8	9	10	11	12
Occorrenze	316	306	316	296	349	566
Percentuale (%)	7.82	7.57	7.82	7.32	8.63	14

Infine si riportano le occorrenze per i giorni della settimana: la **Tabella 6.3** mostra le occorrenze totali nei giorni della settimana. Un numero inferiore di rotture nei giorni di sabato e domenica potrebbe essere giustificato da minori variazioni di pressione e di carico veicolare o, più semplicemente, dal fatto che le segnalazioni ricevute nei weekend vengano rimandate al lunedì successivo.

Tabella 6.3. Occorrenze delle rotture nei diversi giorni della settimana

Giorno	1	2	3	4	5	6	7
Occorrenze	693	729	662	717	632	342	267
Percentuale (%)	17.14	18.04	16.38	17.74	15.64	8.46	6.61

Campo *Diametro*

Si procede ora nel calcolo delle occorrenze per il campo di riferimento dei diametri: come visto nel capitolo precedente, i dati relativi a questa informazione sono presenti nel 91% dei casi e, di conseguenza, 370 interventi sono sprovvisti di tale dato. In aggiunta, il dato sul numero di rotture non è sufficiente per valutare la vulnerabilità di un certo diametro rispetto ad altri. Si considerino ad esempio due condotte di diametri 100 e 150 mm che presentano un numero di rotture rispettivamente pari a 10 e 15. Alla luce di questi dati si può solo affermare che la seconda condotta si è rotta un numero di volte maggiore, ma il calcolo della vulnerabilità deve tenere in conto anche delle lunghezze delle due condotte, per poter valutare un numero di rotture per chilometro di condotta. Si supponga quindi di conoscere le lunghezze delle condotte in esame, rispettivamente 100 e 300 km: con questi dati la prima condotta presenta un numero di rotture per km pari a 0.1, mentre la seconda 0.05 guasti per unità di lunghezza. Nonostante il numero di rotture sia minore, la prima condotta presenta una maggiore vulnerabilità alla rottura rispetto alla seconda.
Si riportano in **Tabella 6.4** le occorrenze per le diverse fasce di diametri.

Tabella 6.4. Occorrenze delle rotture per le differenti fasce di diametri

Diametro (mm)	13-25	30-40	50-60	60-90
Occorrenze	79	47	867	573
Percentuale (%)	2.16	1.28	23.66	15.64

Diametro (mm)	100-150	175-300	350-600	700-1000
Occorrenze	1886	121	81	10
Percentuale (%)	51.47	3.3	2.21	0.27

A partire dal database cartografico descritto nel capitolo precedente, sono stati calcolati i chilometri di condotte caratterizzati da una determinata fascia di diametri. Si riportano i risultati in **Tabella 6.5.** I diametri più frequentemente incontrati appartengono al range tra i 100 e i 150 millimetri.

Tabella 6.5. Lunghezze delle differenti classi di diametri

Diametro (mm)	13-25	30-40	50-60	60-90
Chilometri	2.1	3.18	194.2	191.8
Percentuale (%)	0.123	0.194	11.86	11.74

Diametro (mm)	100-150	175-300	350-600	700-1000
Chilometri	800.4	194.8	215.7	30.7
Percentuale (%)	49.02	11.93	13.21	1.87

Infine, per ogni range di diametri sono state calcolate le occorrenze delle rotture e sono state normalizzate rispetto alla lunghezza delle differenti classi: si ottiene la **Tabella 6.6** che illustra come i diametri più piccoli siano i più vulnerabili. Infatti, le condotte caratterizzate da diametri compresi tra i 13 e i 40 millimetri presentano circa l'83% delle occorrenze totali riscontrate.

Tabella 6.6. Occorrenze normalizzate alle lunghezze complessive delle classi di diametri

Diametro (mm)	13-25	30-40	50-60	60-90
Occorrenze/N° Km	39.27	14.76	4.46	2.98
Percentuale (%)	60.29	22.66	6.85	4.58

Diametro (mm)	100-150	175-300	350-600	700-1000
Occorrenze/N° Km	2.35	0.62	0.37	0.32
Percentuale (%)	3.61	0.95	0.57	0.49

Campo *Materiale*

Questo campo presenta compilazione in 3550 casi con una percentuale di completamento pari all'87%. Per comprendere al meglio la vulnerabilità di un materiale è necessario utilizzare lo stesso approccio descritto per il campo *Diametro*: si valutano,

quindi, le occorrenze per ogni materiale e successivamente si normalizzano rispetto ai chilometri di condotte appartenenti ad una determinata classe di materiale.
Si riporta il numero di occorrenze e la percentuale rispetto al numero totale in **Tabella 6.7**.
Le lunghezze delle rispettive classi di materiale e le relative percentuali rispetto alla lunghezza totale delle condotte della rete sono:

- 4.9 chilometri (0.3%) per *acciaio chameroy*;
- 181 chilometri (11.06%) per *acciaio*;
- 0.7 chilometri (0.04%) per *cemento armato*;
- 54 chilometri (3.29%) per *eternit*:
- 1.8 chilometri (0.11%) per *ferro zincato-nudo;*
- 950.7 chilometri (58.13%) per *ghisa grigia*;
- 417.2 chilometri (25.5%) per *ghisa sferoidale;*
- 8.8 chilometri (0.54%) per *polietilene ad alta densità*;
- 0.06 chilometri (0.003%) per *polietilene a bassa densità;*
- 0.5 chilometri (0.03%) per *piombo;*
- 4 chilometri (0.25%) per *pvc*.

Tabella 6.7. Occorrenze delle rotture per i diversi materiali e percentuale rispetto al numero totale

Materiale	Occorrenze	Percentuale (%)
ACC CHAMEROY	2	0.06
ACCIAIO	216	6.08
ACCIAIO INOX	6	0.17
CLS - CEMENTO	0	0.00
ETERNIT (CEM.AMIANTO)	448	12.62
FERRO ZINCATO-NUDO	10	0.28
GHISA GRIGIA	2436	68.62
GHISA SFEROIDALE	262	7.38
POLIETILENE AD ALTA DENSITA' (PEAD)	107	3.01
POLIETILENE A BASSA DENSITA' (PEBD)	21	0.59
PIOMBO	40	1.13
PVC	1	0.03
VETRO RESINA	1	0.03
TOTALE	3550	100

Si normalizzano le occorrenze rispetto ai chilometri di condotte caratterizzanti ogni materiale e si ottiene la **Tabella 6.8**.
È importante notare che la ghisa grigia è il materiale che presenta il maggior numero di rotture ma, allo stesso tempo, caratterizza il 58.13% delle condotte della rete del comune di Torino. Se si normalizzano le occorrenze rispetto alla lunghezza totale, si riscontra che il materiale più vulnerabile è il polietilene a bassa densità con circa 350

rotture per chilometro di condotta, seguito dal piombo, mentre la ghisa grigia si attesta attorno a 2.56 rotture per chilometro.
È necessario altresì sottolineare che il polietilene a bassa densità ed il piombo sono presenti in percentuale così basse all'interno della rete che non saranno presi in esame da qui in avanti.
Il materiale più performante nei confronti delle rotture risulta essere la ghisa sferoidale, che ha sostituito in tempi recenti la ghisa grigia, caratterizzato da un tasso di rottura pari a 0.63 guasti per chilometro di condotta.

Tabella 6.8. Occorrenze delle rotture per i diversi materiali normalizzate rispetto alle lunghezze

Materiale	Occorrenze/N° km	Percentuale (%)
ACC CHAMEROY	0.41	0.09
ACCIAIO	1.23	0.27
CLS - CEMENTO	0	0.00
ETERNIT (CEM.AMIANTO)	8.3	1.80
FERRO ZINCATO-NUDO	5.56	1.20
GHISA GRIGIA	2.56	0.56
GHISA SFEROIDALE	0.63	0.14
POLIETILENE AD ALTA DENSITA' (PEAD)	12.16	2.64
POLIETILENE A BASSA DENSITA' (PEBD)	350	75.91
PIOMBO	80	17.35
PVC	0.25	0.05
TOTALE	461.09	100.00

Campo *Orestop*

È il campo che descrive la durata dell'interruzione dovuta al guasto e al successivo ripristino. Presenta una copertura dei dati pari al 76% (3094 valori disponibili). Come fatto in precedenza, si riportano le occorrenze delle differenti fasce di età presenti in questo campo. La **Tabella 6.9** riporta che la durata più frequente di interruzione si aggira tra una e cinque ore.

Tabella 6.9. Occorrenze delle rotture per le diverse fasce di durata dell'interruzione di servizio

Orestop (h)	0	0.25-0.75	1	1.05-5
Occorrenze	1	27	1332	1660
Percentuale (%)	0.03	0.87	43.05	53.65

Orestop (h)	5.5-10	11.0-20	20-100
Occorrenze	54	11	9
Percentuale (%)	1.75	0.36	0.29

Campo *Pericolo*

È un campo di minore importanza, esaminato per capire se, a seguito del guasto, ci sia stata una situazione di pericolo. Ciò ha avuto luogo in 4 casi su 4042 e per questo motivo non verrà più preso in considerazione.

Campo *Costiimpresa*

È il campo che quantifica i costi sostenuti da SMAT per gli interventi sulle condotte, a seguito di un guasto. La copertura di questi dati è totale ma, nel caso di lavorazione mediante operatore interno all'azienda, l'ammontare riportato in tabella è pari a 0. Si riportano le fasce di spesa e le relative occorrenze.

Tabella 6.10. Occorrenze delle rotture per le diverse fasce di spesa sostenuta

Costiimpresa (€)	0	100-300	300-500	500-100
Occorrenze	1520	86	795	1110
Percentuale (%)	37.6	2.13	19.67	27.47

Costiimpresa (€)	1000-2000	2000-5000	5000-10000	>10000
Occorrenze	355	150	18	8
Percentuale (%)	8.78	3.71	0.45	0.2

Campo *Pressione*

Come descritto nel capitolo precedente, questo campo è stato integrato grazie alla georeferenziazione dei guasti e all'utilizzo successivo di un modello idraulico in *Epanet* che permette di simulare il funzionamento idraulico della rete in esame. Una volta nota la posizione del guasto, è stato possibile ricavare il carico in metri, espresso come il rapporto tra pressione e peso specifico dell'acqua. A partire quindi dalla pressione massima (p_{max}), la pressione media (p_{media}) e l'intervallo di pressione tra i valori massimo e minimo (Δp), sono stati ricavati i carichi massimi, medi e gli intervalli di carico. I dati a disposizione in questo campo coprono il 91% dei guasti (3707 su 4042) a causa dell'assenza di indirizzo completo per il 9% degli interventi. In assenza delle informazioni riguardanti la latitudine e la longitudine, è stato impossibile localizzare le rotture nel modello idraulico della rete per poterne calcolare i valori di carico. Come per i precedenti campi, sono state calcolate le occorrenze per le differenti fasce di carichi (**Tabella 6.11**, **6.12** e **6.13**). Il numero maggiore di rotture si presenta in corrispondenza di condotte con carichi massimi tra i 30 e i 60 metri e per intervalli di carico tra i 10 e i 15 metri.

Tabella 6.11. Occorrenze delle rotture per le fasce di carichi massimi

Carichi massimi (m)	3-10	15-30	30-40	40-50	50-60
Occorrenze	17	20	548	1135	1188
Percentuale (%)	0.46	0.54	14.79	30.62	32.05

Carichi massimi (m)	60-70	70-80	80-100	100-150	150-300
Occorrenze	549	41	31	67	93
Percentuale (%)	15.28	1.11	0.84	1.81	2.51

Tabella 6.12. Occorrenze delle rotture per le fasce di carichi medi

Carichi medi (m)	3-10	15-30	30-40	40-50	50-60
Occorrenze	16	264	782	1470	828
Percentuale (%)	0.43	7.73	21.11	39.65	22.32

Carichi medi (m)	60-70	70-80	80-100	100-150	150-300
Occorrenze	145	18	31	76	77
Percentuale (%)	3.91	0.49	0.84	2.05	2.08

Tabella 6.13. Occorrenze delle rotture per le fasce di intervalli di carico

Intervalli di carico (m)	0-5	5-10	10-15	15-20	20-50	50-100
Occorrenze	260	443	2487	322	183	12
Percentuale (%)	7.02	11.95	67.11	8.66	4.94	0.32

Campo *Anno di posa delle condotte*

Così come quello precedente, questo campo non fa parte dei dati originariamente a disposizione, ma è stato integrato grazie alla geolocalizzazione dei guasti e all'utilizzo del database cartografico. Nonostante questo, solo il 9% dei guasti contenuti nella tabella *Wwvcondottetorino* presenta questa informazione. Inoltre, non si conosce il criterio mediante il quale questo dato è stato inserito nel database.

Nonostante queste informazioni frammentarie, l'anno di posa può fortemente influenzare il deterioramento delle condotte ed è importante includere questo campo nell'analisi del modello probabilistico.

Si riportano di seguito le occorrenze calcolate per diverse fasce di anni: la **Tabella 6.14** mostra che il 33% dei guasti ha avuto luogo in condotte posate nel decennio 1990-2000, simile al 37% riscontrano nel decennio successivo. Le restanti fasce presentano valori inferiori che non superano mai il 10% delle occorrenze totali.

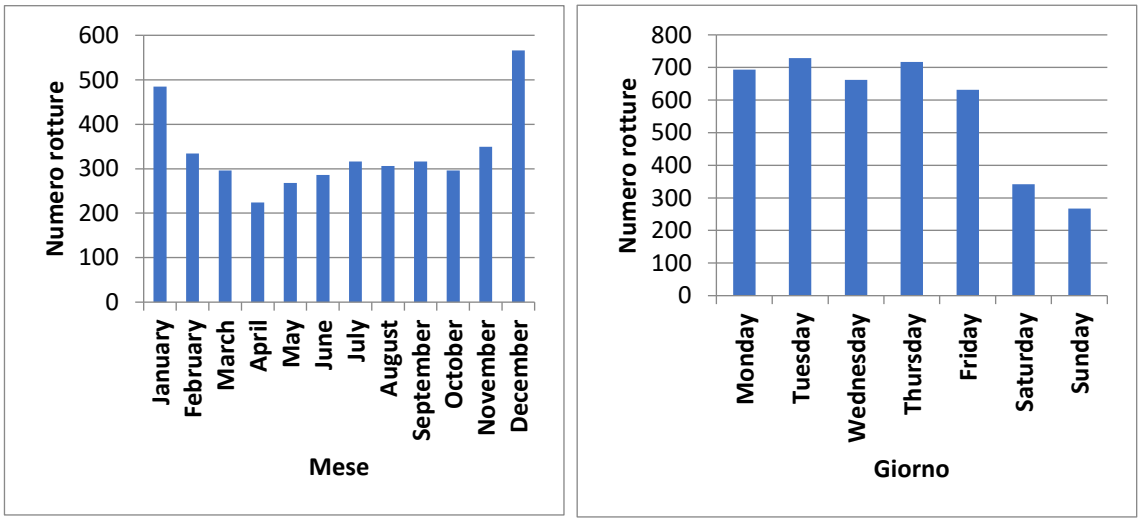
Tabella 6.14. Occorrenze delle rotture per le diverse fasce di età

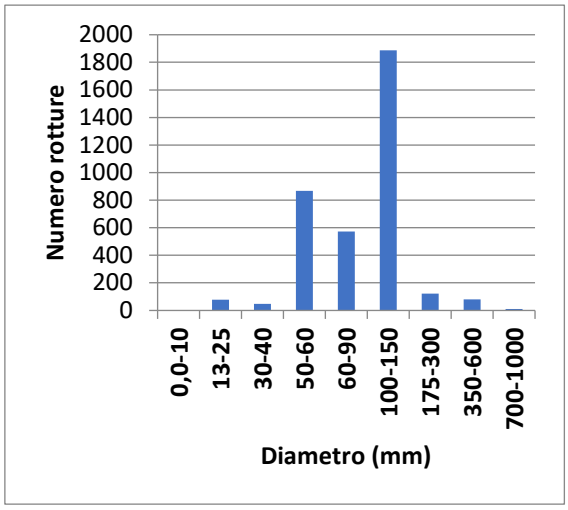
Anno di posa	1871 1900	1900-1930	1930-1960	1960-1990	1990-2000	2000 2005	2005 2010	2010 2015
Occorrenze	5	26	21	26	122	67	68	28
Percentuale (%)	1.38	7.16	5.79	7.16	33.61	18.46	18.73	7.71

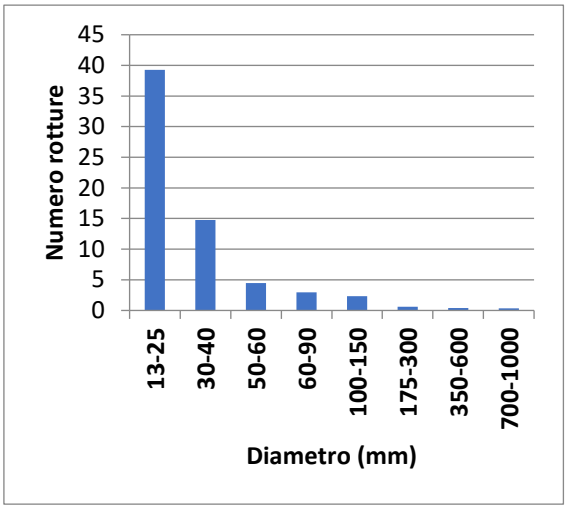
Si ripotano graficamente, infine, le distribuzioni delle occorrenze di tutti i campi esaminati fino a questo punto per una migliore resa grafica (**Figura 6.2**). In ordine, raffigurano il numero di occorrenze per diversi mesi dell'anno, giorni della settimana, nelle differenti fasce di diametri (e relative occorrenze normalizzate rispetto alla lunghezza), per ogni materiale (e relative occorrenze normalizzate rispetto alla lunghezza), per le diverse ore di stop, costi, carichi massimi, medi, intervalli di carico e anni di posa.

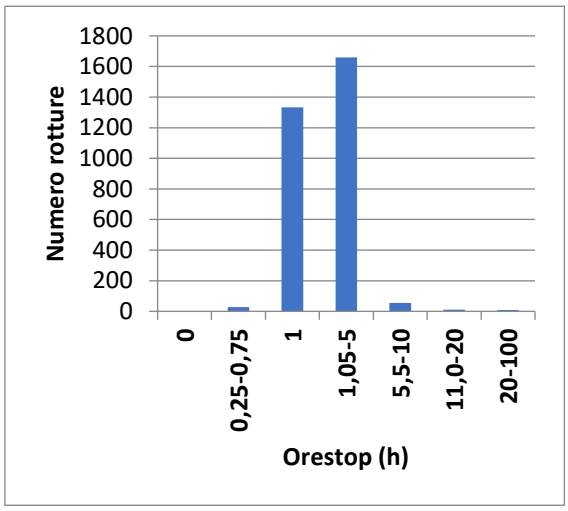
Osservando con attenzione la distribuzione temporale delle occorrenze, sembra sussistere una relazione tra il numero di rotture e la temperatura esterna. Infatti, si riscontrano dei picchi in corrispondenza dei mesi più freddi. Per questo motivo, di seguito verranno calcolate le occorrenze nei diversi mesi dell'anno per ogni materiale, per valutare l'eventuale suscettibilità alla temperatura.

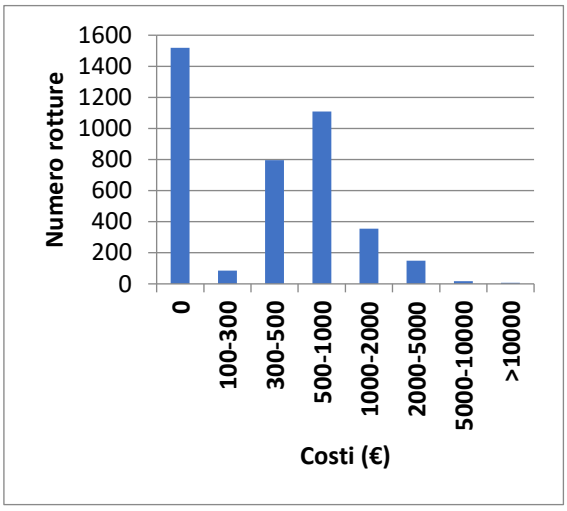
La stessa suddivisione è stata adottata per il calcolo delle occorrenze per fasce di temperature medie mensili, ricavate dal sito dell'*Arpa Piemonte*.

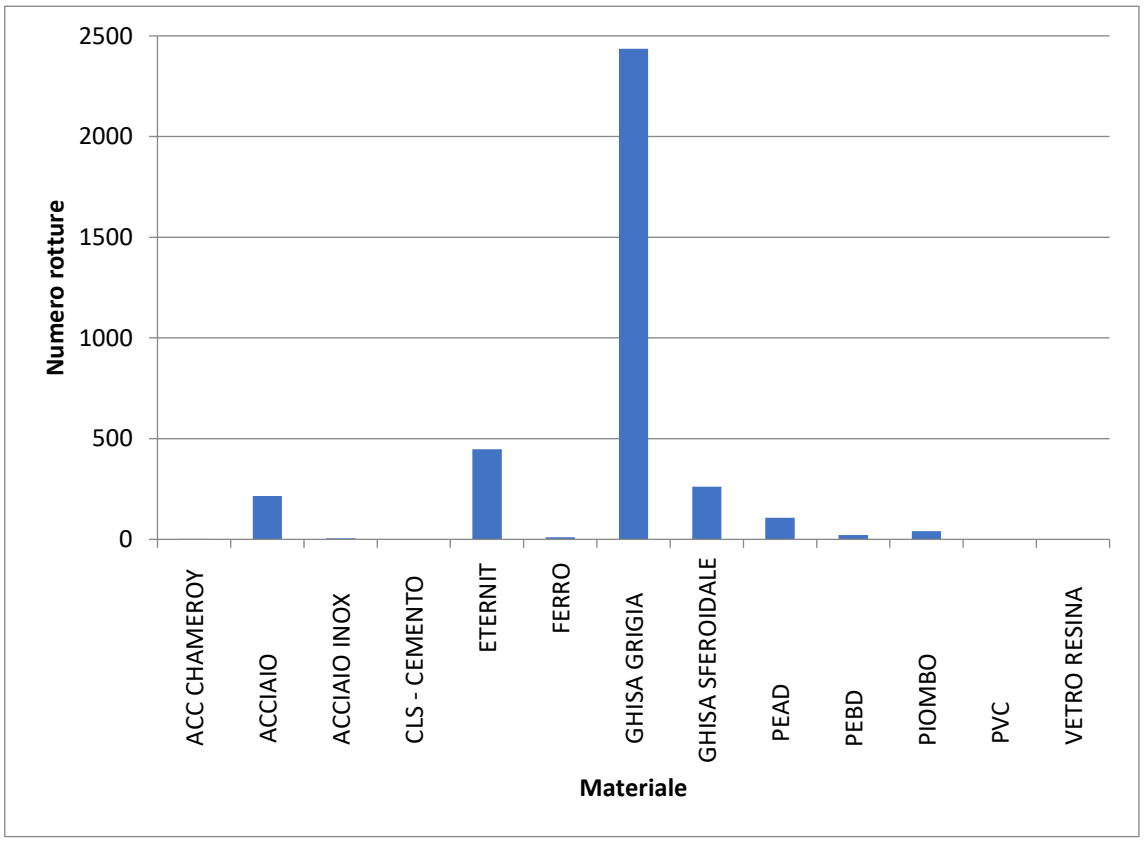



Numero rotture
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0,0-10
13-25
30-40
50-60
60-90
100-150
175-300
350-600
700-1000
Diametro (mm)


Numero rotture
45
40
35
30
25
20
15
10
5
0
13-25
30-40
50-60
60-90
100-150
175-300
350-600
700-1000
Diametro (mm)


Numero rotture
1800
1600
1400
1200
1000
800
600
400
200
0
0
0,25-0,75
1
1,05-5
5,5-10
11,0-20
20-100
Orestop (h)


Numero rotture
1600
1400
1200
1000
800
600
400
200
0
0
100-300
300-500
500-1000
1000-2000
2000-5000
5000-10000
>10000
Costi (€)


Numero rotture
2500
2000
1500
1000
500
0
ACC CHAMEROY
ACCIAIO
ACCIAIO INOX
CLS - CEMENTO
ETERNIT
FERRO
GHISA GRIGIA
GHISA SFEROIDALE
PEAD
PEBD
PIOMBO
PVC
VETRO RESINA
Materiale

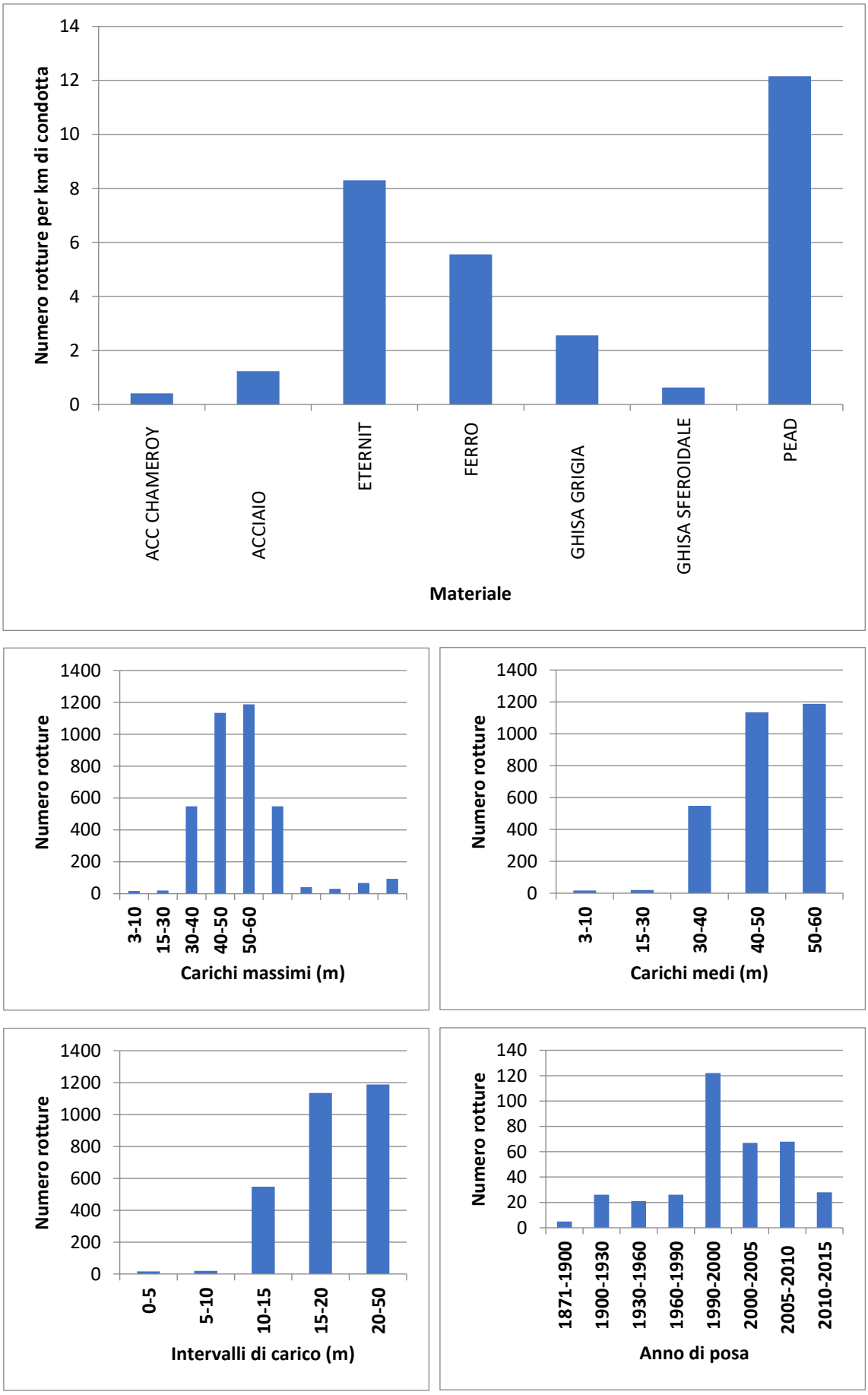


Figura 6.2. Distribuzioni grafiche delle occorrenze per tutti i campi esaminati

Calcolo delle occorrenze in relazione alle temperature medie mensili

È importante confrontare l'andamento delle temperature ed il relativo numero di rotture associato ad ogni mese dell'anno. Questo tipo di analisi mostra chiaramente che il maggior numero di rotture avviene nei mesi invernali ma, per poter trarre delle valide conclusioni, è necessario valutare le occorrenze mensili per ogni materiale preso singolarmente.

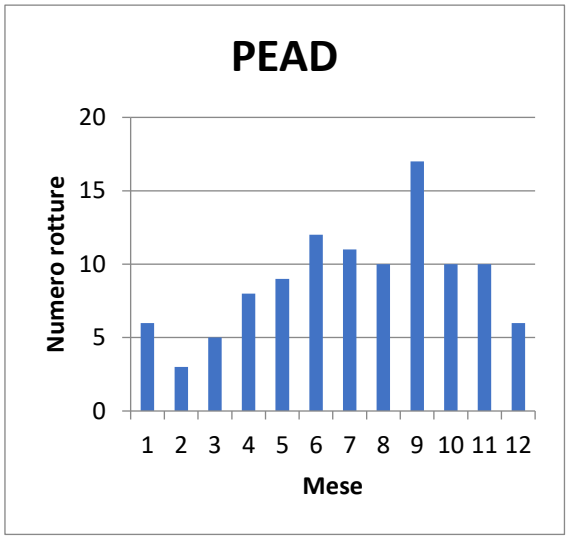
Calcolo delle occorrenze dei singoli materiali nei mesi dell'anno

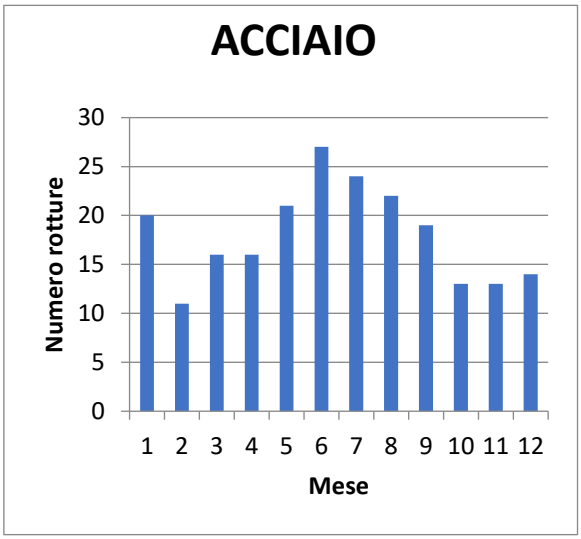
In analogia con quanto fatto per il calcolo delle occorrenze nel campo *Reportdate*, sono stati esaminati singolarmente tutti i materiali che compongono per oltre il 5% la rete di Torino, al fine di studiare la distribuzione temporale delle rotture durante i mesi dell'anno.
Si riportano i valori ricavati in **Tabella 6.15** e graficamente in **Figura 6.3**.

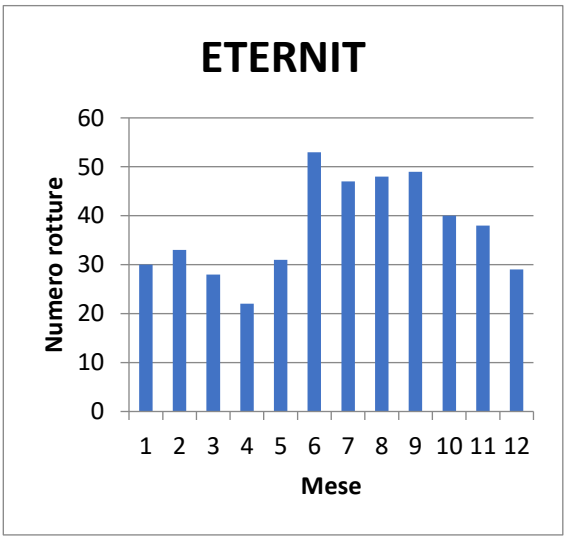
Tabella 6.15. Calcolo delle occorrenze nei diversi mesi per ogni materiale

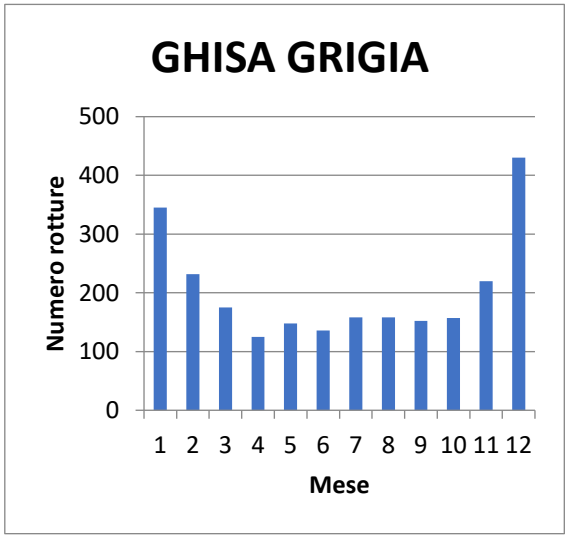
	Mese											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Acciaio	20	11	16	16	21	27	24	22	19	13	13	14
Percentuale (%)	9.26	5.09	7.41	7.41	9.72	12.5	11.11	10.19	8.8	6.02	6.02	6.48
Eternit	30	33	28	22	31	53	47	48	49	40	38	29
Percentuale (%)	6.70	7.37	6.25	4.91	6.92	11.83	10.49	10.71	10.94	8.93	8.48	6.47
Ghisa grigia	345	232	175	125	148	136	158	158	152	157	220	430
Percentuale (%)	14.16	9.52	7.18	5.13	6.08	5.58	6.49	6.49	6.24	6.44	9.03	17.65
Ghisa sferoidale	23	12	16	17	24	20	23	26	33	15	22	31
Percentuale (%)	8.78	4.58	6.11	6.49	9.16	7.63	8.78	9.92	12.60	5.73	8.40	11.83
PEAD	6	3	5	8	9	12	11	10	17	10	10	6
Percentuale (%)	5.61	2.80	4.67	7.48	8.41	11.21	10.28	9.35	15.89	9.35	9.35	5.61

Materiali come acciaio ed eternit presentano dei picchi in corrispondenza dei mesi più caldi (giugno e luglio), mentre il polietilene ad alta densità presenta le maggiori rotture nel mese di settembre. La ghisa sferoidale, invece, non mostra grandi differenze di valori nei diversi mesi, segno della sua non suscettibilità alle differenti temperature. La ghisa grigia, infine, risulta più vulnerabile nei mesi più freddi. Risulta anche il materiale più diffuso all'interno della rete e per questo motivo, nel calcolo delle occorrenze effettuato senza distinzioni di materiali, si è riscontrato il maggior numero di rotture nei mesi di dicembre e gennaio.









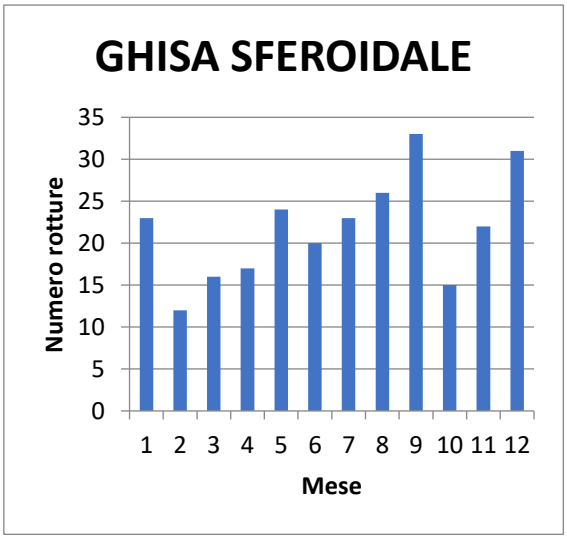


Figura 6.3. Distribuzione delle occorrenze delle rotture nei mesi dell'anno per ogni materiale

Per un reale confronto tra i dati, le occorrenze sono state normalizzate rispetto alla lunghezza totale delle condotte appartenenti ad una determinata classe di materiale. Si rappresentano i risultati in **Figura 6.4**.

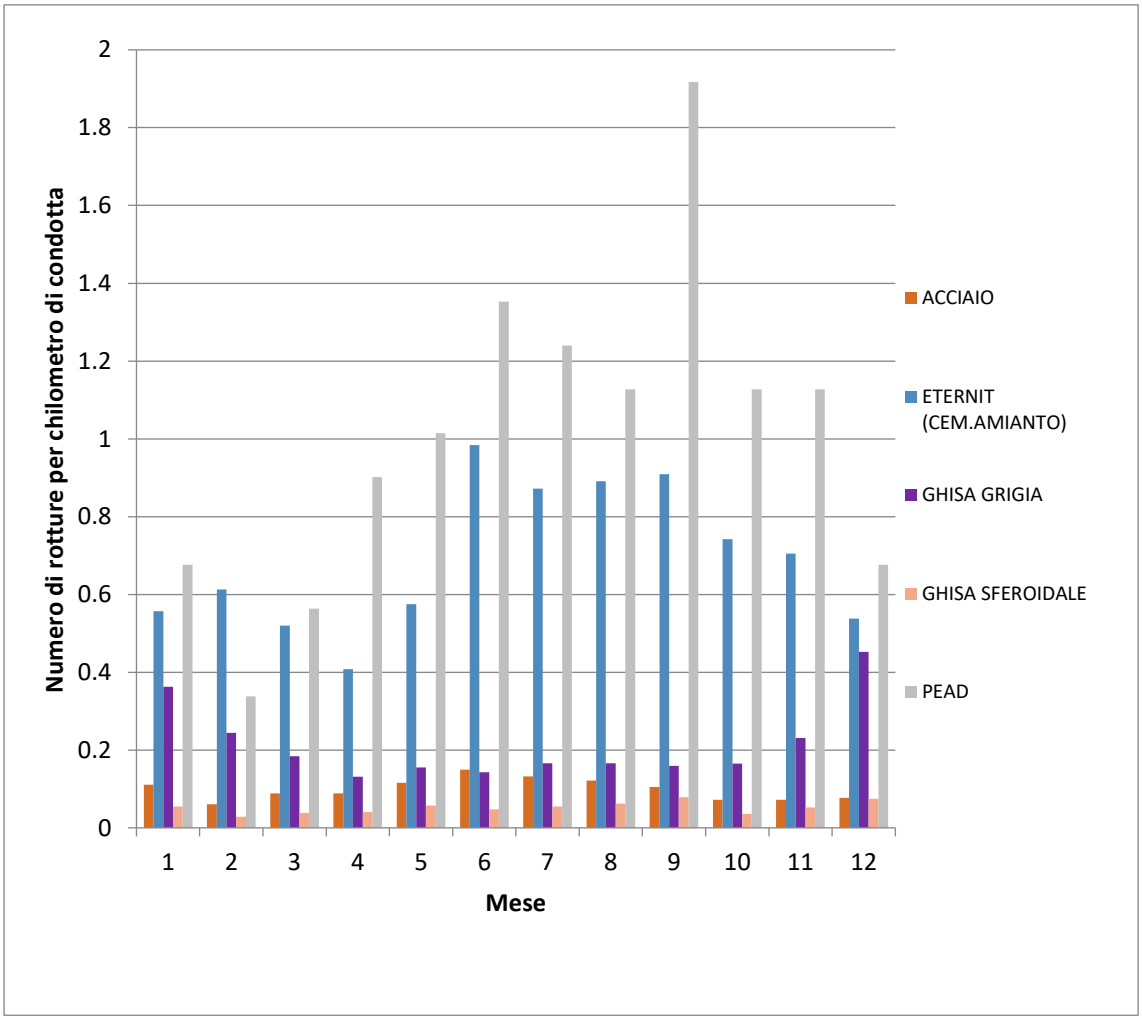


Figura 6.4. Occorrenze delle rotture normalizzate per chilometro di condotta per ogni materiale

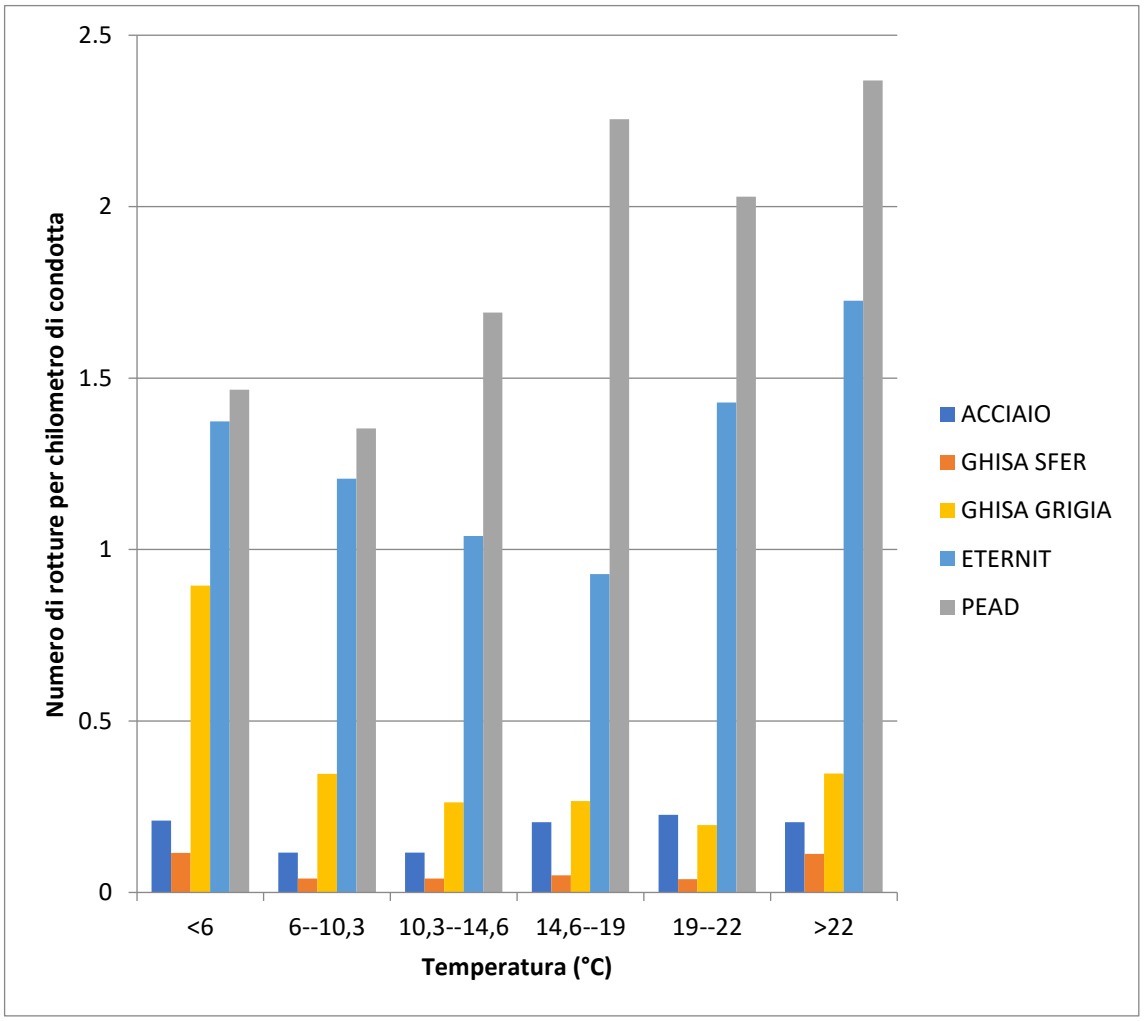
A riprova del diverso comportamento dei materiali in relazione alla temperatura, si riporta la distribuzione delle occorrenze nelle divere fasce di temperature.

Le fasce di temperatura sono state ricavate a partire dalle temperature medie mensili calcolate in corrispondenza di ogni mese, per il periodo dal 2006 al 2016.

Tabella 6.16. Occorrenze delle rotture per le fasce di temperature per ogni materiale

Temperatura (°C)	Acciaio	Ghisa Sferoidale	PEAD	Ghisa grigia	Eternit
<6	38	48	13	851	74
6-10,3	21	17	12	329	65
10,3-14,6	21	17	15	250	56
14,6-19	37	21	20	253	50
19-22	41	16	18	187	77
>22	37	47	21	330	93

Figura 6.5. Distribuzione delle occorrenze delle rotture per le face di temperatura



Risulta evidente che la ghisa grigia sia più vulnerabile a basse temperature, mentre eternit e polietilene presentino più rotture a temperature elevate.

6.2 Correlazione tra le variabili

Prima di procedere con lo studio del modello di regressione è necessario ricercare una eventuale correlazione tra le variabili in gioco poiché, in presenza di stretta correlazione tra due parametri, risulta sufficiente considerarne solo uno.
In presenza di correlazione tra due variabili *X* e *Y,* ad ogni valore della prima variabile è associato un determinato valore della seconda variabile, cioè la prima variabile è funzione della seconda. La correlazione diretta prevede che al crescere di una variabile (o al decrescere), cresca (o decresca) anche la seconda. La correlazione indiretta prevede, invece, che al crescere di una variabile (o al decrescere), decresca (o cresca) la seconda.
Nel caso in esame, le coppie di variabili prese in esame sono:

- mese-materiale;
- mese-diametro;

- mese-carichi in rete;
- mese-età delle condotte;
- materiale-diametro;
- materiale-carichi in rete;
- materiale-età delle condotte;
- diametro-carichi in rete;
- diametro-età delle condotte;
- carichi in rete-età delle condotte.

Il grado di correlazione è espresso mediante l'indice di correlazione *r* (o indice di Pearson): il suo valore varia tra -1 e 1 e, il raggiungimento dei valori estremi di tale intervallo indica forte correlazione tra le variabili. Valori di *r* prossimi a 0 esprimono una sostanziale indipendenza tra i parametri. La sua formulazione è:

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (1)$$

dove σ_X e σ_Y indicano le deviazioni standard e σ_{XY} la covarianza.
Per valutare la correlazione tra la variabile materiale e le altre variabili, è stato assegnato un numero ad ogni materiale presente in rete, come descritto in **Tabella 6.17**.

Tabella 6.17. Numero assegnato ad ogni materiale in rete

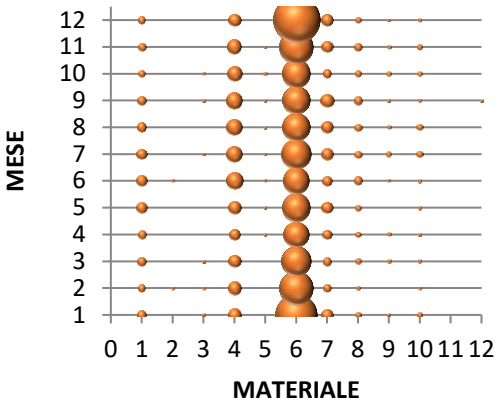
Materiale	Numero
Acciaio	1
Acc Chameroy	2
Acciaio Inox	3
Eternit	4
Ferro zincato-nudo	5
Ghisa grigia	6
Ghisa sferoidale	7
Polietilene ad alta densità (PEAD)	8
Polietilene a bassa densità (PEBD)	9
Piombo	10
Pvc	11
Vetro resina	12

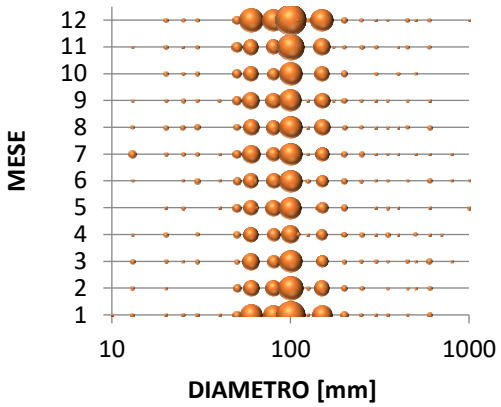
I coefficienti di correlazioni tra le coppie precedentemente menzionate valgono:

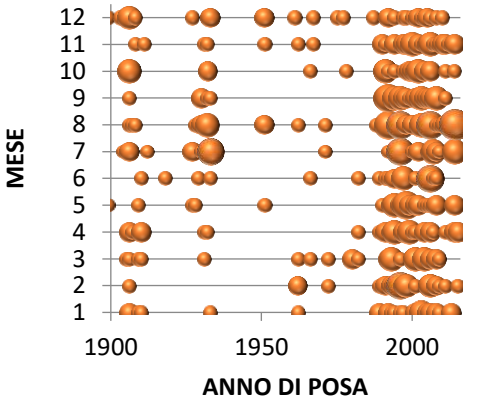
- $r_{\text{mese-materiale}}$=0.028;
- $r_{\text{mese-diametro}}$=-0.015;
- $r_{\text{mese-carichi in rete}}$=-0.005;
- $r_{\text{mese-età delle condotte}}$=-0.08;

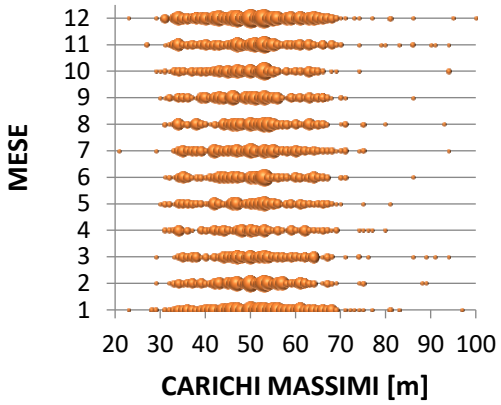
- $r_{materiale\text{-}diametro}$=-0.208;
- $r_{materiale\text{-}carichi\ in\ rete}$=0.003;
- $r_{materiale\text{-}età\ delle\ condotte}$=0.237;
- $r_{diametro\text{-}carichi\ in\ rete}$=0.019;
- $r_{diametro\text{-}età\ delle\ condotte}$=0.069;
- $r_{carichi\ in\ rete\text{-}età\ delle\ condotte}$=0.078.

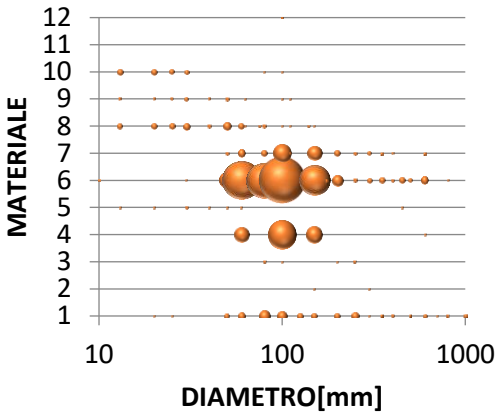
Si riportano in **Figura 6.6** i grafici a bolle ottenuti per tutte le coppie di parametri.
I grafici a bolle sono stati ricavati a partire dalla tabella *Wwvcondottetorino*, una volta estratti i campi delle variabili di interesse. A causa dell'alto numero di dati e della conseguente sovrapposizione di numerosi punti, si è ricorso alla valutazione delle occorrenze di ogni coppia presente nei grafici. In questo modo, la dimensione di ogni bolla fa riferimento al numero di occorrenze in corrispondenza di una determinata coppia di valori: maggiore è la superficie della bolla, maggiori sono le occorrenze di quella coppia.
Si riporta un esempio che possa chiarire quanto appena descritto.
Si estraggano dalla tabella *Wwvcondottetorino* i campi *Materiale* e *Mese* per valutare un'eventuale correlazione tra il materiale e il mese in cui è avvenuta la rottura. Una volta associato un numero ad ogni materiale, si realizza il grafico a dispersione riportato in **Figura 6.7**.

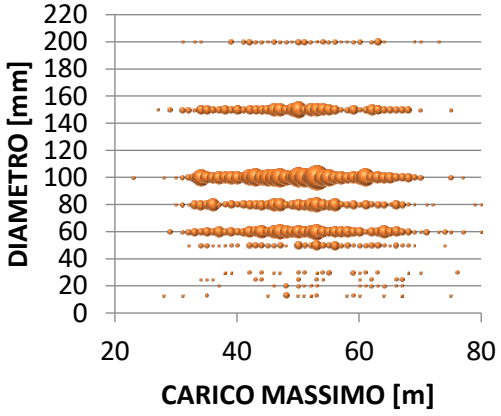

MESE
12
11
10
9
8
7
6
5
4
3
2
1
0 1 2 3 4 5 6 7 8 9 10 11 12
MATERIALE

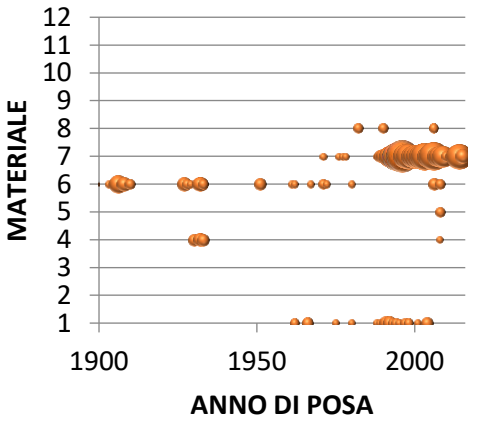

MESE
12
11
10
9
8
7
6
5
4
3
2
1
10
100
1000
DIAMETRO [mm]

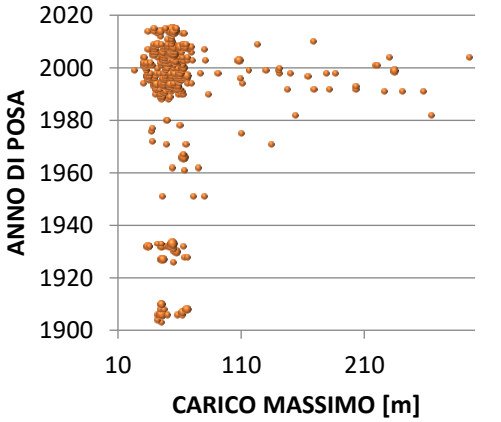

MESE
12
11
10
9
8
7
6
5
4
3
2
1
1900
1950
2000
ANNO DI POSA


MESE
12
11
10
9
8
7
6
5
4
3
2
1
20 30 40 50 60 70 80 90 100
CARICHI MASSIMI [m]


MATERIALE
12
11
10
9
8
7
6
5
4
3
2
1
10
100
1000
DIAMETRO[mm]


DIAMETRO [mm]
220
200
180
160
140
120
100
80
60
40
20
0
20
40
60
80
CARICO MASSIMO [m]


MATERIALE
12
11
10
9
8
7
6
5
4
3
2
1
1900
1950
2000
ANNO DI POSA


ANNO DI POSA
2020
2000
1980
1960
1940
1920
1900
10
110
210
CARICO MASSIMO [m]

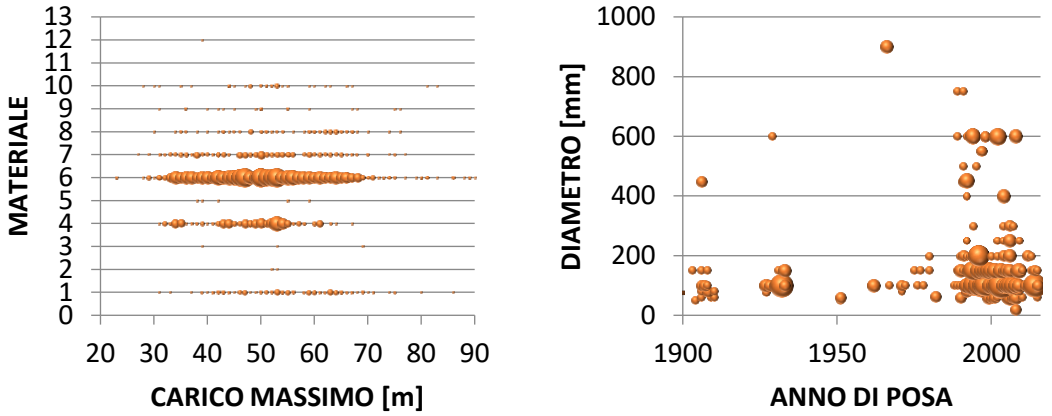


Figura 6.6. Diagrammi di correlazione per i campi di interesse

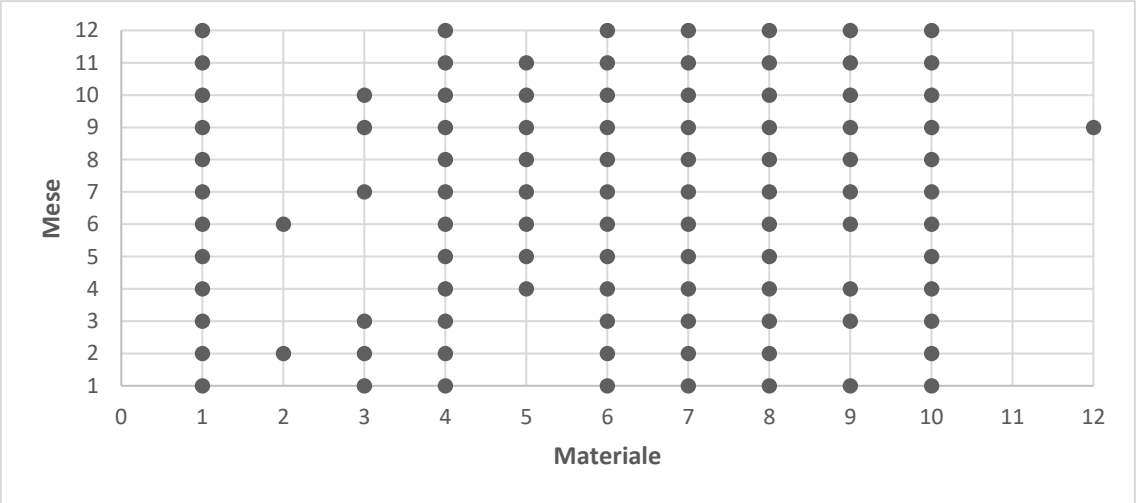


Figura 6.7. Grafico a dispersione relativo ai punti appartenenti ai campi *Materiale* e *Mese*

Il grafico non risulta però adatto ad illustrare questi dati, poiché molti punti si sovrappongono e più coppie sono associate ad un singolo punto. Si calcolano quindi le occorrenze per ogni coppia di punti *Materiale-Mese* e si riportano nel grafico a bolle in **Figura 6.8**.

La figura mostra che il materiale numero 6 (ghisa grigia) presenta il maggior numero di occorrenze e queste si concentrano nei mesi di gennaio e dicembre. Questo risultato conferma quanto ottenuto nel capitolo precedente riguardo il comportamento della ghisa a basse temperature.

Per quanto concerne gli indici di Pearson e i diagrammi di correlazione, è evidente che non vi siano correlazioni tra le variabili poiché i valori di *r* sono prossimi allo 0 e le figure non mostrano delle correlazioni dirette o inverse tra le coppie di parametri, poiché i punti non si dispongono secondo curve definite. Affinché la correlazione sia forte, invece, gli indici devono assumere valori superiori a 0.7

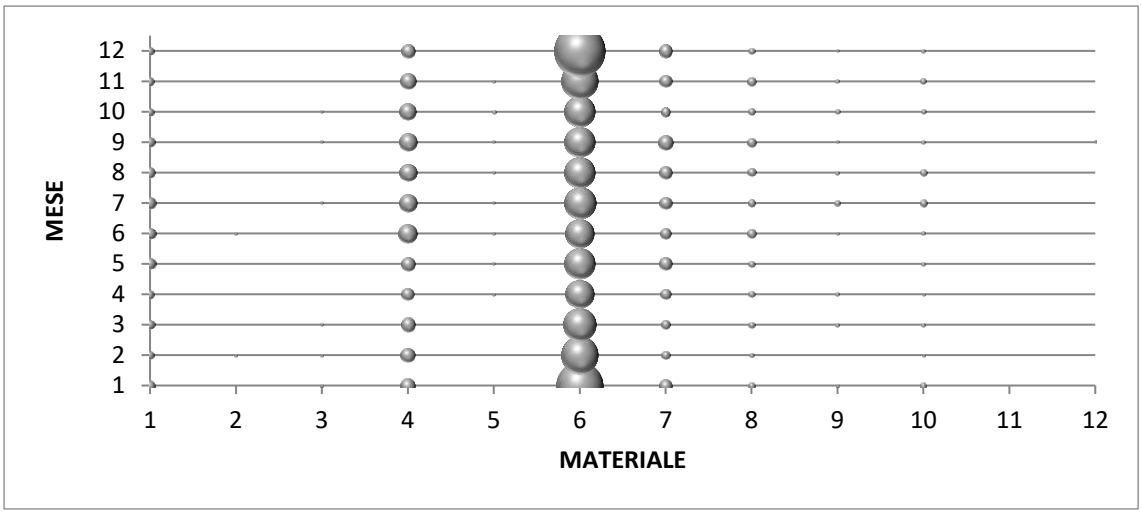


Figura 6.8. Grafico a bolle relativo alle occorrenze *Materiale-Reportdate*

6.3 Analisi della tabella *ReteTorino*

L'analisi della tabella *ReteTorino* è di primaria importanza, in quanto permette di valutare la vulnerabilità delle condotte attraverso la definizione del tasso di fallanza.
Nello specifico, il tasso di fallanza è un indicatore che permette di valutare lo stato delle stesse ed è definito per una determinata classe come:

$$\lambda = \frac{NR}{L} \tag{1}$$

Dove *NR* indica il numero di rotture totali delle condotte appartenenti ad una determinata classe e *L* è la lunghezza totale delle condotte appartenenti alla stessa.
Tale indice è stato calcolato per le classi di condotte appartenenti alla tabella *ReteTorino* poiché questa rappresenta la totalità della rete di Torino e non avrebbe avuto utilità, ai fini del calcolo della vulnerabilità delle condotte, l'applicazione di tale indice alla tabella *Wwvcondottetorino*, contenente esclusivamente gli interventi legati ai guasti. Sono state prese in considerazione le classi di materiali, diametri, anni di posa, carichi massimi e lunghezze. Infine, la stessa procedura è stata adottata esaminando i materiali singolarmente.
Si riporta di seguito un esempio per chiarire la definizione di questo indicatore.
Si supponga di voler calcolare il tasso di fallanza nel caso delle condotte in *Eternit* della rete di Torino. Tali condotte presentano una lunghezza complessiva di *64.37 chilometri* ed un numero totale di rotture che ammonta a 351. Il conseguente tasso di fallanza associato è:

$$\lambda_{Eternit} = \frac{NR}{L} = \frac{64.37}{351} = 5.453$$

Per validare tale dato, è necessario confrontarlo con lo stesso indice calcolato nel caso degli altri materiali. Ad esempio, nel caso della *Ghisa sferoidale*, che caratterizza le condotte della rete per una lunghezza di *425.92 chilometri* e nel periodo di osservazione ha presentato 387 guasti, il valore del tasso di fallanza ammonta a:

$$\lambda_{Ghisa\ sferoidale} = \frac{NR}{L} = \frac{387}{425.92} = 0.909$$

I valori appena ricavati indicano che l'Eternit presenta 5.45 guasti per chilometro di condotta, mentre la Ghisa sferoidale solo 0.91. Di conseguenza, l'Eternit è un materiale più vulnerabile della Ghisa sferoidale.
Questo calcolo è stato effettuato per ogni classe di materiale, diametro, anno di posa, carico massimo e lunghezza. Si può riassumere il procedimento nei seguenti step:

1. A partire dalla tabella *ReteTorino*, sono state suddivise le condotte nelle classi appena citate. Per ogni classe è stata calcolata la lunghezza complessiva delle condotte e il numero di rotture che le appartengono (condotte con flag pari a 1).
2. È stata applicata l'**Equazione 1**, ricavando quindi il tasso di fallanza di ogni classe.

Si riportano di seguito le occorrenze calcolate per tutte le classi della tabella *ReteTorino*. È importante sottolineare che i risultati sono influenzati dall'approssimazione sulle lunghezze delle condotte e dalla filtrazione dettata da una non perfetta geolocalizzazione delle stesse.

Calcolo del tasso di fallanza dei singoli materiali

Una volta suddivise le condotte della tabella *ReteTorino* in classi di materiale e calcolate le lunghezze totali e le occorrenze delle rotture di ogni classe, si ritrovano i risultati riportati in **Tabella 6.18**.
La tabella in esame riporta in terza colonna la percentuale di condotte appartenenti ad una certa classe di materiale, espressa in termini di lunghezza normalizzata rispetto alla lunghezza totale delle condotte. In quarta colonna è invece riportata la percentuale di rotture appartenente ad una determinata classe, rispetto alle rotture totali. L'ultima colonna, di fondamentale importanza, riporta il rispettivo indicatore del tasso di fallanza della classe.
In **Figura 6.9** si riportano invece per via grafica tali risultati, avendo cura di riportate esclusivamente le classi che contraddistinguono almeno lo 0.5% delle condotte della rete.

Tabella 6.18. Calcolo del tasso di fallanza dei singoli materiali presenti nella tabella *ReteTorino*

Materiale	Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
Ghisa grigia	823	51.2	2945	76.0	3.6
Ghisa sferoidale	425.9	26.5	387.0	10.0	0.9
Acciaio	264.2	16.4	173.0	4.5	07
Eternit	64.4	4.0	351.0	9.1	5.5
MATERIALI INDEFINITI	0.06	0.0	0.0	0.0	0.00
PEAD	14.0	0.9	9.0	0.2	0.6
Chameroy	4.8	0.3	4.0	0.1	0.8
PVC	7.6	0.5	4.0	0.1	0.5
Ferro	1.3	0.1	2.0	0.1	1.6
Cemento armato	2.9	0.2	1.0	0.0	0.3
TOT	1608	100	3876	100	

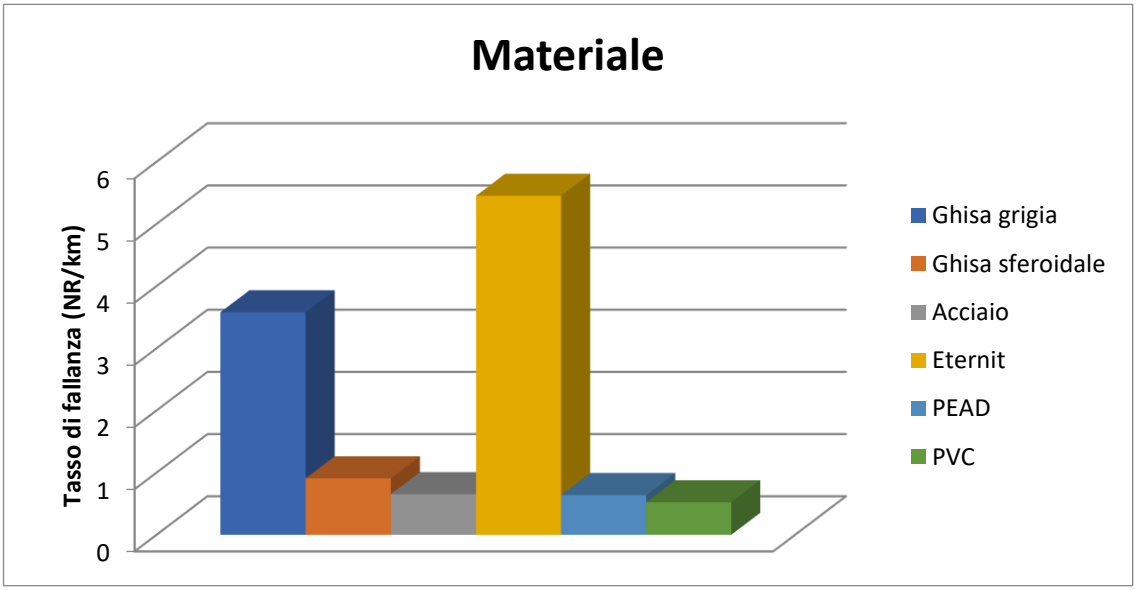


Figura 6.9. Tasso di fallanza dei diversi materiali presenti nella tabella *ReteTorino*

I materiali con tassi di fallanza più alti sono Eternit e Ghisa grigia. I risultati sono differenti da quelli ottenuti per la tabella *Wwvcondottetorino*, in quanto la tabella citata racchiude solo condotte che sono andate incontro a guasti nel periodo 2006-2016, mentre la tabella *ReteTorino* è la tabella raffigurante la totalità delle condotte sul suolo torinese. Inoltre, nella prima analisi il polietilene ad alta densità risultava il materiale più vulnerabile, nettamente in contrasto con quanto ricavato per la tabella *ReteTorino*. Si ricorda, però, che in questa tabella sono stati esclusi le condotte inferiori a 10 metri e, di conseguenza, molti dati riferiti a tale materiale sono stati esclusi. Anche le lievi discrepanze ottenute per gli altri materiali sono da attribuire al criterio adottato per la scelta delle lunghezze delle condotte.

Calcolo del tasso di fallanza delle diverse classi di diametri

Nello stesso modo, sono stati calcolati i tassi di fallanza relativi alle diverse classi di diametri. Si riportano i risultati in **Tabella 6.19** e il relativo grafico in **Figura 6.10**.

Tabella 6.19. Calcolo del tasso di fallanza delle classi di diametri presenti nella tabella *ReteTorino*

Diametro (mm)		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**D ≤ 100 mm**	661.3	41.1	2777	71.6	4.2
CLASSE 2:	**100mm ≤ D ≤ 200 mm**	376.5	23.4	821	21.2	2.2
CLASSE 3:	**D ≥ 200 mm**	570.6	35.5	283	7.3	0.5
	TOT	1608.4	100	3881	100	

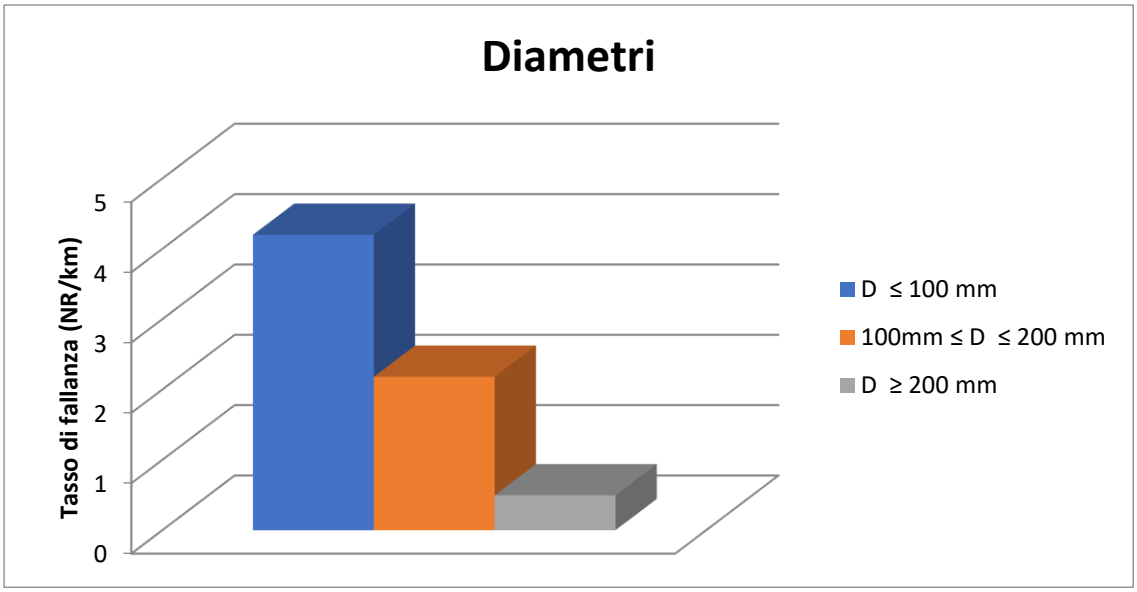


Figura 6.10. Tasso di fallanza delle diverse classi di diametri presenti nella tabella *ReteTorino*

I risultati mostrano che il tasso di rottura aumenta al diminuire del diametro. Lo stesso risultato è stato ottenuto nell'analisi della tabella *Wwvcondottetorino*.

Calcolo del tasso di fallanza relativo all'anno di posa delle condotte

Lo stesso procedimento è stato adottato per le informazioni relative agli anni di posa. Nello specifico, tali informazioni sono state classificate in classi ventennali e successivamente sono state calcolate le lunghezze totali e le occorrenze delle rotture di ogni classe (**Tabella 6.20**).

Tabella 6.20. Calcolo del tasso di fallanza delle classi di posa presenti nella tabella *ReteTorino*

Anno di posa		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**1870 - 1890**	1.1	0.3	3	0.7	2.7
CLASSE 2:	**1890 - 1910**	11.9	3.2	35	8.4	2.9
CLASSE 3:	**1910 - 1930**	13.0	3.5	16	3.8	1.2
CLASSE 4:	**1930 - 1950**	3.3	0.9	23	5.5	6.8
CLASSE 5:	**1950 - 1970**	6.9	1.9	18	4.3	2.6
CLASSE 6:	**1970 - 1990**	34.9	9.4	24	5.7	0.7
CLASSE 7:	**1990 - 2010**	264.3	71.0	269	64.4	1.0
CLASSE 8:	**2010 →**	36.4	9.8	30	7.2	0.8
	TOT	372.1	100	418	100	

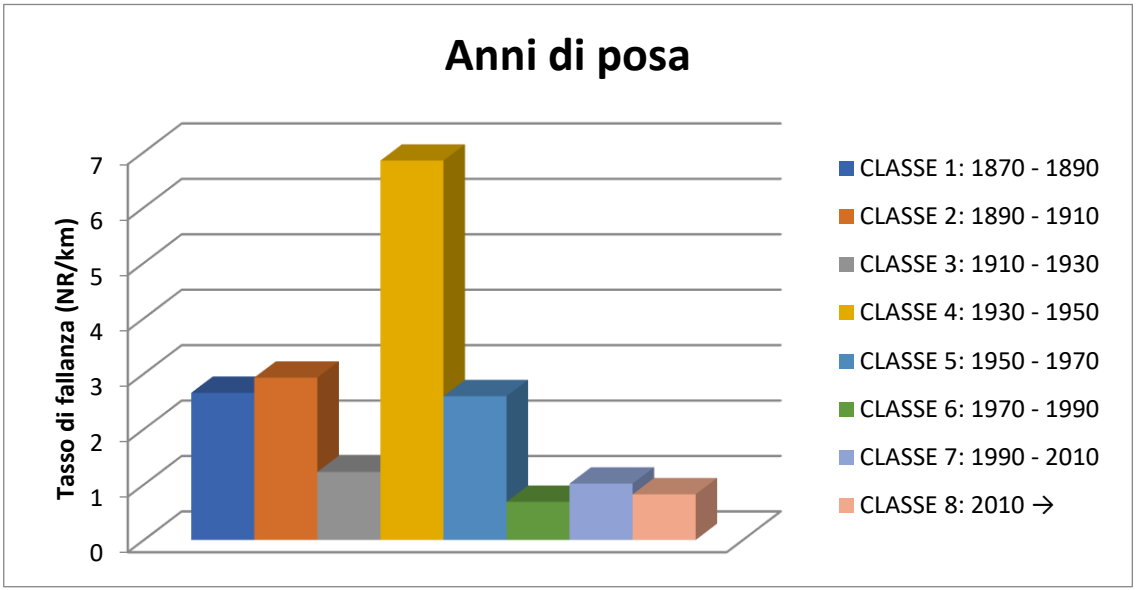


Figura 6.11. Tasso di fallanza delle classi di posa presenti nella tabella *ReteTorino*

La maggior parte delle condotte dotate di età è stata posata nel ventennio 1990-2010, mentre il picco di rotture è stato registrato tra le condotte posate negli anni 1930-1950.

Calcolo del tasso di fallanza delle classi di carico massimo

Si riportano in **Tabella 6.21** i risultati per le diverse classi di carico.

Tabella 6.21. Calcolo del tasso di fallanza delle classi di carico massimo presenti nella tabella *ReteTorino*

Carico massimo (m)		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**(0, 40]**	224.4	16.7	576	15.6	2.6
CLASSE 2:	**(40, 80]**	1063.2	78.9	2997	81.2	2.8
CLASSE 3	**(80, 120]**	16.5	1.2	56	1.5	3.4
CLASSE 4:	**(120,160]**	13.1	0.9	63	1.7	4.8

CLASSE 5:	**>160**	29.3	2.1	110	3.0	3.8
	TOT	1346.6	100	3692	100	

La classe di carico con tasso di fallanza più elevata è la quarta, con carichi compresi tra i 120 e i 160 metri.

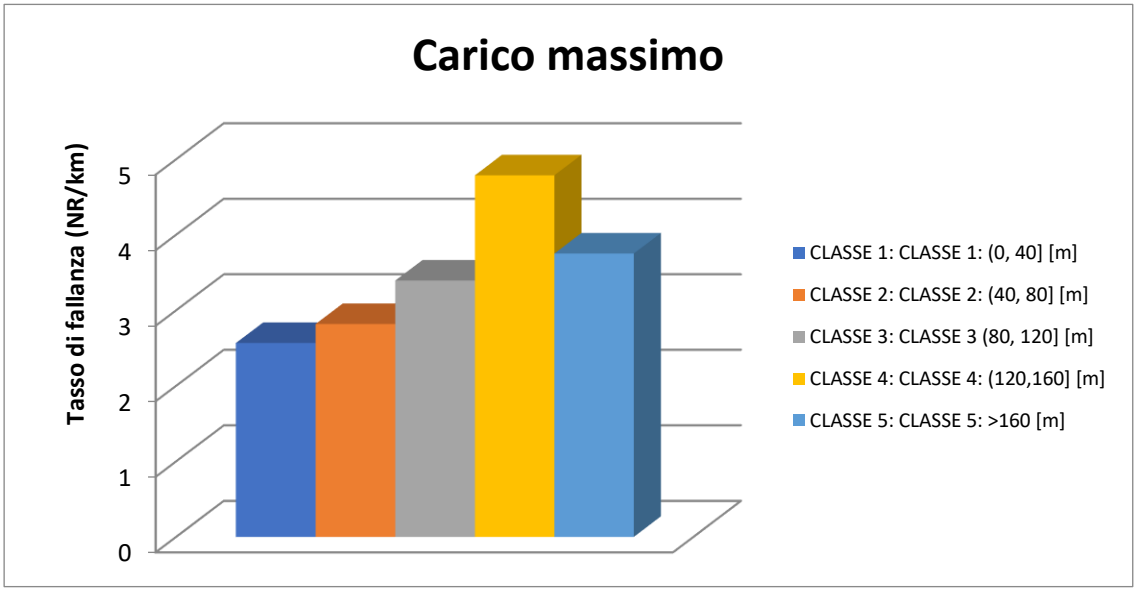


Figura 6.12. Tasso di fallanza delle classi carico massimo presenti nella tabella *ReteTorino*

Calcolo del tasso delle classi di lunghezza

Si riportano in **Tabella 6.22** i risultati per le diverse classi di lunghezza.
È possibile notare che l'85% delle condotte è caratterizzato da lunghezze comprese entro i 400 metri. I tassi di fallanza maggiori si ritrovano nelle prime tre classi: minore è la lunghezza di una condotta, maggiore è il numero di rotture. Questa informazione è coerente con quella relativa ai diametri: infatti, a lunghezze minori, in genere, corrispondono diametri minori.

Tabella 6.22. Calcolo del tasso di fallanza delle classi di lunghezza nella tabella *ReteTorino*

Lunghezza (m)		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**(0, 400]**	137	85.7	3840	98.9	2.8
CLASSE 2:	**(400, 800]**	106.2	6.6	34	0.8	0.3
CLASSE 3:	**(800, 1200]**	58.2	3.6	6	0.1	0.1
CLASSE 4:	**(1200,160]**	24.8	1.5	1	0.1	0.1
CLASSE 5:	**(1600, 2000]**	14.2	0.9	0	0	0.00
CLASSE 6:	**(2000, 2400]**	10.3	0.6	0	0	0.00
CLASSE 7:	**(2400, 2800]**	5.7	0.3	0	0	0.00
CLASSE 8:	**(2800, 3200]**	2.7	0.1	0	0	0.00
CLASSE 9:	**(3200, 3600]**	3.3	0.2	0	0	0.00
CLASSE 10:	**(3600, 4000]**	4.3	0.2	1	0.01	0.2
	TOT	1608	100	3882	100	

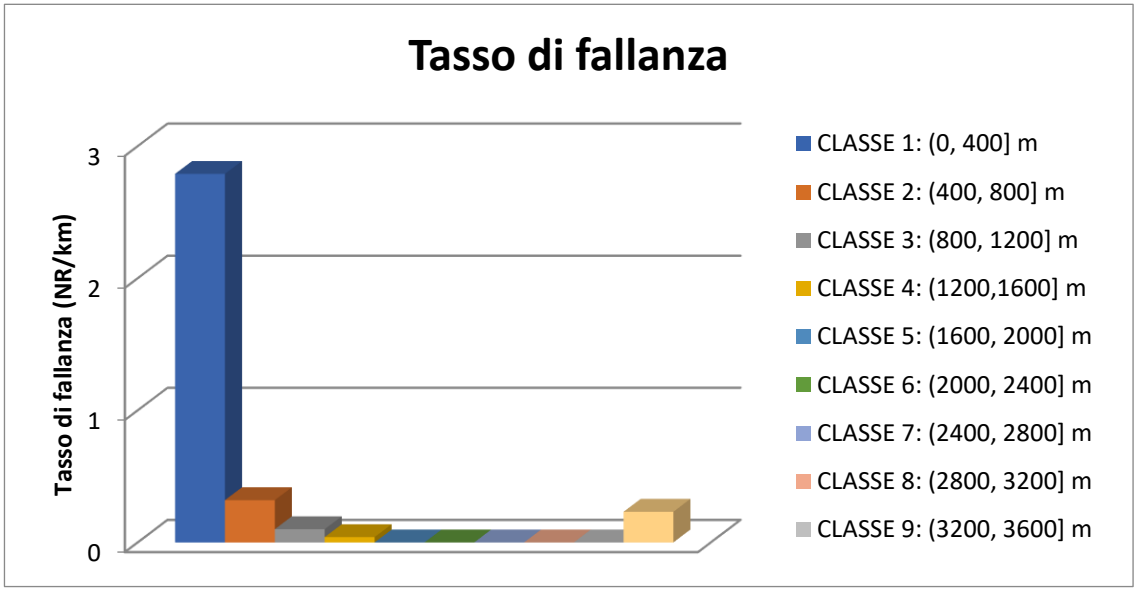


Figura 6.13. Tasso di fallanza delle classi di lunghezza presenti nella tabella *ReteTorino*

Lo stesso procedimento è stato adottato prendendo in considerazione i materiali singolarmente, per i quali è stato ricalcolato il tasso di fallanza delle classi di diametro, anno di posa, carico massimo e lunghezza.
Ogni sotto-tabella è stata ricavata mediante il software di programmazione *Python.* Sono state denominate come segue:

- *ReteTorinoGhisaGrigia;*
- *ReteTorinoGhisaSferoidale;*
- *ReteTorinoPead;*
- *ReteTorinoEternit;*
- *ReteTorinoAcciaio.*

Per ogni materiale è stato valutato il numero di condotte presente in rete, il numero di rotture avvenute e le occorrenze delle classi di diametro, anno di posa, carico massimo e lunghezza. In seguito, è stato valutato il tasso di fallanza delle diverse classi.
La scelta di una suddivisione dell'analisi in base al materiale deriva dal diverso comportamento degli stessi e dalla necessità di comprendere l'eventuale vulnerabilità di determinati materiali in concomitanza alle altre caratteristiche prese in esame. Sono stati analizzati solo i materiali che caratterizzano per almeno lo 0.5% la lunghezza totale della rete.

Tabella *ReteTorinoGhisaGrigia*

Una volta estratte le sole condotte contraddistinte dal materiale Ghisa grigia, si analizzano le classi di diametro, anno di posa e carico massimo.
Si riporta l'analisi delle classi di diametro.

Tabella 6.23. Calcolo del tasso di fallanza delle classi di diametro presenti nella tabella *ReteTorinoGhisaGrigia*

Diametro (mm)		Lunghezza (Km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**D ≤ 100**	459.1	55.8	2232	75.8	4.9
CLASSE 2:	**100 ≤ D ≤ 200**	179.8	21.9	575	19.5	3.2
CLASSE 3:	**D ≥ 200**	183.7	22.3	138	4.7	0.8
	TOT	822.6	100	2945	100	

La ghisa grigia caratterizza la rete per oltre il 50% con diametri inferiori ai 100 millimetri. Questi ultimi sono anche quelli con tasso di fallanza maggiore. All'aumentare del diametro, il tasso di fallanza diminuisce.

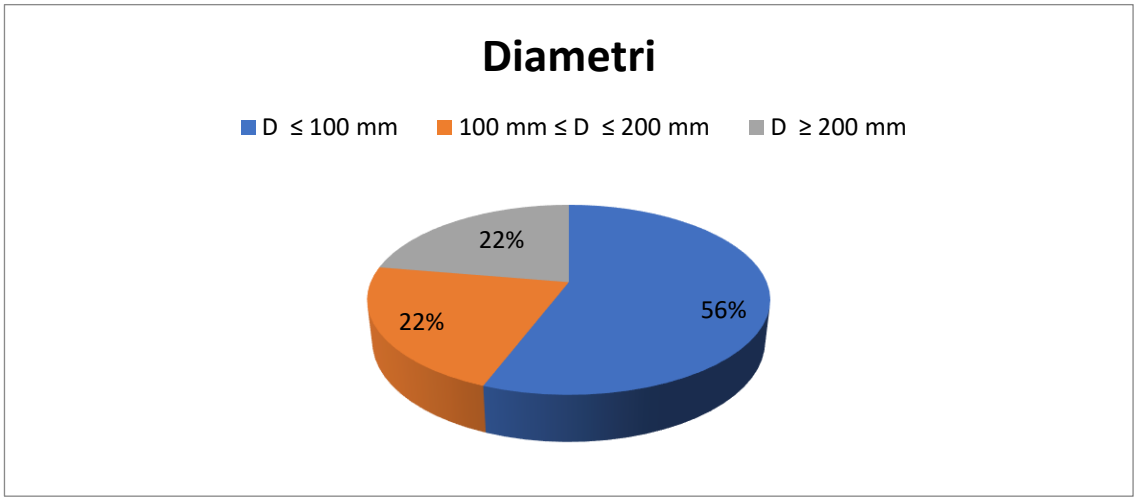


Figura 6.14. Composizione in diametri delle condotte della tabella *ReteTorinoGhisaGrigia*

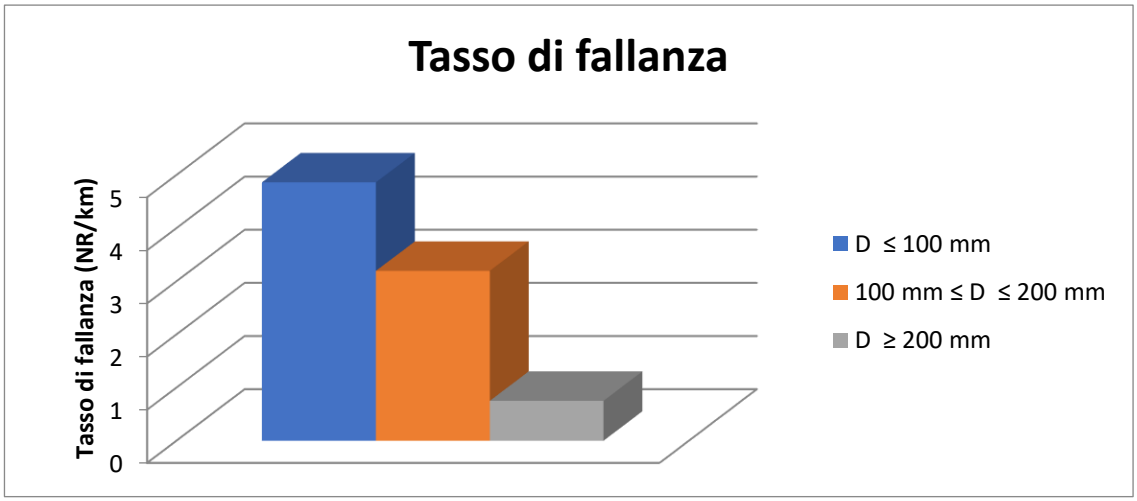


Figura 6.14. Tasso di fallanza delle classi di diametro della tabella *ReteTorinoGhisaGrigia*

La stessa procedura è stata adottata per le classi relative all'anno di posa.

Solo il 4% delle condotte è provvisto dell'informazione relativa all'anno di posa. La maggior parte di queste condotte è stata posata prima del 1950, poiché la ghisa grigia è

stato un materiale ampiamente utilizzato in passato e recentemente sostituito dalla ghisa sferoidale.

Tabella 6.24. Calcolo del tasso di fallanza delle classi relative all'anno di posa delle condotte presenti nella tabella *ReteTorinoGhisaGrigia*

Anno di posa		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**1870 - 1890**	1.1	3.4	3	3.3	2.7
CLASSE 2:	**1890 - 1910**	11.2	33.4	35	38.9	3.1
CLASSE 3:	**1910 - 1930**	12.3	36.8	11	12.2	0.9
CLASSE 4:	**1930 - 1950**	2.9	8.7	13	14.4	4.5
CLASSE 5:	**1950 - 1970**	1.5	4.7	11	12.2	7.0
CLASSE 6:	**1970 - 1990**	0.8	2.5	4	4.4	4.8
CLASSE 7:	**1990 - 2010**	3.5	10.6	13	14.4	3.7
	TOT	33.5	100	90	100	

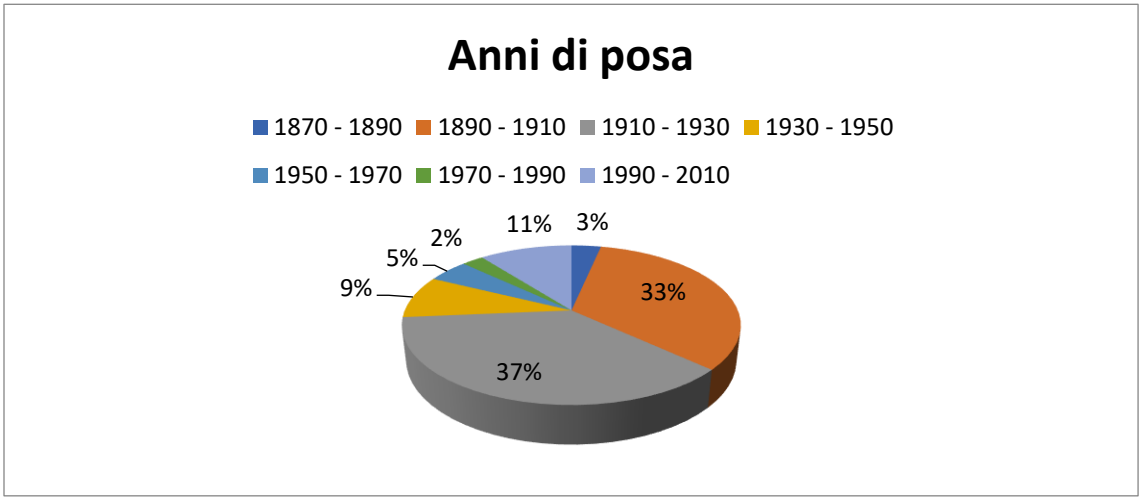


Figura 6.15. Composizione classi di posa delle condotte della tabella *ReteTorinoGhisaGrigia*

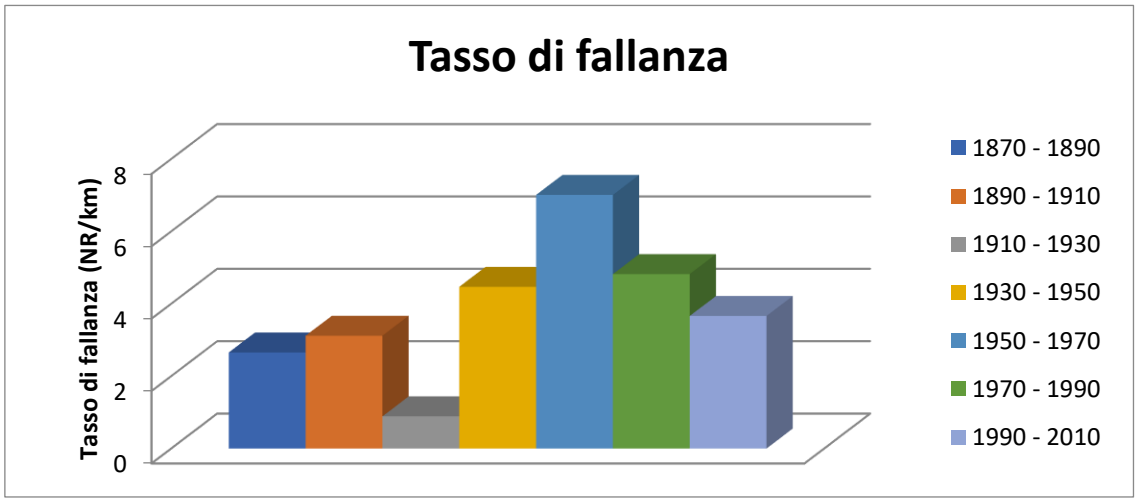


Figura 6.16. Tasso di fallanza delle classi di posa della tabella *ReteTorinoGhisaGrigia*

La stessa procedura è stata adottata per le classi relative al carico massimo.

Tabella 6.25. Calcolo del tasso di fallanza delle classi relative al carico massimo delle condotte presenti nella tabella *ReteTorinoGhisaGrigia*

Carico massimo [m]		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**(0, 40]**	108.0	14.0	450	15.6	4.2
CLASSE 2:	**(40, 80]**	644.0	83.4	2295	79.8	3.6
CLASSE 3:	**(80, 120]**	4.7	0.6	40	1.4	8.4
CLASSE 4:	**(120,160]**	5.1	0.6	33	1.1	6.4
CLASSE 5:	**>160**	9.5	1.2	59	2.1	6.2
	TOT	771.5	100	2877	100	

È possibile notare che il tasso di rottura aumenta all'aumentare della classe di carico, ma la classe caratterizzata da un tasso di fallanza maggiore è quella tra gli 80 e i 120 metri di carico.

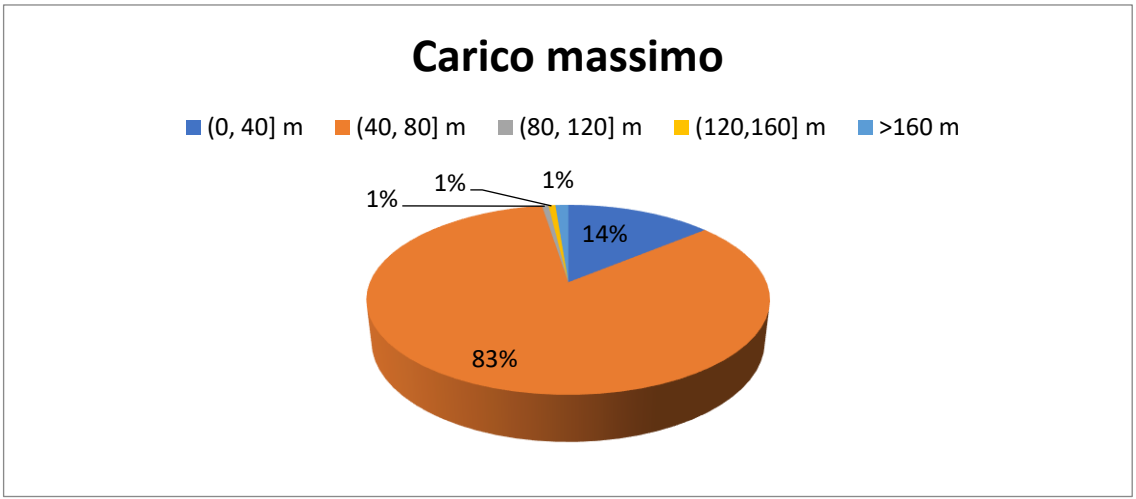


Figura 6.17. Composizione classi di carico delle condotte della tabella *ReteTorinoGhisaGrigia*

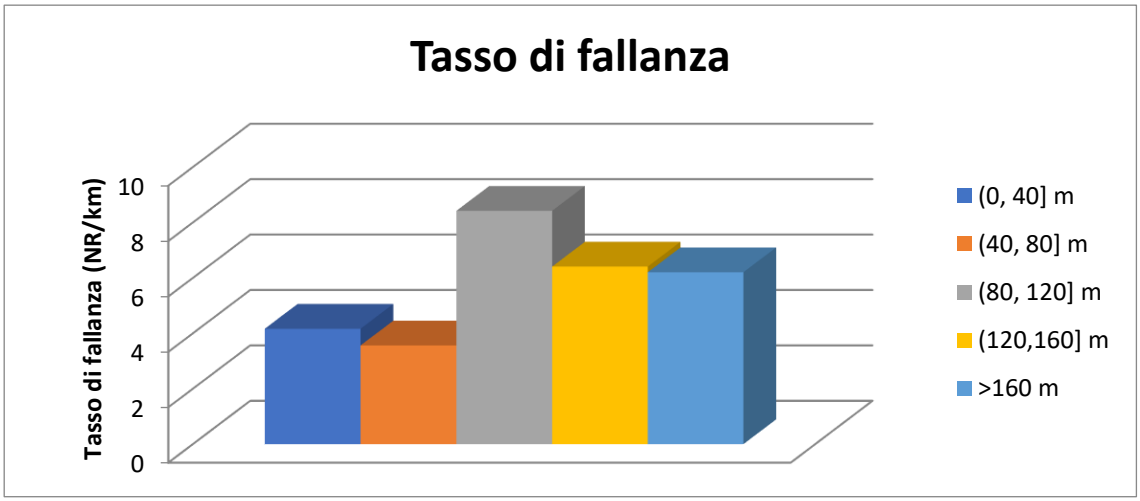


Figura 6.18. Tasso di fallanza delle classi di carico delle condotte della tabella *ReteTorinoGhisaGrigia*

La stessa procedura è stata adottata per le classi relative alla lunghezza delle condotte.

Tabella 6.26. Calcolo del tasso di fallanza delle classi relative alla lunghezza delle condotte presenti nella tabella *ReteTorinoGhisaGrigia*

Lunghezza (m)		Lunghezze (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**(0, 400]**	756.4	91.9	2929	99.5	3.9
CLASSE 2:	**(400, 800]**	38.1	4.6	15	0.5	0.4
CLASSE 3:	**(800, 1200]**	15.5	1.9	1	0.0	0.1
CLASSE 4:	**(1200,1600]**	9.5	1.2	0	0.0	0.0
CLASSE 5:	**(1600, 2000]**	0	0.00	0	0.0	/
CLASSE 6:	**(2000, 2400]**	0	0.00	0	0.0	/
CLASSE 7:	**(2400, 2800]**	0	0.00	0	0.0	/
CLASSE 8:	**(2800, 3200]**	0	0.00	0	0.0	/
CLASSE 9:	**(3200, 3600]**	3.4	0.4	0	0.0	0.0
CLASSE 10:	**(3600, 4000]**	0	0.00	0	0.0	/
	TOT	823	100	2945	100	

La tabella appena proposta fornisce importanti indicazioni sull'influenza della lunghezza delle condotte. Condotte più corte sono più vulnerabili a rotture.

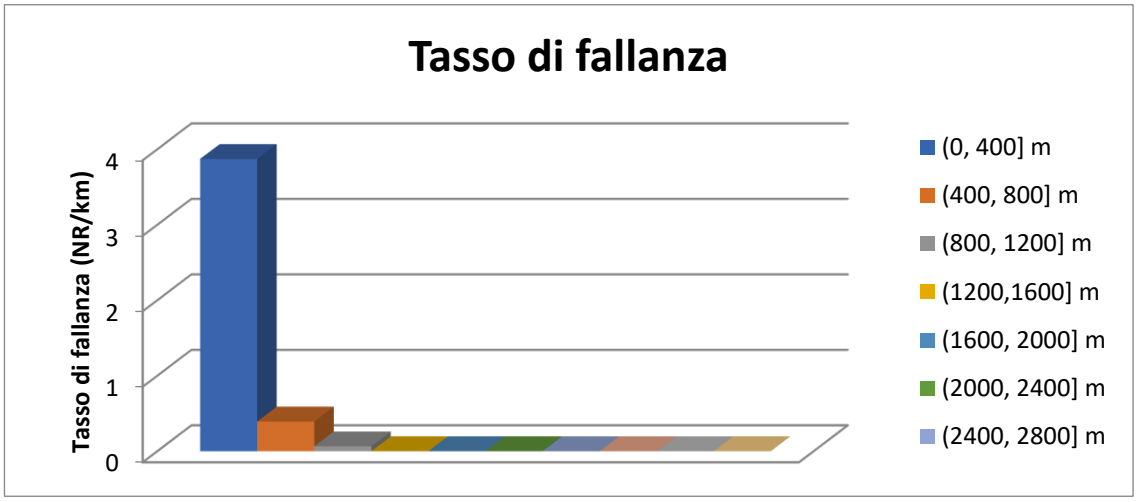


Figura 6.19. Tasso di fallanza delle classi di lunghezza della tabella *ReteTorinoGhisaGrigia*

Tabella *ReteTorinoGhisaSferoidale*

Come nel caso precedente, questa tabella è stata ottenuta come estrazione dalla tabella *ReteTorino*.
Si analizzano i campi diametro, anno di posa, carico massimo e lunghezza, come fatto nel caso della Ghisa grigia.

Tabella 6.27. Calcolo del tasso di fallanza delle classi di diametro presenti nella tabella *ReteTorinoGhisaSferoidale*

Diametro (mm)		Lunghezza (Km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**D ≤ 100**	142.6	33.5	197	50.9	1.4
CLASSE 2:	**100mm ≤ D ≤ 200**	161.9	38.0	140	36.2	0.9
CLASSE 3:	**D ≥ 200**	121.5	28.5	50	12.9	0.4
	TOT	426	100	387	100	

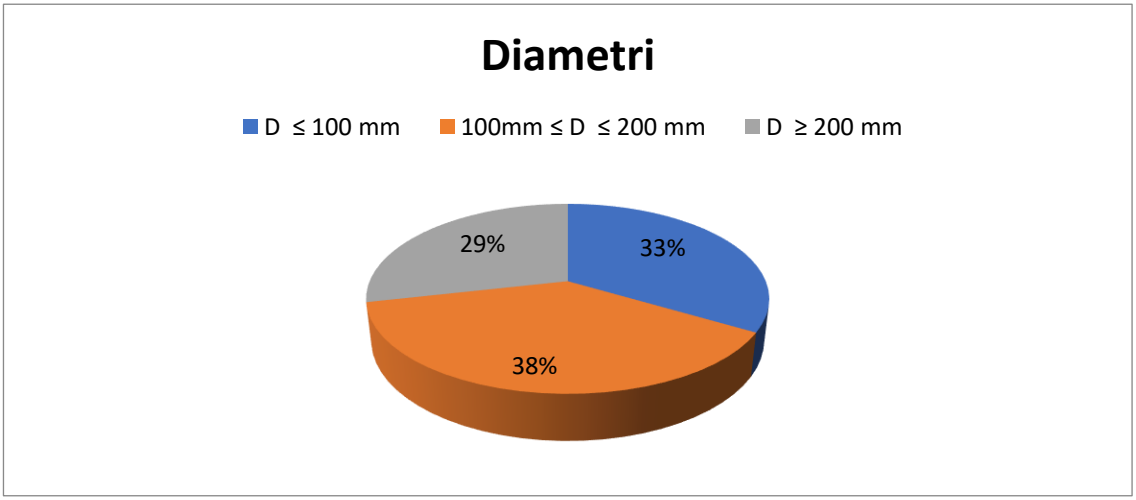


Figura 6.20. Composizione in diametri della tabella *ReteTorinoGhisaSferoidale*

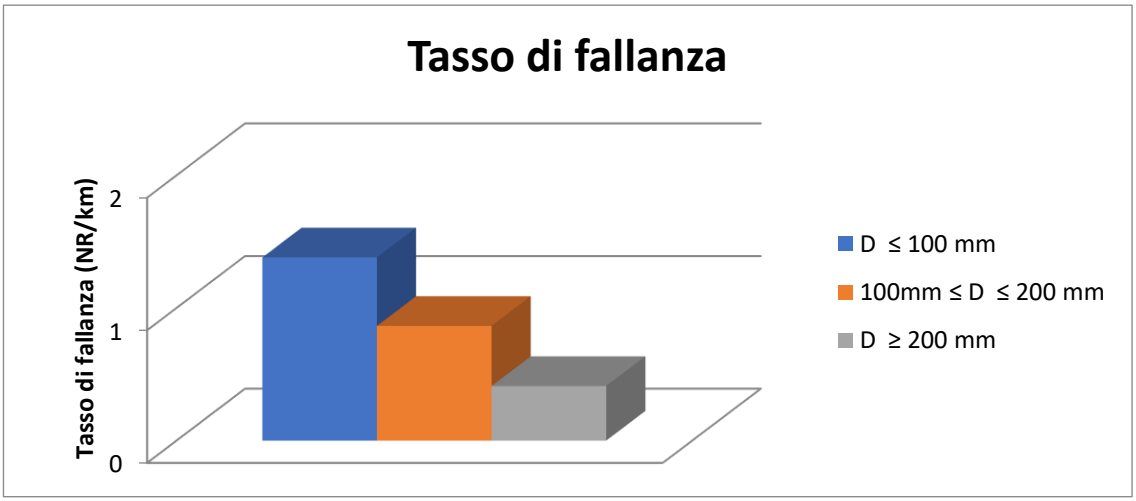


Figura 6.21. Tasso di fallanza delle classi di diametro della tabella *ReteTorinoGhisaSferoidale*

Il tasso di rottura diminuisce all'aumentare del diametro. La maggior parte delle condotte nella rete di Torino è caratterizzata da diametri compresi tra 100 millimetri e 200 millimetri.

La stessa procedura è stata adottata per le classi relative all'anno di posa.

Tabella 6.28. Calcolo del tasso di fallanza delle classi di posa presenti nella tabella *ReteTorinoGhisaSferoidale*

Anno di posa		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**1950 - 1970**	0.2	0.1	0	0.0	0.0
CLASSE 2:	**1970 - 1990**	20.8	7.7	13	5.1	0.6
CLASSE 3:	**1990 - 2010**	216.3	80.1	212	83.1	1.0
CLASSE 4:	**2010 →**	32.61	12.1	30	11.8	0.9
	TOT	372.1	100	418	100	

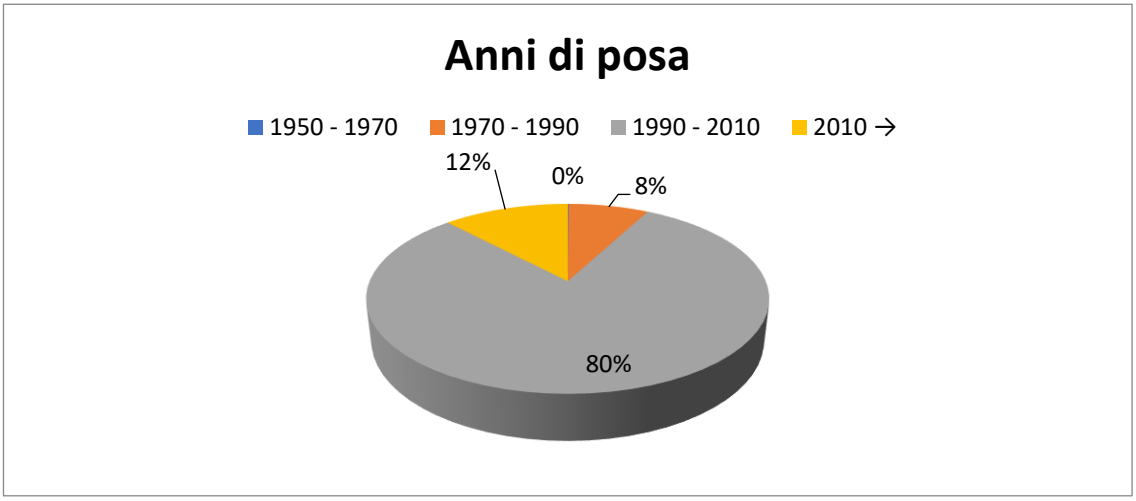


Figura 6.22. Composizione in anni di posa della tabella *ReteTorinoGhisaSferoidale*

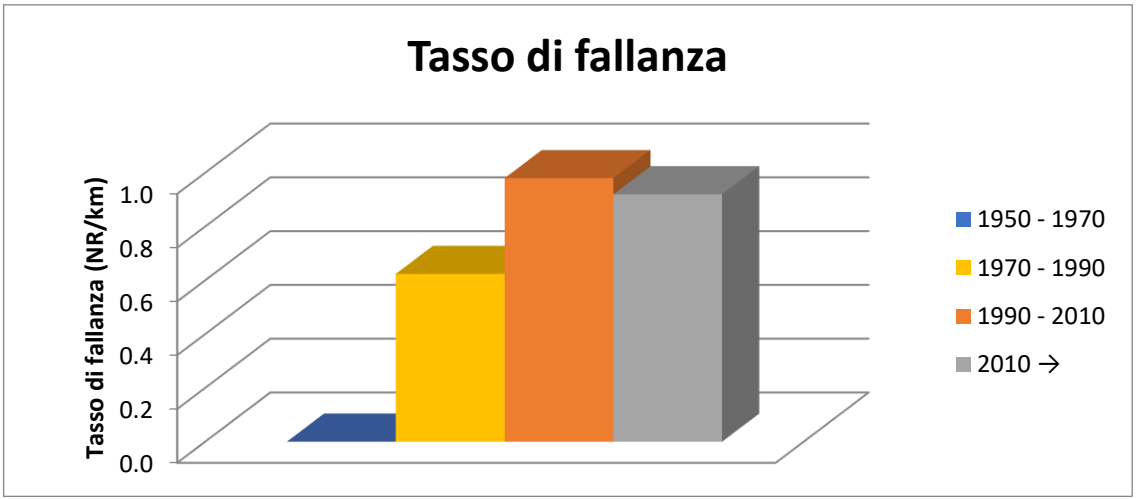


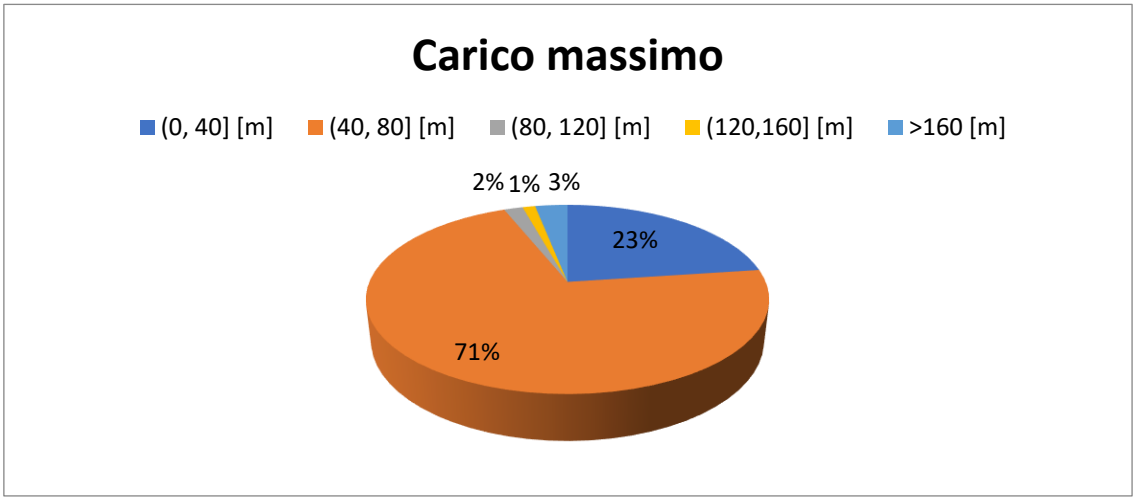
Figura 6.23. Tasso di fallanza delle singole classi di posa della tabella *ReteTorinoGhisaSferoidale*

Dato il recente impiego della ghisa sferoidale, il 92% delle condotte è posteriore al 1990. Oltre il 64% delle condotte dispone inoltre di tale dato.
I grafici mostrano che le condotte recenti presentano tassi di fallanza più elevati.
La stessa procedura è stata adottata per le classi relative al carico massimo.

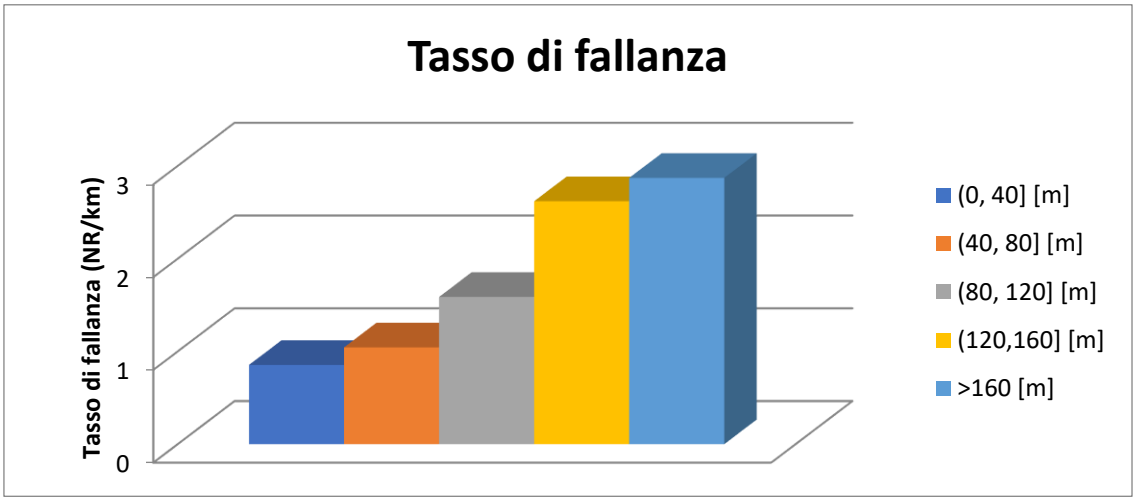
Tabella 6.29. Calcolo del tasso di fallanza delle classi relative al carico massimo delle condotte presenti nella tabella *ReteTorinoGhisaSferoidale*

Carico massimo [m]		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**(0, 40]**	79.6	22.9	68	18.0	0.9
CLASSE 2:	**(40, 80]**	247.2	71.0	258	68.3	1.0
CLASSE 3:	**(80, 120]**	6.3	1.8	10	2.6	1.6
CLASSE 4:	**(120,160]**	4.2	1.2	11	2.9	2.6
CLASSE 5:	**>160**	10.8	3.1	31	8.2	2.9
	TOT	348.1	100	378	100	

Figura 6.24. Composizione classi di carico delle condotte della tabella *ReteTorinoGhisaSferoidale*



6.25. Fallanza delle classi di carico delle condotte della tabella *ReteTorinoGhisaSferoidale*



Il tasso di rottura aumenta all'aumentare del carico. Il 71% della rete è caratterizzato da carichi tra i 40 e gli 80 metri.

La stessa procedura è stata adottata per le classi relative alla lunghezza delle condotte.

Tabella 6.30. Calcolo del tasso di fallanza delle classi relative alla lunghezza delle condotte presenti nella tabella *ReteTorinoGhisasferoidale*

Lunghezza (m)		Lunghezze (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**(0, 400] m**	376.6	86.5	384	99.2	1.0
CLASSE 2:	**(400, 800] m**	33.2	7.6	2	0.5	0.1
CLASSE 3:	**(800, 1200] m**	14.9	3.4	1	0.3	0.1
CLASSE 4:	**(1200,1600] m**	3.9	0.9	0	0.0	0.0
CLASSE 5:	**(1600, 2000] m**	3.7	0.8	0	0.0	0.0
CLASSE 6:	**(2000, 2400] m**	0	0.0	0	0.0	/
CLASSE 7:	**(2400, 2800] m**	0	0.0	0	0.0	/
CLASSE 8:	**(2800, 3200] m**	3	0.7	0	0.0	0.0
CLASSE 9:	**(3200, 3600] m**	0	0.0	0	0.0	/
CLASSE 10:	**(3600, 4000] m**	0	0.0	0	0.0	/
	TOT	435.3	100	387	100	

La tabella appena proposta fornisce importanti indicazioni sull'influenza della lunghezza delle condotte. Condotte più corte sono più vulnerabili a rotture.

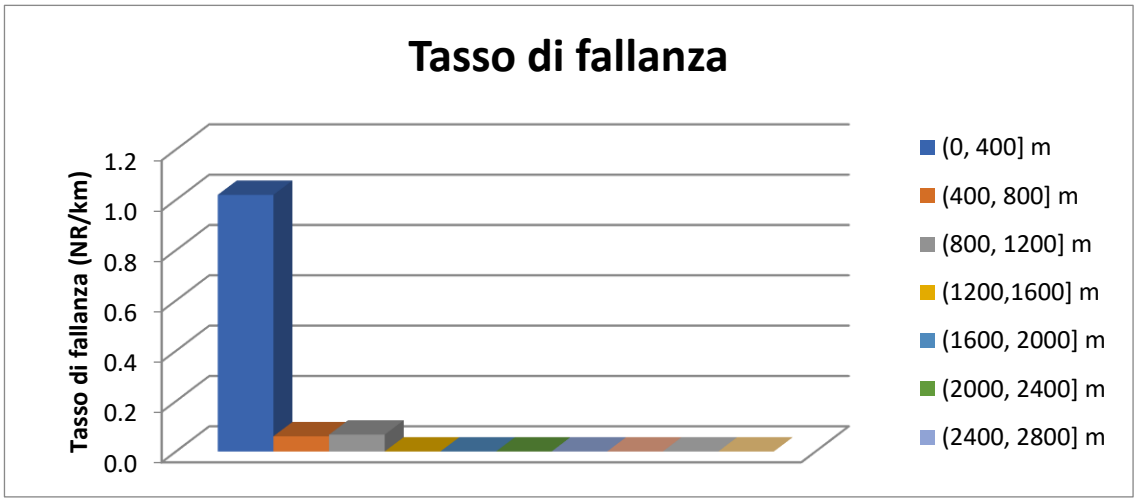


Figura 6.26. Tasso di fallanza delle classi di lunghezza della tabella *ReteTorinoGhisaSferoidale*

Lo stesso risultato è riportato nel diagramma a barre in **Figura 6.26**: all'aumentare della lunghezza delle condotte, diminuisce il tasso di fallanza delle stesse.

Tabella *ReteTorinoPEAD*

Come nel caso precedente, questa tabella è stata ottenuta come estrazione dalla tabella *ReteTorino*.
Si analizzano i campi diametro, anno di posa, carico massimo e lunghezza.

Tabella 6.31. Calcolo del tasso di fallanza delle classi di diametro presenti nella tabella *ReteTorinoPEAD*

Diametro (mm)		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**D ≤ 100**	7.28	51.7	8	88.9	1.1
CLASSE 2:	**100mm ≤ D ≤ 200**	0.5	3.6	1	11.1	2.2
CLASSE 3:	**D ≥ 200**	6.3	44.7	0	0.0	0.0
	TOT	14.1	100	9	100	

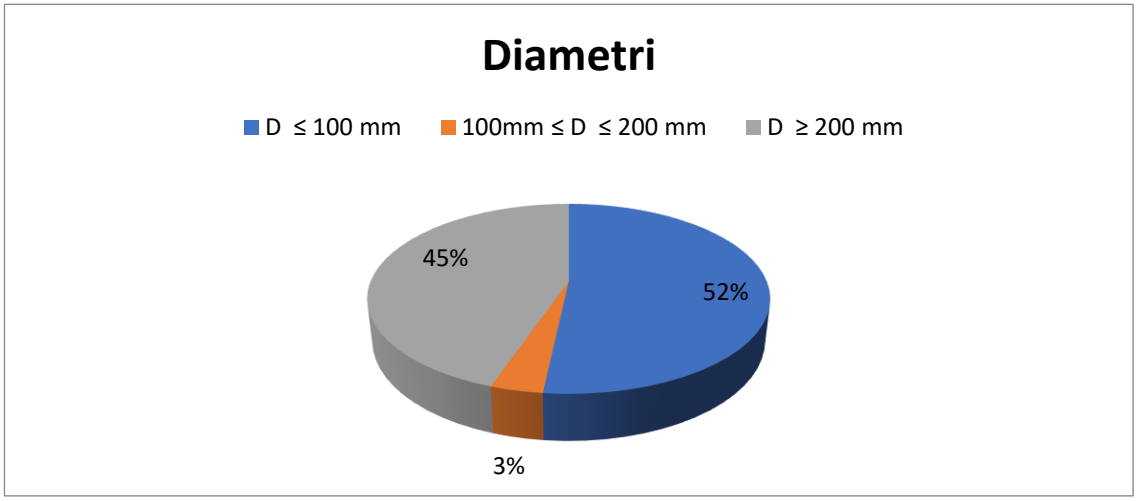


Figura 7.27. Composizione in diametri della tabella *ReteTorinoPEAD*

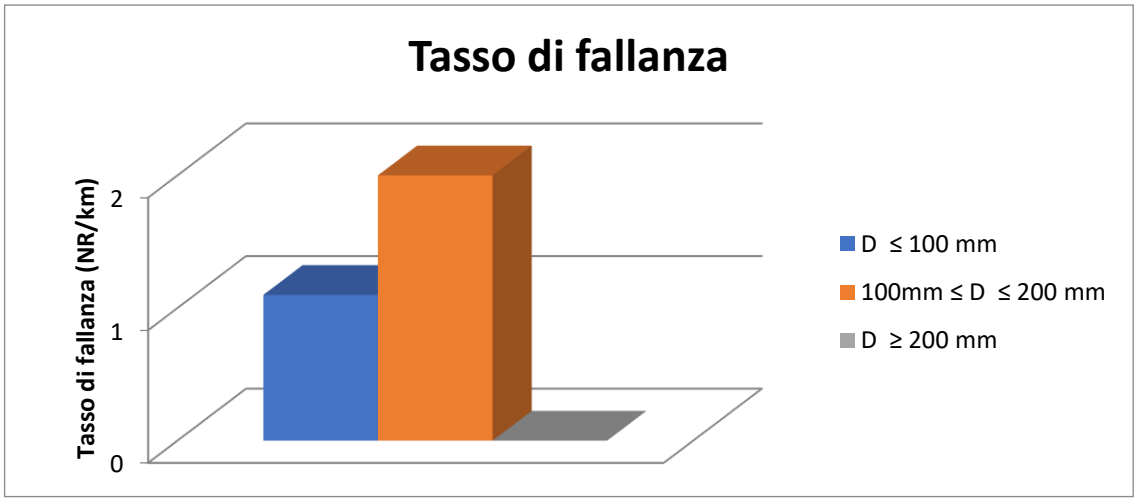


Figura 6.28. Tasso di fallanza delle classi di diametro della tabella *ReteTorinoPEAD*

Il PEAD è un materiale utilizzato per condotte limitate in lunghezza e, come espressamente detto in precedenza, tali lunghezze di frequente risultano minori di 10 metri. Per tale motivo, molte informazioni su questo materiale sono state eliminate: basti pensare al numero di guasti del PEAD nella tabella *Wwvcondottetorino* e confrontarlo con quello ricavato nella tabella *ReteTorino* (9). Le informazioni che si ricavano sono quindi poco significative. Il tasso di rottura ha un trend non costante all'aumentare del diametro.

La stessa procedura è stata adottata per le classi relative all'anno di posa.

Tabella 6.32. Calcolo del tasso di fallanza delle classi di posa presenti nella tabella *ReteTorinoPEAD*

Anno di posa		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**1970 - 1990**	1.3	35.1	4	50	3.1
CLASSE 2:	**1990 - 2010**	2.3	62.2	4	50	1.7
CLASSE 3:	**2010 →**	0.1	2.7	0	0	0.0
	TOT	3.7	100	8	100	

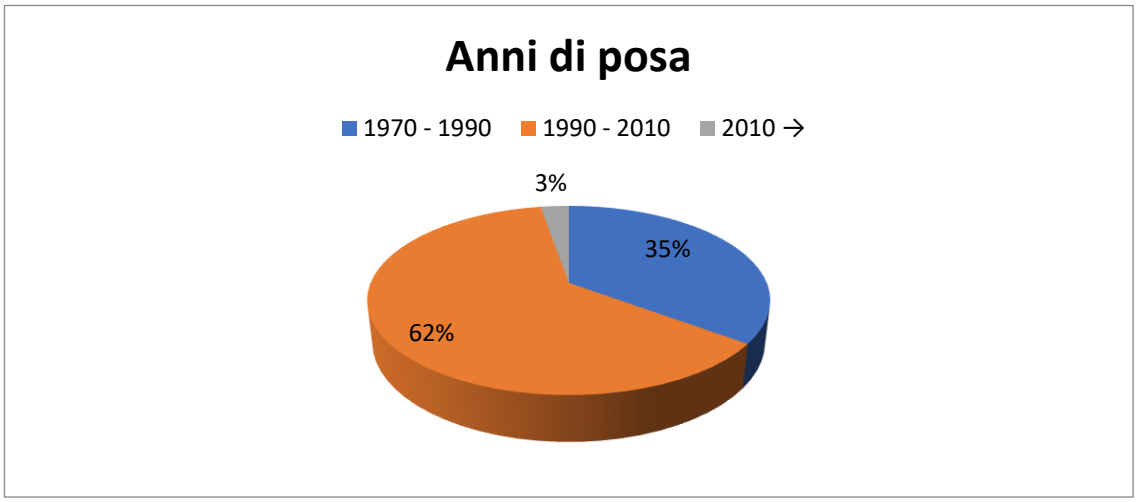


Figura 6.29. Composizione in anni di posa della tabella *ReteTorinoPEAD*

Così come la ghisa sferoidale, anche il PEAD è un materiale di recente tecnologia e per questo motivo non risultano condotte posate prima del 1970.
È doveroso sottolineare che questa informazione sull'età è stata scarsamente riportata dagli operatori nel tempo.
I grafici mostrano che le condotte meno recenti presentano tassi di fallanza più elevati.

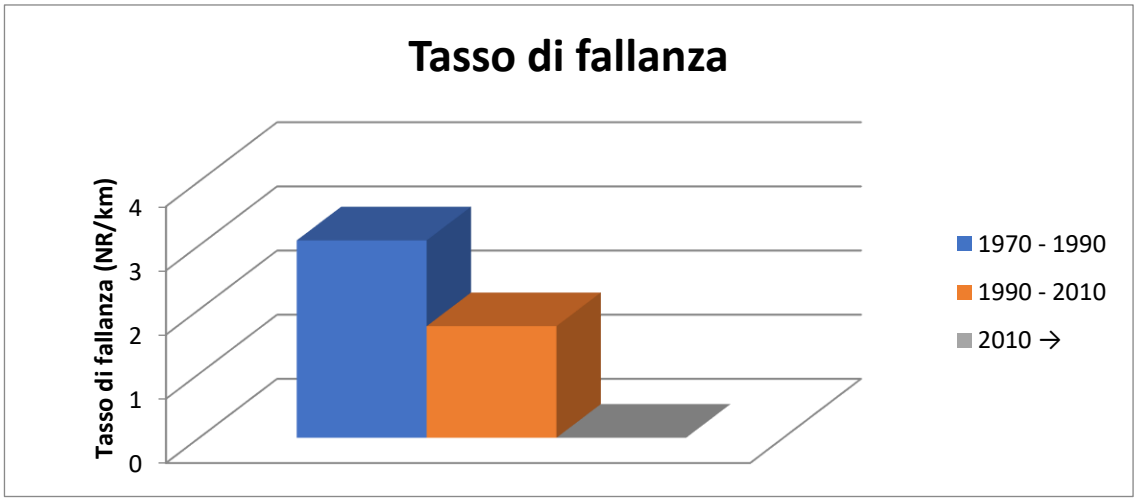


Figura 6.30. Tasso di fallanza delle singole classi di posa della tabella *ReteTorinoPEAD*

La stessa procedura è stata adottata per le classi relative al carico massimo.

Tabella 6.33. Calcolo del tasso di fallanza delle classi relative al carico massimo delle condotte presenti nella tabella *ReteTorinoPEAD*

Carico massimo [m]		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**(0, 40]**	0.43	5.58	1	11.1	2.3
CLASSE 2:	**(40, 80]**	6.74	87.42	4	44.4	0.6
CLASSE 3:	**(80, 120]**	0.23	2.98	2	22.2	8.7
CLASSE 4:	**> 160**	0.31	4.02	2	22.2	6.5
	TOT	7.71	100	9	100	

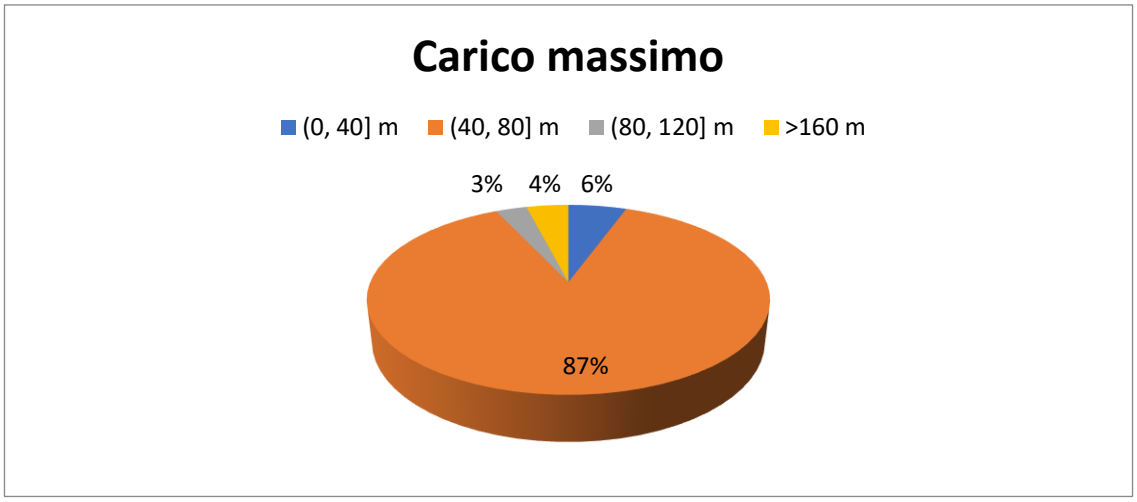
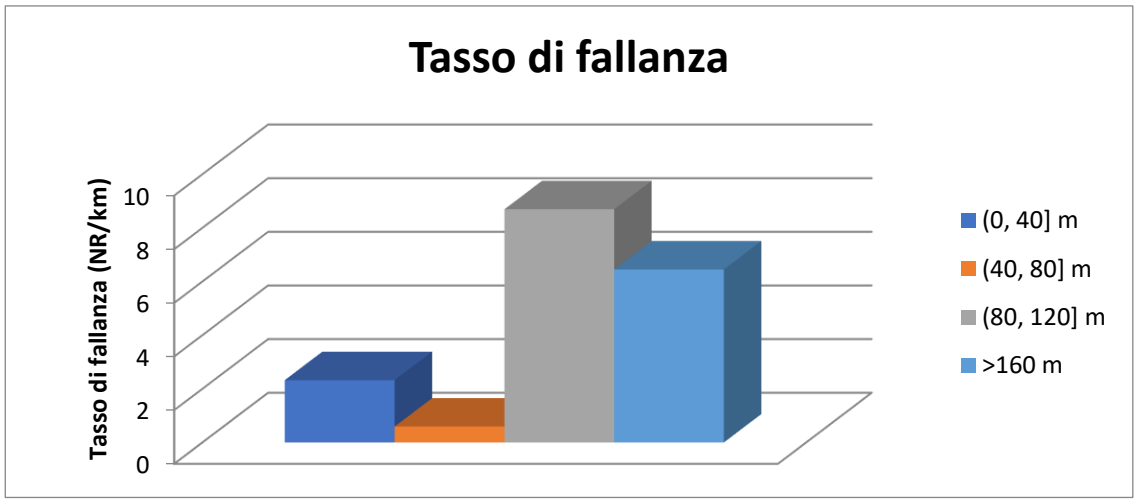


Figura 6.31. Composizione classi di carico delle condotte della tabella *ReteTorinoPEAD*

Il tasso di rottura ritrova un picco in corrispondenza di carichi massimi tra 80 e 120 metri. L'87% delle condotte è caratterizzato da condizioni di carico massimo comprese tra 40 e 80 metri.



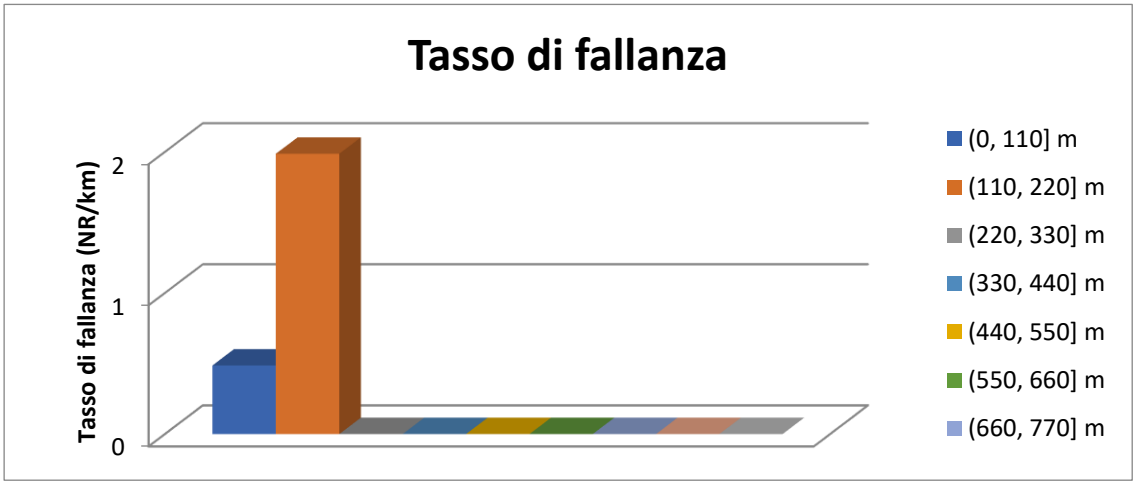
6.32. Tasso di fallanza delle classi di carico delle condotte della tabella *ReteTorinoPEAD*

La stessa procedura è stata adottata per le classi relative alle lunghezze delle condotte.

Tabella 6.34. Calcolo del tasso di fallanza delle classi relative alla lunghezza delle condotte presenti nella tabella *ReteTorinoPEAD*

Lunghezza (m)		Lunghezze (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**(0, 110]**	6.2	45.6	3	33.3	0.5
CLASSE 2:	**(110, 220]**	3.0	22.3	6	66.7	2.0
CLASSE 3:	**(220, 330]**	0.8	5.8	0	0.0	0.0
CLASSE 4:	**(330, 440]**	0.4	3.0	0	0.0	0.0
CLASSE 5:	**(440, 550]**	0.0	0.0	0	0.0	/
CLASSE 6:	**(550, 660]**	0.6	4.3	0	0.0	0.0
CLASSE 7:	**(660, 770]**	0.0	0.0	0	0.0	/
CLASSE 8:	**(770, 880]**	1.6	11.8	0	0.0	0.0
CLASSE 9:	**(880, 990]**	1.0	7.2	0	0.0	0.0
	TOT	13.5	100	9	100	

Si nota nuovamente come il PEAD caratterizzi principalmente condotte con limitate lunghezze. I guasti, inoltre, si concentrano nelle condotte più corte per il 100%.



7.33. Tasso di fallanza delle classi di lunghezza delle condotte della tabella *ReteTorinoPEAD*

Tabella *ReteTorinoEternit*

Come nel caso precedente, questa tabella è stata ottenuta come estrazione dalla tabella *ReteTorino*.
Si analizzano i campi diametro, anno di posa, carico massimo e lunghezza.

Tabella 6.35. Calcolo del tasso di fallanza delle classi di diametro presenti nella tabella *ReteTorinoEternit*

Diametro (mm)		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**D ≤ 100**	25.7	39.9	268	76.4	10.4
CLASSE 2:	**100mm ≤ D ≤ 200**	9.15	14.2	74	21.1	8.1
CLASSE 3:	**D ≥ 200**	29.5	45.9	9	2.6	0.3
	TOT	64.37	100	351	100	

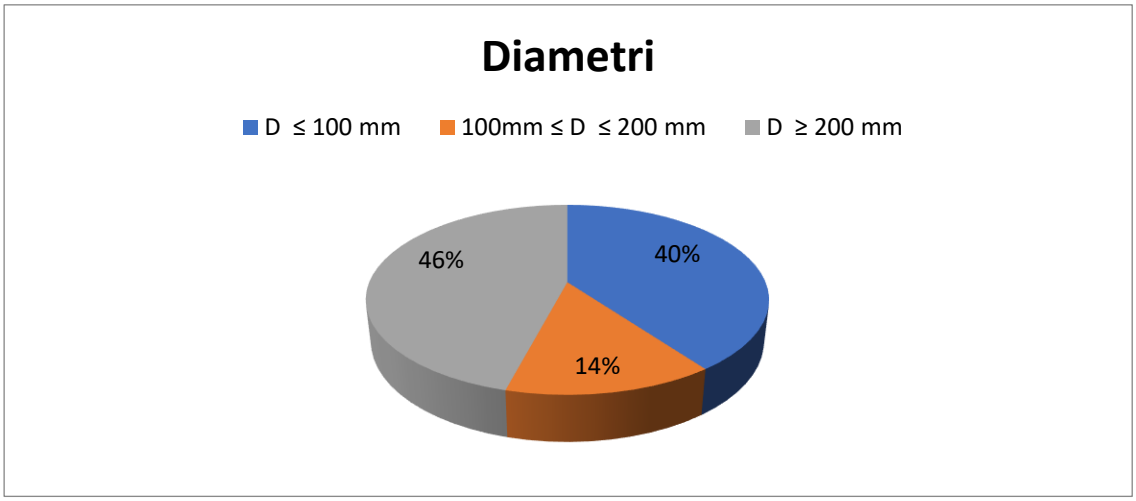


Figura 6.34. Composizione in diametri della tabella *ReteTorinoEternit*

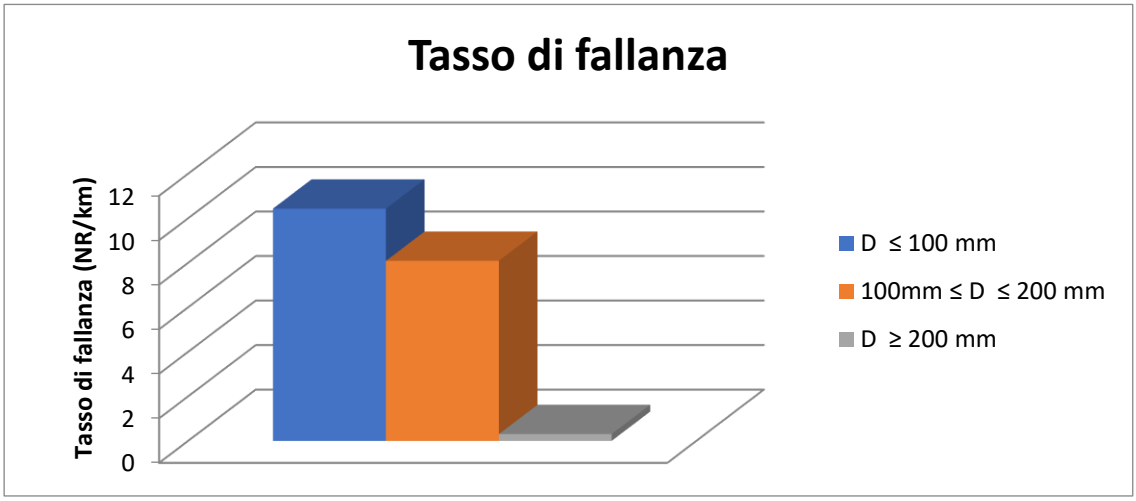


Figura 6.35. Tasso di fallanza delle classi di diametro della tabella *ReteTorinoEternit*

Il tasso di rottura diminuisce all'aumentare del diametro. La quasi totalità delle condotte è caratterizza da diametri inferiori ai 200 millimetri.
La stessa procedura è stata adottata per le classi relative all'anno di posa.

Tabella 6.36. Calcolo del tasso di fallanza delle classi di posa presenti nella tabella *ReteTorinoEternit*

Anno di posa		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**1950 - 1970**	0.3	19.2	5	0.0	17.9
CLASSE 2:	**1970 - 1990**	0.45	30.8	10	5.1	22.2
CLASSE 3:	**1990 - 2010**	0.45	30.8	0	83.1	0
CLASSE 4:	**2010 →**	0.28	19.2	0	11.8	0
	TOT	1.46	100	15	100	

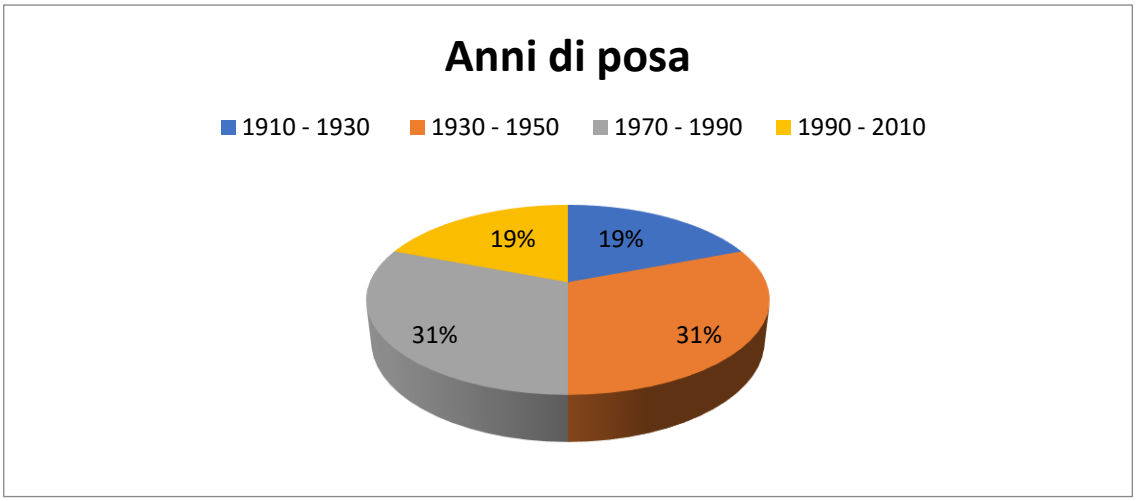


Figura 6.36. Composizione in anni di posa della tabella *ReteTorinoEternit*

Una limitata percentuale di condotte in eterni è provvista dell'informazione riguardante l'anno di posa. Inoltre, ne è stato vietato l'utilizzo a partire dal 1992. I tassi più elevati caratterizzano le condotte più vecchie.

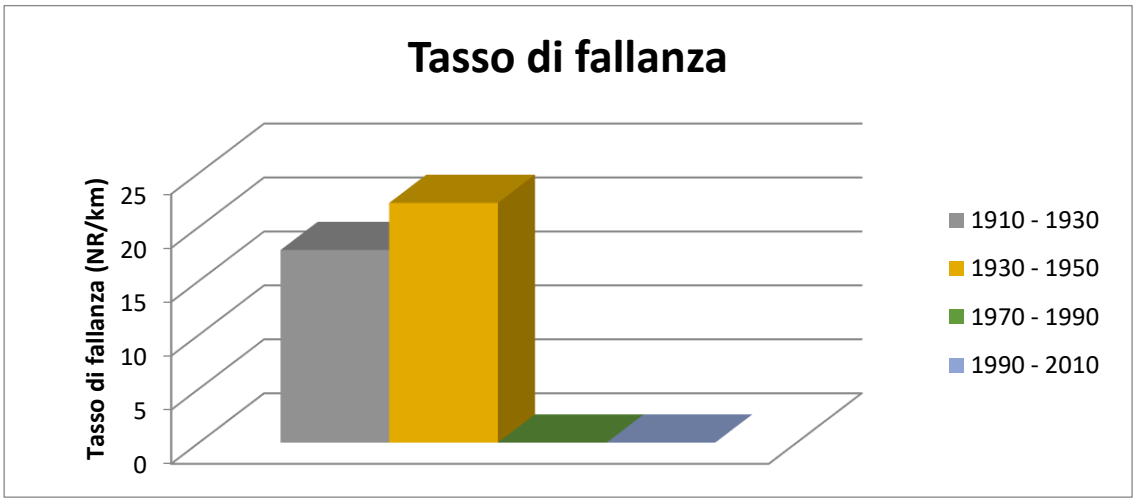


Figura 6.37. Tasso di fallanza delle singole classi di posa della tabella *ReteTorinoEternit*

La stessa procedura è stata adottata per le classi relative al carico massimo.

Tabella 6.37. Calcolo del tasso di fallanza delle classi relative al carico massimo delle condotte presenti nella tabella *ReteTorinoEternit*

Carico massimo (m)		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**(0, 40]**	13.1	27.4	48	13.7	3.7
CLASSE 2:	**(40, 80]**	34.4	72.0	303	86.3	8.8
CLASSE 3:	**> 80**	0.32	0.6	0	0	0
	TOT	47.76	100	351	100	

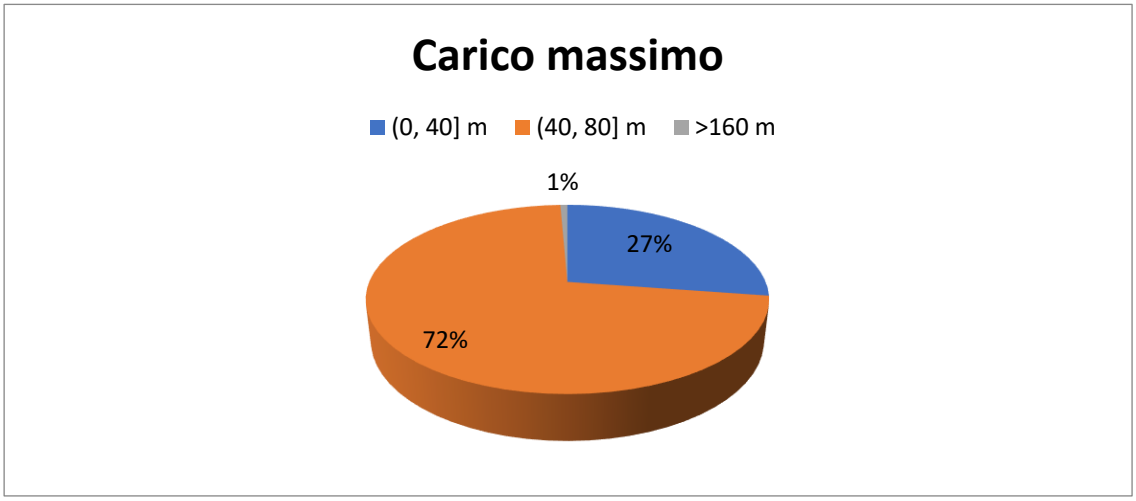
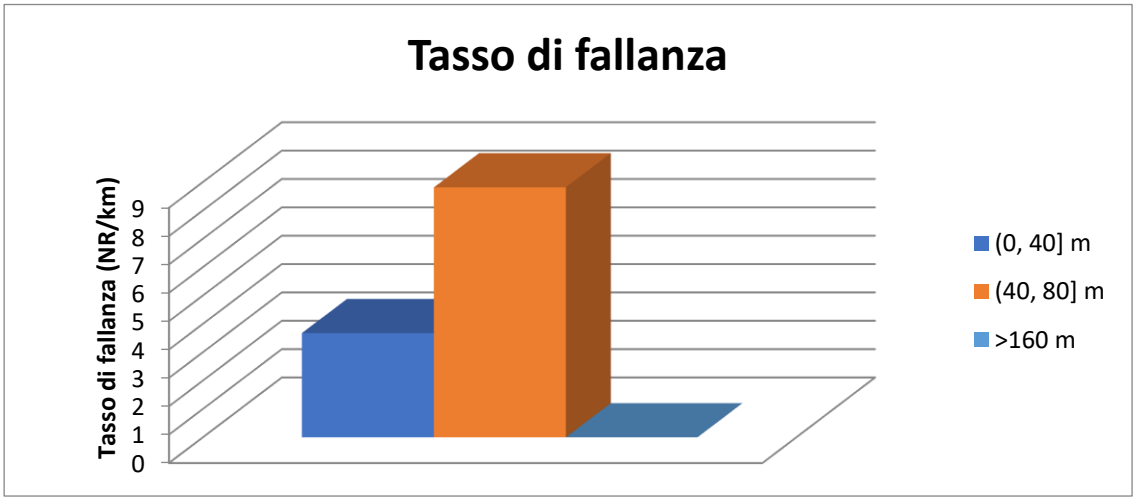


Figura 6.38. Composizione classi di carico delle condotte della tabella *ReteTorinoEternit*

Il tasso di rottura ritrova un picco in corrispondenza di carichi massimi tra 40 e 80 metri. Nonostante ciò, gran parte delle condotte in Eternit è utilizzata in presenza di carichi massimi inferiori agli 80 metri.



6.39. Tasso di fallanza delle classi di carico delle condotte della tabella *ReteTorinoEternit*

La stessa procedura è stata adottata per le classi relative alle lunghezze delle condotte.

Tabella 6.38. Calcolo del tasso di fallanza delle classi relative alla lunghezza delle condotte presenti nella tabella *ReteTorinoEternit*

Lunghezza		Lunghezze (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**(0, 300]**	166.8	56.5	347	98.9	2.1
CLASSE 2:	**(300.600]**	61.6	20.9	4	1.1	0.1
CLASSE 3:	**(600, 900]**	25.8	8.7	0	0	0
CLASSE 4:	**(900, 1200]**	22.1	8.7	0	0	0
CLASSE 5:	**(1200, 1500]**	8.2	2.8	0	0.0	0.0
CLASSE 6:	**(1500, 1800]**	6.6	2.2	0	0.0	0.0
CLASSE 7:	**(1800,2100]**	0.0	0.0	0	0.0	/
CLASSE 8:	**(2100,2400]**	4.3	1.4	0	0.0	0.0
	TOT	295.4	100	177	100	

L'eternit caratterizza condotte limitate in lunghezza: infatti, oltre il 56% ha una lunghezza inferiore a 300 metri.
Il tasso di fallanza diminuisce all'aumentare della lunghezza.

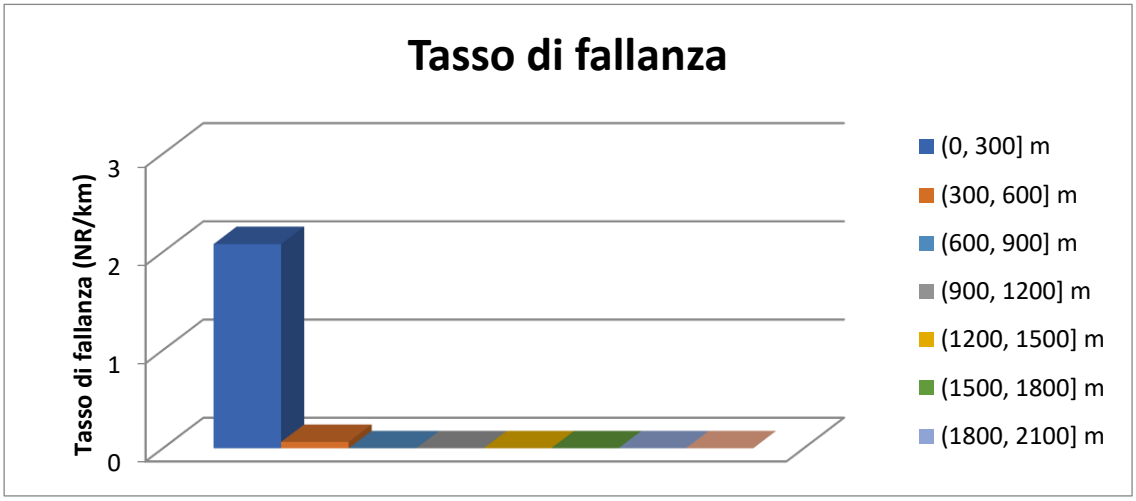


Figura 6.40. Tasso di fallanza delle classi di lunghezza delle condotte della tabella *ReteTorinoEternit*

Tabella *ReteTorinoAcciaio*

Si analizzano di seguito i campi diametro, anno di posa, carico massimo e lunghezza.

Tabella 6.39. Calcolo del tasso di fallanza delle classi di diametro presenti nella tabella *ReteTorinoAcciaio*

Diametro (mm)		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**D ≤ 100**	21.7	8.2	61	35.3	2.8
CLASSE 2:	**100mm ≤ D ≤ 200**	24.4	9.2	29	16.8	1.2
CLASSE 3:	**D ≥ 200**	218.1	82.5	83	48	0.4
	TOT	264.2	100	173	100	

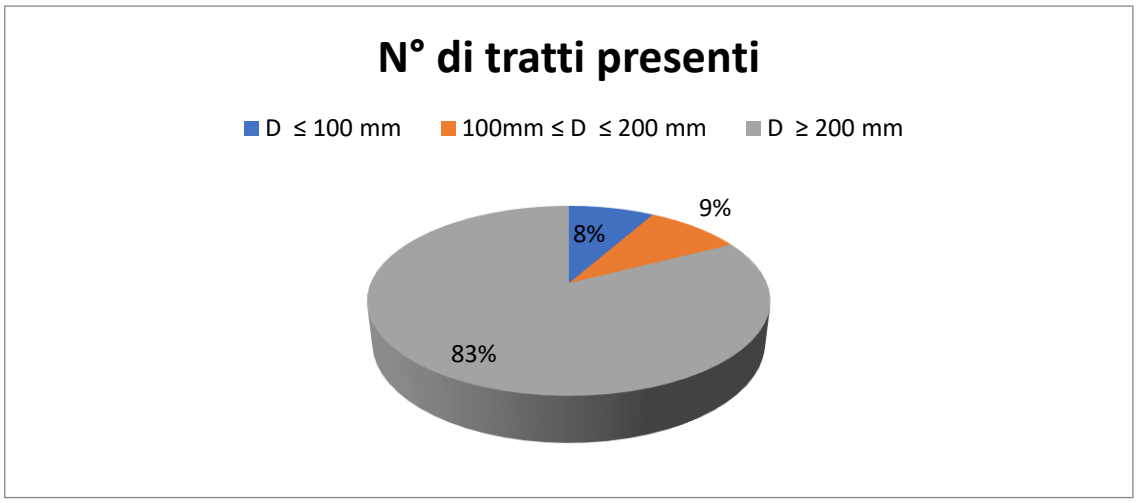


Figura 6.41. Composizione in diametri della tabella *ReteTorinoAcciaio*

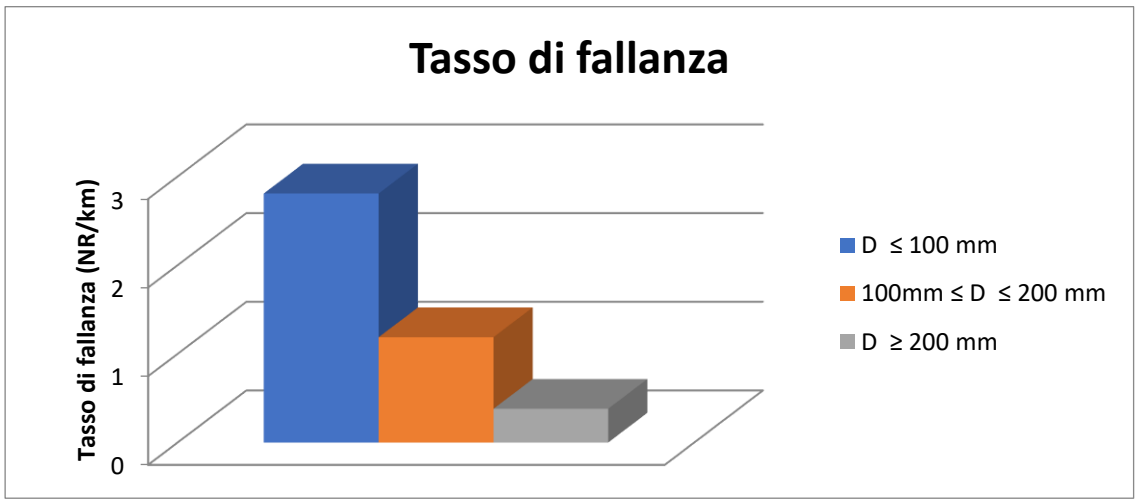


Figura 6.42. Tasso di fallanza delle classi di diametro della tabella *ReteTorinoAcciaio*

La maggior parte delle condotte in acciaio ha un diametro superiore ai 200 millimetri. Anche in questo caso il tasso di rottura diminuisce all'aumentare del diametro.
La stessa procedura è stata adottata per le classi relative all'anno di posa.

Tabella 6.41. Calcolo del tasso di fallanza delle classi di posa presenti nella tabella *ReteTorinoAcciaio*

Anno di posa		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**(1890-1910]**	0.01	0	0	0	0
CLASSE 2:	**(1910-1930]**	0.32	0.5	0	0.0	0.0
CLASSE 3:	**(1950-1970]**	5.18	8.3	7	14.6	1.4
CLASSE 4:	**(1970-1990]**	11.5	18.4	3	6.3	0.3
CLASSE 5:	**(1990-2010]**	41.88	66.9	38	79.2	0.9
CLASSE 6:	**2010→**	3.71	5.9	0	0.0	0.0
	TOT	62.6	100	48	100	

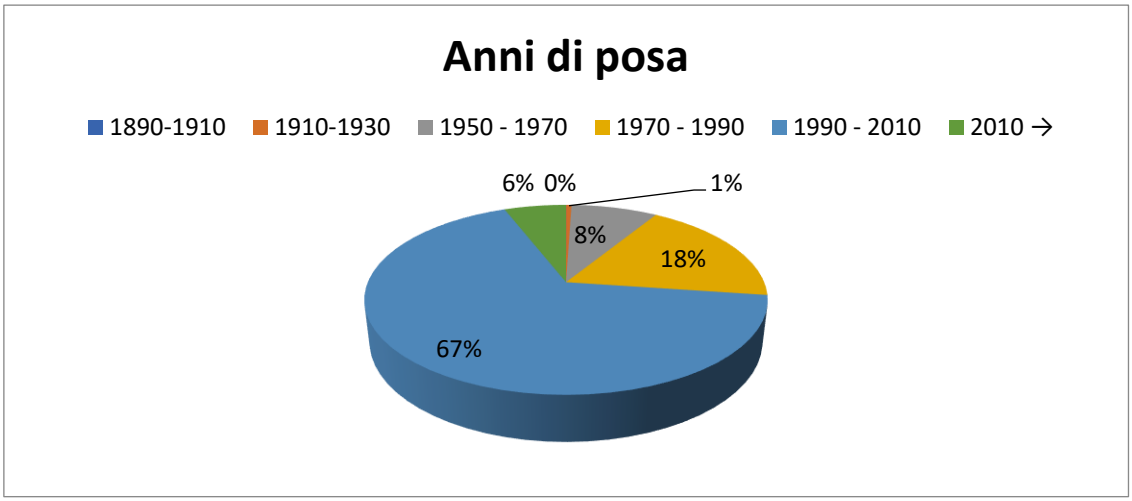


Figura 6.43. Composizione in anni di posa della tabella *ReteTorinoAcciaio*

L'acciaio è un materiale posato per la maggiore tra gli anni 1990 e 2010. Questo dato probabilmente fornisce un'indicazione sulla migliore attenzione da parte degli operatori nel riportare tale informazione in tempi recenti. Il tasso di fallanza più elevato si riscontra per le condotte posate tra il 1950 e il 1970.

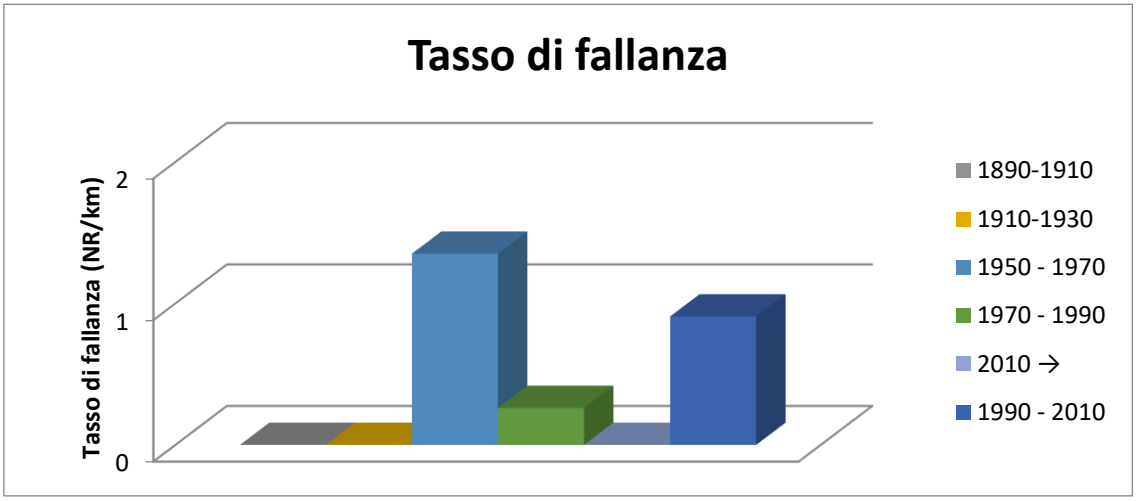


Figura 6.44. Tasso di fallanza delle singole classi di posa della tabella *ReteTorinoAcciaio*

La stessa procedura è stata adottata per le classi relative al carico massimo.

Tabella 6.42. Calcolo del tasso di fallanza delle classi relative al carico massimo delle condotte presenti nella tabella *ReteTorinoAcciaio*

Carico massimo (m)		Lunghezza (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**(0, 40]**	32.4	20.2	19	11.1	0.6
CLASSE 2:	**(40,80]**	112.93	70.4	114	66.7	1.0
CLASSE 3:	**(80. 120]**	4.36	2.72	11	6.4	2.5
CLASSE 4:	**(120,160]**	4.2	2.6	12	7.0	2.9
CLASSE 5:	**> 160**	66.6	4.2	15	8.8	2.3
	TOT	160.5	100	171	100	

I carichi tra i 40 e gli 80 metri caratterizzano il 70% della lunghezza delle condotte. Il tasso di fallanza sembra aumentare all'aumentare del carico massimo, anche se il maggior valore di questo indice si ritrova nella classe 4.

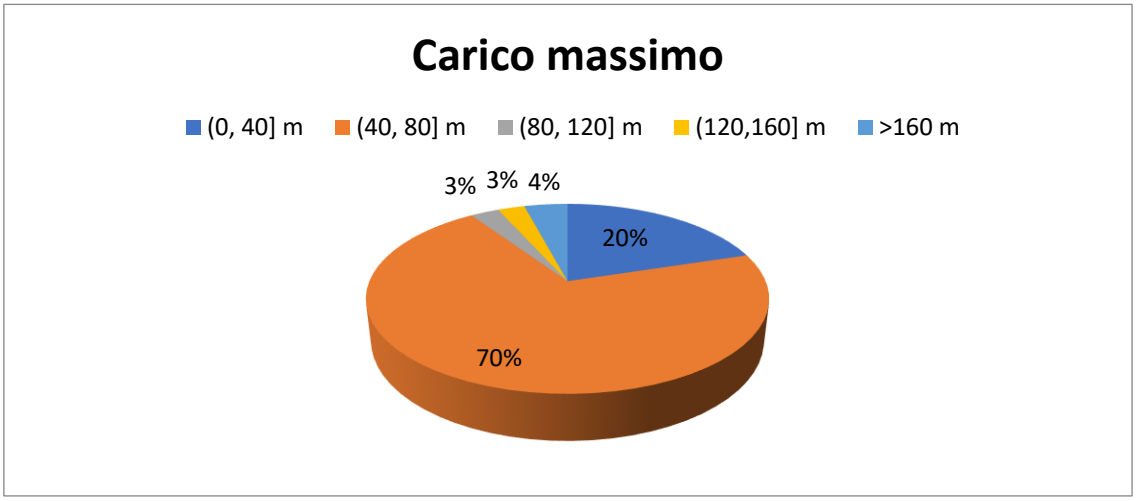
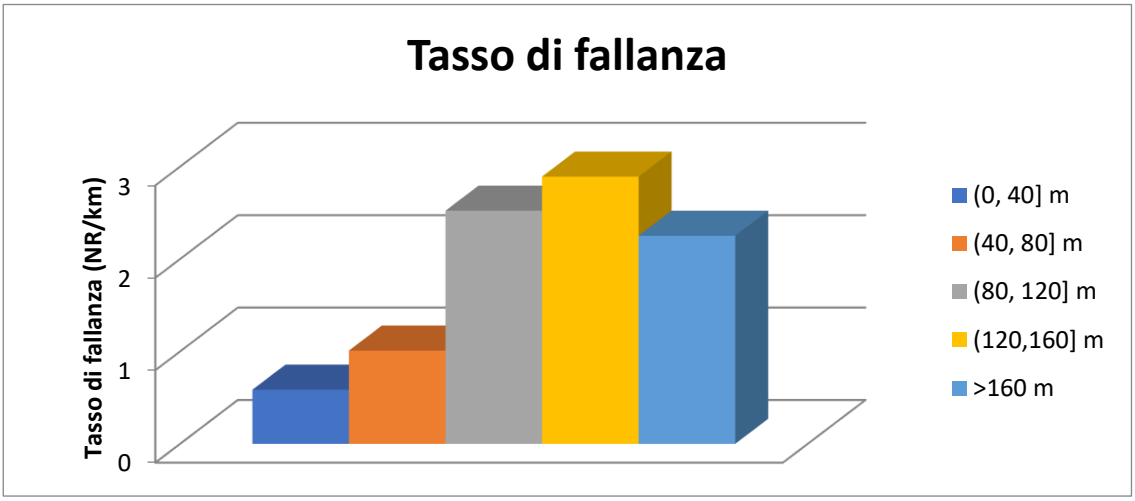


Figura 6.45. Composizione classi di carico delle condotte della tabella *ReteTorinoAcciaio*

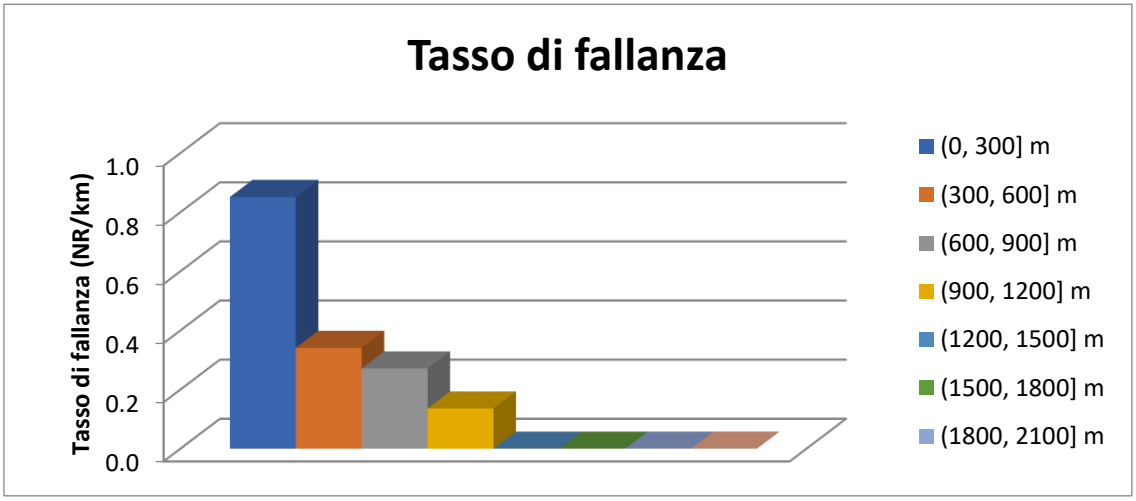


6.46. Tasso di fallanza delle classi di carico delle condotte della tabella *ReteTorinoAcciaio*

La stessa procedura è stata adottata per le classi relative alla lunghezza delle condotte. Oltre la metà della lunghezza totale delle condotte in acciaio presenta lunghezze tra 0 e 300 metri. Un ulteriore 20.9% presenta lunghezze fino a 600 metri.
Il tasso di fallanza diminuisce all'aumentare della lunghezza.

Tabella 6.43. Calcolo del tasso di fallanza delle classi relative alla lunghezza delle condotte presenti nella tabella *ReteTorinoAcciaio*

Lunghezza (m)		Lunghezze (km)	%	N° di guasti	%	Tasso di fallanza (NR/km)
CLASSE 1:	**(0, 300]**	166.8	56.5	142	82.1	0.9
CLASSE 2:	**(300.600]**	61.6	20.9	21	12.1	0.3
CLASSE 3:	**(600, 900]**	25.8	8.7	7	4.0	0.3
CLASSE 4:	**(900, 1200]**	22.1	8.7	3	1.7	0.1
CLASSE 5:	**(1200, 1500]**	8.2	2.8	0	0.0	0.0
CLASSE 6:	**(1500, 1800]**	6.6	2.2	0	0.0	0.0
CLASSE 7:	**(1800,2100]**	0.0	0.0	0	0.0	/
CLASSE 8:	**(2100,2400]**	4.3	1.4	0	0.0	0.0
	TOT	295.4	100	177	100	



6.47. Tasso di fallanza delle classi di lunghezza delle condotte della tabella *ReteTorinoAcciaio*

6.4 Analisi della tabella *Esatta*

Come descritto nel capitolo precedente, la tabella *Esatta* è la tabella di partenza contenente i dati sui quali tarare i modelli di regressione. A differenza della tabella *ReteTorino*, dalla quale questa deriva, contiene le sole condotte caratterizzate da un quadro completo di informazioni nei campi materiale, diametro, anno di posa, carico massimo e lunghezza. Contiene al suo interno 5855 condotte e 106 guasti.

Si riporta anche in questo caso l'analisi della tabella in esame e il tasso di fallanza associato alle diverse classi.

Questa tabella sarà di fondamentale importanza per valutare un modello statistico e la successiva capacità dello stesso di prevedere un guasto.

Calcolo delle occorrenze dei materiali

In accordo con quanto fatto per la tabella *ReteTorino*, sono stati presi in esame solo i materiali che caratterizzano la rete di Torino per almeno lo 0.5% della sua estensione e nello specifico sono: acciaio, eternit, ghisa grigia, ghisa sferoidale e PEAD.
Per ogni materiale è stata calcolata la lunghezza complessiva delle condotte, il numero di guasti e il relativo tasso di fallanza, inteso come rapporto tra il numero di rotture e i chilometri di condotte caratterizzanti un certo materiale. Si riporta un esempio per chiarire la procedura per il calcolo del tasso di fallanza.

Esempio 1

Si prenda in esame l'acciaio. Attraverso il software *Python* è possibile calcolare il numero di chilometri di condotte per ogni materiale e il numero di guasti. 48.78 chilometri di condotte nella tabella *Esatta* sono in acciaio e tale materiale ha mostrato 5 rotture nel periodo di osservazione. Il tasso di fallanza è dato da 5/48.78 = 0.1.

L'acciaio presenta 0.1 rotture per chilometro di condotta

Tale procedura è stata applicata per ogni materiale.
Si riportano i risultati in **Tabella 6.44**. Da questa analisi si ricava che il materiale predominante nella rete è la ghisa sferoidale, in contrasto con quanto ritrovato nella tabella *ReteTorino*, che riportava la ghisa grigia come materiale con maggiore presenza in rete. In realtà, quest'ultimo presentava molte condotte prive di informazioni complete, specialmente per il campo relativo all'anno di posa. Questa informazione è stata riportata con maggiore frequenza negli ultimi anni e, di conseguenza, i materiali più recenti come la ghisa sferoidale presentano una maggiore copertura di tale campo.
Per quanto concerne il tasso di fallanza, eternit e ghisa grigia presentano i valori maggiori.

Tabella 6.44. Calcolo del tasso di fallanza dei materiali della tabella *Esatta*

Materiale	N° assegnato	Lunghezza (km)	N° di guasti	Tasso di fallanza (NR/km)
Acciaio	1	48.78	5	0.10
Eternit	4	1.72	10	5.8
Ghisa grigia	6	27.75	59	2.1
Ghisa sferoidale	7	224.69	32	0.1
PEAD	8	3.59	0	0.0
TOT	**/**	306.62	106	**/**

Si riportano di seguito l'istogramma relativo ai tassi di fallanza.

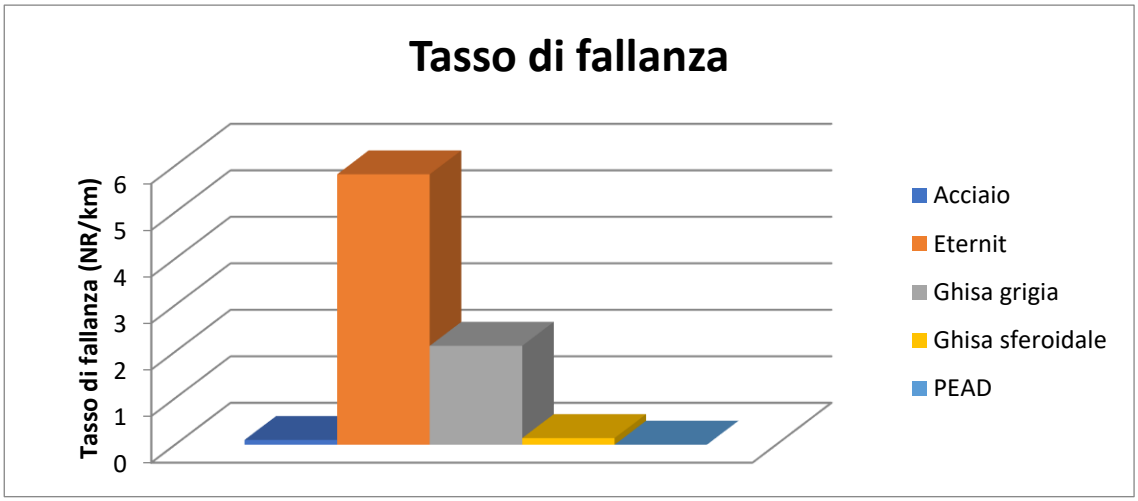


Figura 6.48. Tasso di fallanza dei materiali

Calcolo delle occorrenze dei diametri

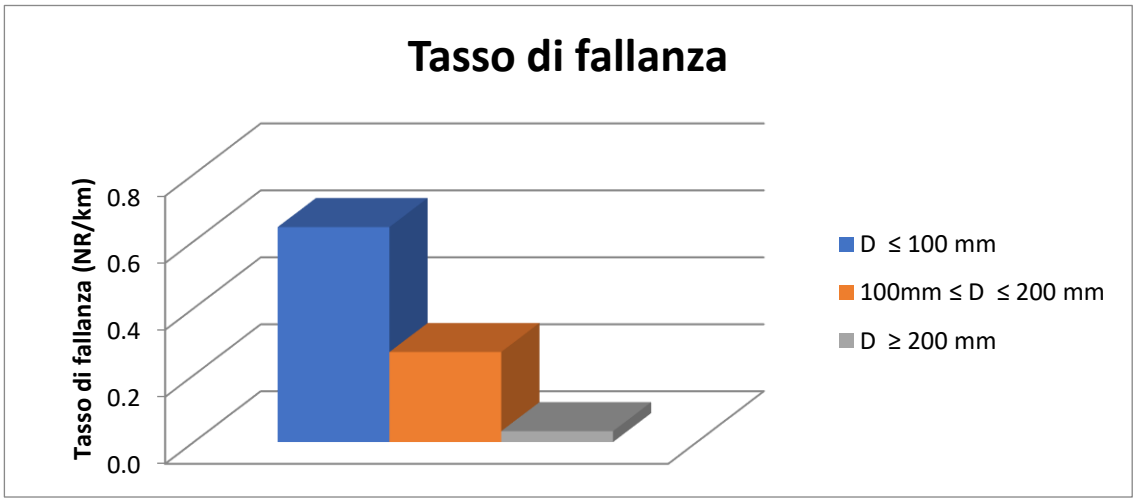
Come per i materiali, è stato valutato il tasso di fallanza di 3 diverse classi di diametri, attraverso il software *Python*.

Tabella 6.45. Calcolo delle occorrenze dei diametri della tabella *Esatta*

Diametro		Lunghezza (km)	N° di guasti	Tasso di fallanza (NR/km)
CLASSE 1:	**D ≤ 100 mm**	121.45	78	0.6
CLASSE 2:	**100mm ≤ D ≤ 200 mm**	92.92	25	0.3
CLASSE 3:	**D ≥ 200 mm**	92.25	3	0.03
	TOT	306.62	106	**/**

Si riporta il grafico relativo in **Figura 6.48.**

Figura 6.49. Tasso di fallanza delle classi di diametri



Come riscontrato nella tabella *ReteTorino,* i diametri più piccoli vanno incontro a rotture più frequentemente. Inoltre, gran parte della tabella *Esatta* presenta diametri inferiori a 200 millimetri.

Calcolo delle occorrenze per gli anni di posa

Come per i materiali, è stato valutato il tasso di fallanza di 8 diverse classi di età, attraverso il software *Python*.

Tabella 6.46. del tasso di fallanza delle classi di età della tabella *Esatta*

Anno di posa		Lunghezza (km)	N° di guasti	Tasso di fallanza (NR/km)
CLASSE 1:	**1870 - 1890**	1.11	1	0.9
CLASSE 2:	**1890 - 1910**	11.71	21	1.8
CLASSE 3:	**1910 - 1930**	7.59	13	1.7
CLASSE 4:	**1930 - 1950**	3.2	14	4.4
CLASSE 5:	**1950 - 1970**	5.85	11	1.9
CLASSE 6:	**1970 - 1990**	33.11	6	0.2
CLASSE 7:	**1990 - 2010**	227.44	39	0.2
CLASSE 8:	**2010 →**	16.21	1	0.1
	TOT	306.22	106	**/**

La maggior parte delle condotte dotate dell'informazione sull'età è posteriore al 1990, a riprova che tale dato è stato riportato con maggiore frequenza solo recentemente.
Si riporta in **Figura 6.50** il tasso di fallanza delle diverse classi.

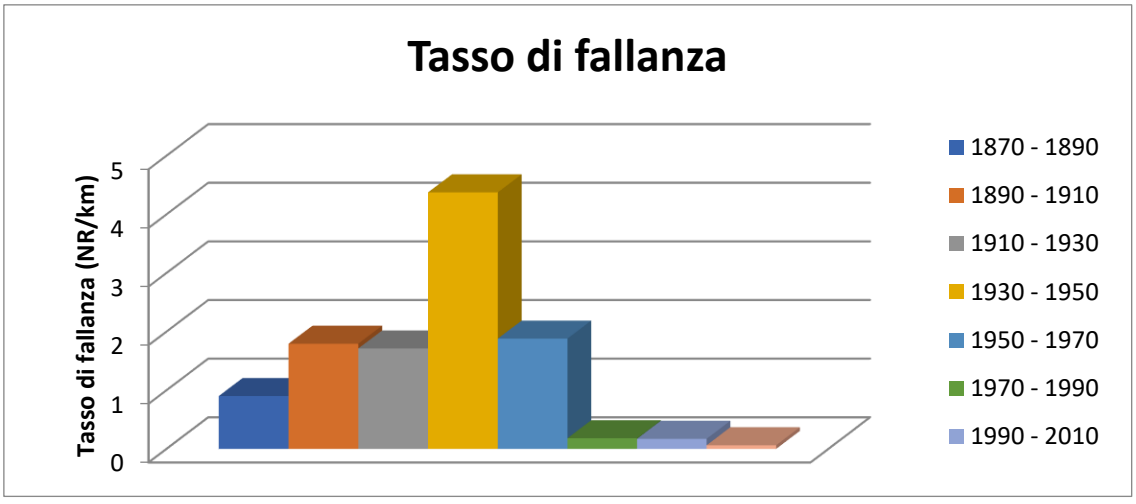


Figura 6.50. Occorrenze degli anni di posa presenti nella tabella *Esatta,* percentuale dei guasti rispetto al numero totale di guasti e percentuale dei guasti rispetto al numero totale di condotte di ogni classe

Il biennio 1930-1950 presenta il tasso di fallanza maggiore, pari a 4.4 rotture per chilometro.

Calcolo delle occorrenze per i carichi massimi

Anche nel caso del carico massimo è stato valutato il tasso di fallanza per 7 differenti classi.

Tabella 6.47. Calcolo delle occorrenze dei carichi massimi della tabella *Esatta*

Carico massimo		Lunghezza (km)	N° di guasti	Tasso di fallanza (NR/km)
CLASSE 1:	**(0, 40] m**	55.96	10	0.2
CLASSE 2:	**(40, 80] m**	226.9	88	0.4
CLASSE 3:	**(80, 120] m**	6.38	2	0.3
CLASSE 4:	**(120,160] m**	4.04	1	0.2
CLASSE 5:	**(160,200] m**	4.43	1	0.2
CLASSE 6:	**(200,240] m**	3.23	2	0.6
CLASSE 7:	**>240 m**	5.68	2	0.4
	TOT	306.62	106	**/**

La maggior parte delle condotte è caratterizzata da un carico massimo compreso tra 40 e 80 metri. In corrispondenza di questa classe, inoltre, si trova il maggior numero di guasti (l'83% dei guasti totali contenuti nella tabella *Esatta*).

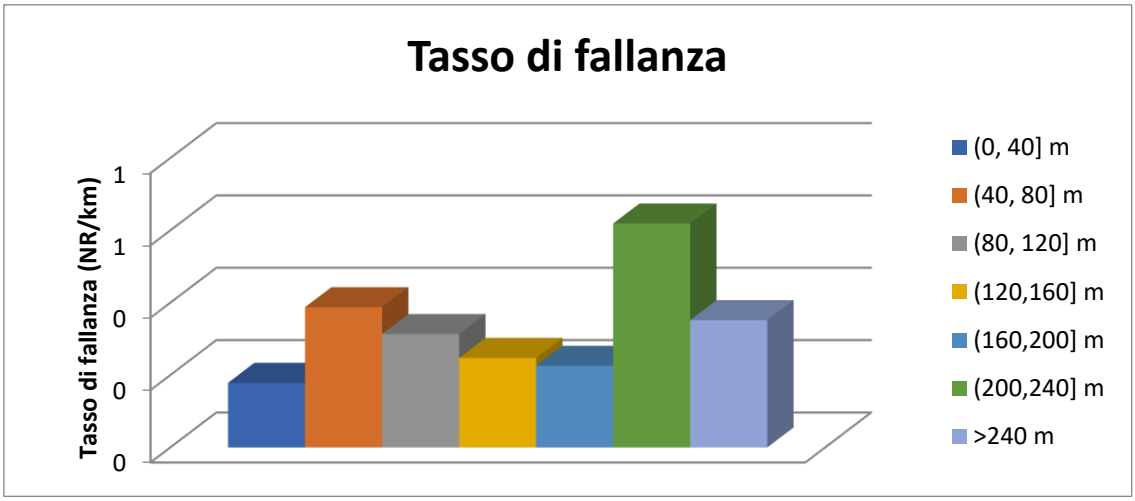


Figura 6.51. Tasso di fallanza delle diverse classi di carico massimo

Il tasso di fallanza più elevato si ritrova nella classe di carico compresa tra i 200 e i 240 metri.

Calcolo delle occorrenze per le classi di lunghezze

Infine, si valutano le occorrenze dei diversi campi in relazione alle lunghezze delle condotte. Sono state adottate 6 classi di lunghezze, come riportato in **Tabella 6.48**.

Tabella 6.48. Calcolo delle occorrenze delle lunghezze della tabella *Esatta*

Lunghezza		Lunghezza complessiva (km)	N° di guasti	Tasso di fallanza (NR/km)
CLASSE 1:	**(0, 150] m**	234.5	96	0.4
CLASSE 2:	**(150, 300] m**	43.7	7	0.2
CLASSE 3:	**(300, 450] m**	12.71	1	0.1
CLASSE 4:	**(450, 600] m**	8.99	2	0.2
CLASSE 5:	**(600, 750] m**	4.76	0	0.0
CLASSE 6:	**> 750 m**	1.56	0	0.0
	TOT	306.22	106	**/**

La quasi totalità delle condotte è caratterizzata da lunghezze inferiori a 150 metri. Il 90.6% dei guasti contenuti nella tabella *Esatta* fanno riferimento alla classe 1. Inoltre, all'aumentare della lunghezza, i guasti diminuiscono in quasi tutti i casi.
Questi risultati sono coerenti con quanto ritrovato dall'analisi della tabella *ReteTorino*.

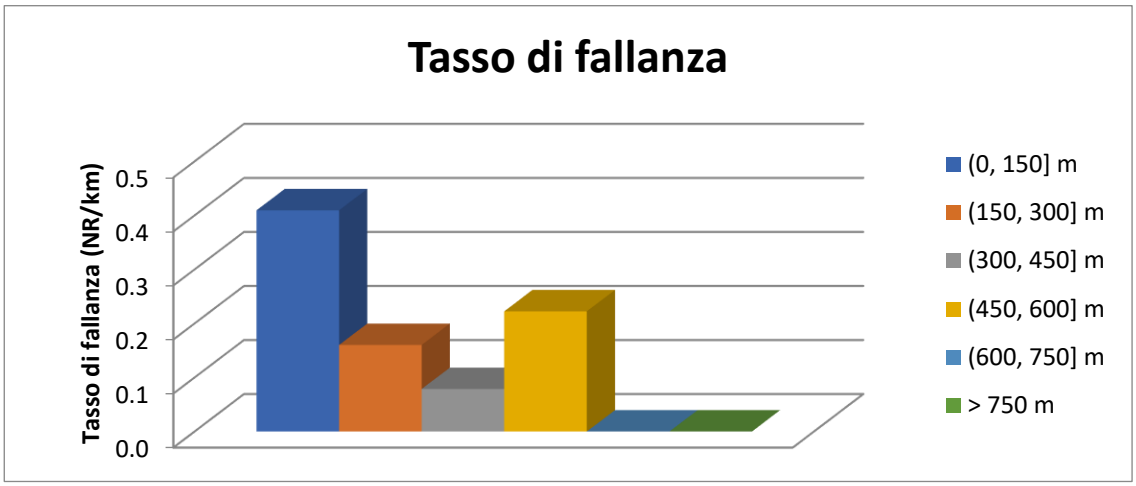


Figura 6.52. Tasso di fallanza delle diverse classi di lunghezza

6.5 Conclusioni

In questo capitolo sono state analizzate le informazioni contenute nelle tabelle fondamentali *Wwvcondottetorino, ReteTorino* e *Esatta*.
Per quanto concerne la prima tabella, in ordine, sono stati percorsi i seguenti passi:

- calcolo delle occorrenze per tutti i campi della tabella;
- calcolo delle occorrenze nei mesi dell'anno;
- calcolo delle occorrenze nei mesi dell'anno e nelle diverse fasce di temperatura per ogni materiale preso singolarmente,
- ricerca di eventuali correlazioni tra i parametri attraverso il calcolo dell'indice di *Pearson* e definizione dei relativi grafici a bolle.

A seguito di questa analisi è possibile affermare che i diametri che presentano il maggior numero di rotture risultano quelli tra i 100 e i 150 millimetri. Normalizzando tali valori rispetto ai chilometri di condotte, i più vulnerabili risultano quelli tra 13 e 40 millimetri.
Allo stesso modo, la ghisa grigia è il materiale che presenta il maggior numero di guasti ma, normalizzando le occorrenze rispetto ai chilometri di condotta, i materiali più vulnerabili risultano l'eternit e il polietilene.
Gli interventi di ripristino comportano, nella maggior parte dei casi, un'interruzione di servizio inferiore a 5 ore, in assenza di pericolo.
Il maggior numero di rotture si osserva in corrispondenza dei carichi tra 30 e 60 metri e per condotte posate tra il 1990 e il 2010. Come già affermato in precedenza, però, il dato relativo all'anno di posa è spesso mancante (nel 91% dei casi).
Per quanto riguarda le temperature, la ghisa grigia presenta più guasti nei mesi invernali, mentre polietilene ad alta densità, eternit ed acciaio sono più vulnerabili al di sopra dei 20°C.
Infine, il calcolo degli indici di correlazione non segnala particolari relazioni tra le variabili: sarà quindi necessario prendere in considerazione tutti i campi nello sviluppo del modello di regressione.
Successivamente è stato calcolato il tasso di fallanza delle diverse classi di diametri, materiali, anni di posa, carichi massimi e lunghezze della tabella *ReteTorino*. Si è concluso che:

- il materiale con il tasso di fallanza più elevato è l'eternit con 5.5 rotture per chilometro di condotta;
- il tasso di fallanza aumenta al diminuire del diametro;
- l'informazione relativa all'anno di posa delle condotte è spesso mancante e le indicazioni fornite dalla sua analisi variano da caso a caso;
- in generale, il tasso di fallanza aumenta all'aumentare del carico massimo in condotta;

- il tasso di fallanza diminuisce all'aumentare della lunghezza delle condotte. La classe più vulnerabile di condotte è quella relativa alle lunghezze tra 0 e 400 metri con 2.8 rotture su chilometro;
- l'analisi del tasso di rottura, una volta estratte le tabelle relative ai diversi materiali presenti, conduce alle stesse conclusioni ricavate dall'analisi della tabella *ReteTorino* nella sua complessità.

Infine, è stata analizzata la tabella *Esatta*, ottenuta dalla tabella *ReteTorino*, una volta prese in considerazione le sole condotte dotate di tutte le informazioni necessarie all'applicazione dei modelli di regressione.
Nello specifico, è stata calcolata la lunghezza complessiva e il tasso di fallanza delle diverse classi di diametri, materiali, anni di posa, carichi massimi e lunghezze. Si è concluso che:

- il materiale con il tasso di fallanza più elevato è l'eternit con 5.8 rotture per chilometro;
- il tasso di fallanza aumenta al diminuire del diametro;
- l'informazione relativa all'anno di posa delle condotte è spesso mancante e le indicazioni fornite dalla sua analisi variano da caso a caso;
- il tasso di rottura non sembra seguire una legge specifica all'aumentare del carico massimo;
- in 3 casi su 4, il tasso di fallanza diminuisce all'aumentare della lunghezza delle condotte. La classe di lunghezze fino ai 150 metri presenta il tasso di fallanza più elevato e pari a 0.4 rotture per chilometro.

Capitolo 7

Applicazione del modello di regressione logistica

Nel seguente capitolo verranno descritti i risultati ottenuti dall'applicazione del modello di regressione logistica, descritto in precedenza, alla tabella *Esatta Finale*, contenente tutte le condotte con la totale copertura delle informazioni riguardanti diametro, materiale, anno di posa, carico massimo e lunghezza totale delle stesse.
Come ampiamente illustrato, il modello di regressione logistica applicato a questo caso studio permette di definire la probabilità di rottura per ogni condotta, a partire da una serie di variabili indipendenti che caratterizzano le unità della rete. Nello specifico, la variabile dipendente dicotomica *Y* assume valore unitario in presenza di rottura e valore nullo in assenza di non-rottura, mentre le variabili esplicative indipendenti per ogni condotta sono:

- diametro;
- materiale;
- carico massimo;
- anno di posa;
- lunghezza.

Il punto di partenza per la costruzione del modello di regressione logistica è la tabella *Esatta*, costituita da 5855 righe e 12 colonne, tra cui 106 rotture e 5749 condotte mai rotte nell'arco temporale tra il 2006 e il 2016. Di conseguenza, questa presenta una percentuale di rottura pari all'1.8%.
Per una maggiore chiarezza, si riportano di seguito i passi percorsi in questo capitolo per la stima e la verifica del modello di regressione logistica.

1. Sono state realizzate 20 sotto-estrazioni, a partire dalla tabella *Esatta*, e sono stati stimati i parametri di regressione logistica di ognuna, attraverso il software *Gretl*, una volta prese in considerazione le variabili indipendenti *diametro*, *materiale*, *anno di posa*, *carico massimo* e *lunghezza.* In seguito, è stata valutata la significatività di ogni singolo parametro e la bontà di adattamento complessiva del modello. (**Paragrafo 7.1**).
2. Da ogni sotto-estrazione è stato estratto l'80% degli elementi e sono stati stimati nuovamente i parametri del modello (fase di *fitting*). Anche in questo caso, è

stata valutata la significatività di ogni parametro e la bontà di adattamento del modello. Nel caso di non-significatività di un parametro, il modello è stato stimato escludendo tale parametro. I modelli così composti, sono stati applicati al restante 20% delle sotto-estrazioni ed è stata calcolata la probabilità di rottura di ogni condotta facente parte di questo insieme (fase di *testing*). Tali probabilità sono state riportate in grafici a bolle, per valutare la capacità del modello di prevedere una rottura (**Paragrafo 7.2**). Infatti, un modello è in grado di prevedere una rottura se associa probabilità elevate di rottura a condotte realmente guaste e basse probabilità di rottura a condotte integre.

3. È stato definito il valore di soglia (o di *cut-off*) della probabilità, limite al di là del quale una condotta può essere considerata suscettibile di rottura (**Paragrafo 7.3**).

7.1. Realizzazione di 20 estrazioni e stima del modello di regressione logistica

Come appena illustrato, la tabella *Esatta* è caratterizzata da una percentuale rotture pari all'1.8% mentre la tabella *Rete Torino* presenta una percentuale di rotture pari al 15%.
Di conseguenza, sono state definite 20 estrazioni a partire dalla tabella *Esatta Finale* in modo tale che questa percentuale potesse rimanere inalterata, poiché il valore 1.8% è frutto di un filtraggio adottato sui dati e non rispecchia la reale percentuale di rottura.
Per mantenere tale percentuale immutata, sono stati seguiti i seguenti passaggi:

- a partire dalla tabella *Esatta* sono state estratte le 106 condotte con flag pari all'unità;
- a partire dalla tabella *Esatta* sono stati estratti in maniera casuale 20 campioni costituiti da 601 condotte non rotte;
- l'unione tra i 601 tratti non rotti e i 106 tratti rotti restituisce 20 estrazioni contraddistinte da una percentuale di rottura pari al 15%.

A questo punto, è stato stimato il modello di regressione logistica per ognuna delle 20 estrazioni. La scelta di estrarre 20 campioni deriva dalla necessità di estrarre in maniera casuale 601 tratti con flag nullo e, di conseguenza, un numero minore di campioni sarebbe stato riduttivo e probabilmente poco significativo nelle stime.
Ognuna delle venti estrazioni è stata denominata come *EstrazioneEsatta*, seguita dal numero dell'estrazione. Per ognuna, il modello di regressione è stato mediante il software *Gretl*.
A titolo di esempio, per l'*EstrazioneEsatta 1* si richiamano i passi esposti nel capitolo *"La regressione logistica"* per la definizione delle variabili del modello, la stima dei parametri e la valutazione della bontà di adattamento. Inoltre, sono riportati i dati di output del software *Gretl*.

EstrazioneEsatta 1

Si definiscono, in prima istanza, le variabili del modello:

- la variabile risposta dipendente *Y*, condizionata alla rottura o non rottura della generica condotta (rispettivamente flag pari a 1 o 0);
- la variabile indipendente x_1 "*materiale"*;
- la variabile indipendente x_2 *"anno di posa"*;
- la variabile indipendente x_3 *"diametro"*;
- la variabile indipendente x_4 "*pMax*", relativa al carico massimo in condotta;
- la variabile indipendente x_5 *"lunghezza*".

Come descritto in precedenza, nel caso della variabile indipendente "*materiale"*, data la sua natura qualitativa, è stato necessario assegnare un numero ad ogni materiale presente in rete da inserire tra i dati di input del modello.
Una volta definite le variabili, il modello di regressione logistica multivariata assume la seguente forma:

$$logit[P(Y=1)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

in cui β_0 è la costante del modello, β_1, β_2, β_3, β_4 e β_5 rappresentano l'effetto delle variabili indipendenti sulla probabilità di rottura della condotta. La stima di questi parametri avviene attraverso la massimizzazione della funzione di verosimiglianza ed è implementata in *Gretl*, i cui dati di output sono riportati in **Figura 7.1**.
In relazione ai risultati ottenuti, il modello di regressione stimato assume la seguente forma:

$$logit = 90.2118 - 0.4429x_1 - 0.0449x_2 - 0.0095x_3 + 0.0138x_4 + 0.0057x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5)}}{1 + e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 x_5)}}$$

$$\hat{\pi}(x) = \frac{e^{(90.2118-0.4429x_1-0.0449x_2-0.0095x_3+0.0138x_4+0.0057x_5)}}{1 + e^{(90.2118-0.4429x_1-0.0449x_2-0.0095x_3+0.0138x_4+0.0057x_5)}}$$

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;

- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

Figura 7.1. Output di *Gretl* relativo all'*EstrazioneEsatta 1*

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ----------------------------------------------------------
  flag = 1
  const       90.2118        9.62802        9.370   7.27e-021 ***
  materiale   -0.442948      0.107466      -4.122   3.76e-05  ***
  anno_posa   -0.0448802     0.00496270    -9.044   1.52e-019 ***
  diametro    -0.00947000    0.00276653    -3.423   0.0006    ***
  pMax         0.0138202     0.00523332     2.641   0.0083    ***
  lunghezza    0.00570549    0.00157628     3.620   0.0003    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -181.4546   Criterio di Akaike     374.9093
Criterio di Schwarz    402.2755   Hannan-Quinn           385.4831
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 649 (91.8%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 234.628 [0.0000]
```

Il programma fornisce, inoltre, nella colonna *'z'* i valori del *Test Z* in relazione alle singole variabili. Affinché sia respinta l'ipotesi nulla H_0 è necessario che la variabile test soddisfi la relazione:

$$|Z| > z_{\frac{\alpha}{2}}$$

che, con un livello di significatività α pari a 0.05, diventa:

$$|Z| > 1.96$$

Tale relazione risulta soddisfatta per tutte le variabili indipendenti del modello. La stessa indicazione è fornita dal *p-value*, nell'omonima colonna e dal numero di asterischi che lo seguono, strettamente connessi alla significatività di ogni singolo parametro. Infatti, maggiore è il numero di asterischi (fino ad un massimo di 3), maggiore è la significatività di ogni parametro. Un numero di asterischi inferiori a 2 indica la non-significatività di un parametro.

In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)* e indicata in questo lavoro con *G*. Nello specifico, il numero tra parentesi indica il numero di gradi di libertà ed il valore della statistica ammonta a 234.628. Questa soddisfa le seguenti relazioni:

$$G = 234.628 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

Dove il livello di significatività α è pari a 0.05 ed il *p-value* segue invece il valore del *Chi-quadro (5)* in parentesi quadre.
È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti sopra citate.
La totalità di questi passaggi è stata eseguita per le restanti 19 estrazioni, delle quali saranno riportati i risultati di *Gretl* ed una breve descrizione degli stessi.

EstrazioneEsatta 2

Si riporta in **Figura 7.2** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 2*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 72.9772 - 0.3898x_1 - 0.0363x_2 - 0.0090x_3 + 0.0090x_4 + 0.0051x_5$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ----------------------------------------------------------
  flag = 1
  const       72.9772        7.80023        9.356   8.30e-021 ***
  materiale   -0.389843      0.0966531     -4.033   5.50e-05  ***
  anno_posa   -0.0362826     0.00403094    -9.001   2.24e-019 ***
  diametro    -0.00930620    0.00267156    -3.483   0.0005    ***
  pMax         0.00901223    0.00495414     1.819   0.0689    *
  lunghezza    0.00509401    0.00139839     3.643   0.0003    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -201.1121   Criterio di Akaike     414.2242
Criterio di Schwarz    441.5904   Hannan-Quinn           424.7980
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 629 (89.0%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 195.313 [0.0000]
```

Figura 7.2. Output di *Gretl* relativo all'*EstrazioneEsatta 2*

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(72.9772-0.3898x_1-0.0363x_2-0.0090x_3+0.0090x_4+0.0051x_5)}}{1+e^{(72.9772-0.3898x_1-0.0363x_2-0.0090x_3+0.0090x_4+0.0051x_5)}}$$

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri ad eccezione del carico massimo risultano significativi. La stessa conclusione si trae valutando il valore del *p-value* relativo alla variabile *'pMax'*.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 195.313. Questa soddisfa le seguenti relazioni:

$$G = 195.313 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione le variabili *"materiale"*, *"anno di posa"*, *"diametro"* e *"lunghezza".*

EstrazioneEsatta 3

Si riporta in **Figura 7.3** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 3*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 77.0484 - 0.3634x_1 - 0.0385x_2 - 0.0088x_3 + 0.0106x_4 + 0.0067x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(77.0484 - 0.3634x_1 - 0.0385x_2 - 0.0088x_3 + 0.0106x_4 + 0.0067x_5)}}{1 + e^{(77.0484 - 0.3634x_1 - 0.0385x_2 - 0.0088x_3 + 0.0106x_4 + 0.0067x_5)}}$$

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ----------------------------------------------------------
  flag = 1
  const       77.0484        8.20677        9.388   6.09e-021 ***
  materiale   -0.363412      0.0959101     -3.789   0.0002    ***
  anno_posa   -0.0385175     0.00423211    -9.101   8.93e-020 ***
  diametro    -0.00882081    0.00256588    -3.438   0.0006    ***
  pMax         0.0105846     0.00496431     2.132   0.0330    **
  lunghezza    0.00665305    0.00170333     3.906   9.39e-05  ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -194.0032   Criterio di Akaike     400.0065
Criterio di Schwarz    427.3726   Hannan-Quinn           410.5802
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 636 (90.0%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 209.531 [0.0000]
```

Figura 7.3. Output di *Gretl* relativo all'*EstrazioneEsatta 3*

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 209.531. Questa soddisfa le seguenti relazioni:

$$G = 209.531 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 4

Si riporta in **Figura 7.4** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 4*.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

              coefficiente   errore std.      z       p-value
  ----------------------------------------------------------
  flag = 1
  const        88.9612        9.28683        9.579   9.77e-022 ***
  materiale    -0.313939      0.0917771     -3.421   0.0006    ***
  anno_posa    -0.0447496     0.00478074    -9.360   7.94e-021 ***
  diametro     -0.00778428    0.00248259    -3.136   0.0017    ***
  pMax          0.0131652     0.00503695     2.614   0.0090    ***
  lunghezza     0.00460246    0.00139494     3.299   0.0010    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -190.0503   Criterio di Akaike     392.1006
Criterio di Schwarz    419.4668   Hannan-Quinn           402.6744
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 647 (91.5%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 217.437 [0.0000]
```

Figura 7.4. Output di *Gretl* relativo all'*EstrazioneEsatta 4*

Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 88.9612 - 0.3140x_1 - 0.0447x_2 - 0.0078x_3 + 0.0132x_4 + 0.0047x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(88.9612 - 0.3140x_1 - 0.0447x_2 - 0.0078x_3 + 0.0132x_4 + 0.0047x_5)}}{1 + e^{(88.9612 - 0.3140x_1 - 0.0447x_2 - 0.0078x_3 + 0.0132x_4 + 0.0047x_5)}}$$

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 217.437. Questa soddisfa le seguenti relazioni:

$$G = 217.437 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 5

Si riporta in **Figura 7.5** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 5*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 89.8137 - 0.3660x_1 - 0.0450x_2 - 0.0081x_3 + 0.0125x_4 + 0.0059x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(88.9612 - 0.3140x_1 - 0.0447x_2 - 0.0078x_3 + 0.0132x_4 + 0.0047x_5)}}{1 + e^{(88.9612 - 0.3140x_1 - 0.0447x_2 - 0.0078x_3 + 0.0132x_4 + 0.0047x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------
  flag = 1
  const       89.8137        9.38134        9.574   1.03e-021 ***
  materiale   -0.365956      0.0979710     -3.735   0.0002    ***
  anno_posa   -0.0450023     0.00483334    -9.311   1.27e-020 ***
  diametro    -0.00809831    0.00256069    -3.163   0.0016    ***
  pMax         0.0125480     0.00505543     2.482   0.0131    **
  lunghezza    0.00594798    0.00156523     3.800   0.0001    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -185.2242   Criterio di Akaike     382.4484
Criterio di Schwarz    409.8146   Hannan-Quinn           393.0222
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 646 (91.4%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 227.089 [0.0000]
```

Figura 7.5. Output di *Gretl* relativo all'*EstrazioneEsatta 5*

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 227.089. Questa soddisfa le seguenti relazioni:

$$G = 227.089 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 6

Si riporta in **Figura 7.6** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 6*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 72.0590 - 0.3717x_1 - 0.0359x_2 - 0.0089x_3 + 0.0101x_4 + 0.0045x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(72.0590 - 0.3717x_1 - 0.0359x_2 - 0.0089x_3 + 0.0101x_4 + 0.0045x_5)}}{1 + e^{(72.0590 - 0.3717x_1 - 0.0359x_2 - 0.0089x_3 + 0.0101x_4 + 0.0045x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ----------------------------------------------------------
  flag = 1
  const       72.0590        7.84652        9.184   4.17e-020 ***
  materiale   -0.374714      0.0925965     -4.047   5.19e-05  ***
  anno_posa   -0.0358853     0.00405982    -8.839   9.65e-019 ***
  diametro    -0.00885911    0.00261381    -3.389   0.0007    ***
  pMax         0.0100839     0.00491702     2.051   0.0403    **
  lunghezza    0.00446027    0.00131234     3.399   0.0007    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -199.7831   Criterio di Akaike     411.5661
Criterio di Schwarz    438.9323   Hannan-Quinn           422.1399
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 631 (89.3%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 197.971 [0.0000]
```

Figura 7.6. Output di *Gretl* relativo all'*EstrazioneEsatta 6*

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;

- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 197.971. Questa soddisfa le seguenti relazioni:

$$G = 197.971 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 7

Si riporta in **Figura 7.7** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 7*.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

              coefficiente   errore std.      z       p-value
  ----------------------------------------------------------
  flag = 1
  const        80.6626        8.48935        9.502   2.07e-021 ***
  materiale    -0.358419      0.0928396     -3.861   0.0001    ***
  anno_posa    -0.0404225     0.00437366    -9.242   2.41e-020 ***
  diametro     -0.00735053    0.00238526    -3.082   0.0021    ***
  pMax          0.0129600     0.00501514     2.584   0.0098    ***
  lunghezza     0.00394510    0.00135956     2.902   0.0037    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -194.4984   Criterio di Akaike     400.9968
Criterio di Schwarz    428.3630   Hannan-Quinn           411.5706
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 643 (90.9%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 208.54 [0.0000]
```

Figura 7.7. Output di *Gretl* relativo all'*EstrazioneEsatta 7*

Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 80.6626 - 0.3584x_1 - 0.0404x_2 - 0.0074x_3 + 0.0130x_4 + 0.0039x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(80.6626 - 0.3584x_1 - 0.0404x_2 - 0.0074x_3 + 0.0130x_4 + 0.0039x_5)}}{1 + e^{(80.6626 - 0.3584x_1 - 0.0404x_2 - 0.0074x_3 + 0.0130x_4 + 0.0039x_5)}}$$

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 208.54. Questa soddisfa le seguenti relazioni:

$$G = 208.54 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 8

Si riporta in **Figura 7.8** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 8*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 80.1108 - 0.3684x_1 - 0.0399x_2 - 0.0095x_3 + 0.0104x_4 + 0.0036x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(80.1108 - 0.3684x_1 - 0.0399x_2 - 0.0095x_3 + 0.0104x_4 + 0.0036x_5)}}{1 + e^{(80.1108 - 0.3684x_1 - 0.0399x_2 - 0.0095x_3 + 0.0104x_4 + 0.0036x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

              coefficiente   errore std.      z       p-value
  ---------------------------------------------------------------
  flag = 1
  const        80.1108        8.45629        9.474   2.71e-021 ***
  materiale    -0.368406      0.0942643     -3.908   9.30e-05  ***
  anno_posa    -0.0399066     0.00436253    -9.148   5.82e-020 ***
  diametro     -0.00951312    0.00283422    -3.357   0.0008    ***
  pMax          0.0104226     0.00495023     2.105   0.0352    **
  lunghezza     0.00359854    0.00126566     2.843   0.0045    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -195.3042   Criterio di Akaike     402.6084
Criterio di Schwarz    429.9746   Hannan-Quinn           413.1822
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 638 (90.2%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 206.929 [0.0000]
```

Figura 7.8. Output di *Gretl* relativo all'*EstrazioneEsatta 8*

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 206.929. Questa soddisfa le seguenti relazioni:

$$G = 206.929 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 9

Si riporta in **Figura 7.9** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 9*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 81.7267 - 0.3258x_1 - 0.0408x_2 - 0.0095x_3 + 0.0110x_4 + 0.0024x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(81.7267 - 0.3258x_1 - 0.0408x_2 - 0.0095x_3 + 0.0110x_4 + 0.0024x_5)}}{1 + e^{(81.7267 - 0.3258x_1 - 0.0408x_2 - 0.0095x_3 + 0.0110x_4 + 0.0024x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ------------------------------------------------------------
  flag = 1
  const       81.7267        8.52824        9.583   9.42e-022 ***
  materiale   -0.325756      0.0935159     -3.483   0.0005    ***
  anno_posa   -0.0408145     0.00438896    -9.299   1.41e-020 ***
  diametro    -0.00952639    0.00262825    -3.625   0.0003    ***
  pMax         0.0110497     0.00493481     2.239   0.0251    **
  lunghezza    0.00244215    0.00112103     2.178   0.0294    **

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -193.8545   Criterio di Akaike     399.7090
Criterio di Schwarz    427.0751   Hannan-Quinn           410.2827
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 643 (90.9%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 209.828 [0.0000]
```

Figura 7.9. Output di *Gretl* relativo all'*EstrazioneEsatta 9*

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;

- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 209.828. Questa soddisfa le seguenti relazioni:

$$G = 209.828 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta10

Si riporta in **Figura 7.10** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 10*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 77.4805 - 0.3799x_1 - 0.0386x_2 - 0.0091x_3 + 0.0111x_4 + 0.0047x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(77.4805 - 0.3799x_1 - 0.0386x_2 - 0.0091x_3 + 0.0111x_4 + 0.0047x_5)}}{1 + e^{(77.4805 - 0.3799x_1 - 0.0386x_2 - 0.0091x_3 + 0.0111x_4 + 0.0047x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------
  flag = 1
  const       77.4805        8.22007        9.426   4.27e-021 ***
  materiale   -0.379852      0.0959282     -3.960   7.50e-05  ***
  anno_posa   -0.0386173     0.00423600    -9.116   7.76e-020 ***
  diametro    -0.00905991    0.00258156    -3.509   0.0004    ***
  pMax         0.0111166     0.00495961     2.241   0.0250    **
  lunghezza    0.00471845    0.00148775     3.172   0.0015    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -194.6713   Criterio di Akaike     401.3426
Criterio di Schwarz    428.7088   Hannan-Quinn           411.9163
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 638 (90.2%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 208.195 [0.0000]
```

Figura 7.10. Output di *Gretl* relativo all'*EstrazioneEsatta 10*

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 208.195. Questa soddisfa le seguenti relazioni:

$$G = 208.195 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 11

Si riporta in **Figura 7.11** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 11*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 75.3364 - 0.3992x_1 - 0.0376x_2 - 0.0077x_3 + 0.0111x_4 + 0.0049x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(75.3364 - 0.3992x_1 - 0.0376x_2 - 0.0077x_3 + 0.0111x_4 + 0.0049x_5)}}{1 + e^{(75.3364 - 0.3992x_1 - 0.0376x_2 - 0.0077x_3 + 0.0111x_4 + 0.0049x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

              coefficiente   errore std.     z       p-value
  ----------------------------------------------------------
  flag = 1
  const        75.3364        8.17896        9.211   3.23e-020 ***
  materiale    -0.399229      0.0998090     -4.000   6.34e-05  ***
  anno_posa    -0.0375682     0.00423214    -8.877   6.88e-019 ***
  diametro     -0.00772560    0.00235396    -3.282   0.0010    ***
  pMax          0.0111829     0.00494053     2.264   0.0236    **
  lunghezza     0.00492968    0.00147039     3.353   0.0008    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -196.2179   Criterio di Akaike     404.4357
Criterio di Schwarz    431.8019   Hannan-Quinn           415.0095
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 635 (89.8%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 205.101 [0.0000]
```

Figura 7.11. Output di *Gretl* relativo all'*EstrazioneEsatta 11*

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;

- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 205.101. Questa soddisfa le seguenti relazioni:

$$G = 205.101 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 12

Si riporta in **Figura 7.12** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 12*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 79.3144 - 0.4501x_1 - 0.0392x_2 - 0.0108x_3 + 0.0110x_4 + 0.0048x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(79.3144 - 0.4501x_1 - 0.0392x_2 - 0.0108x_3 + 0.0110x_4 + 0.0048x_5)}}{1 + e^{(79.3144 - 0.4501x_1 - 0.0392x_2 - 0.0108x_3 + 0.0110x_4 + 0.0048x_5)}}$$

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;

- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ----------------------------------------------------------
  flag = 1
  const       79.3144        8.44059        9.397   5.63e-021 ***
  materiale   -0.450112      0.104646      -4.301   1.70e-05  ***
  anno_posa   -0.0391949     0.00435255    -9.005   2.16e-019 ***
  diametro    -0.0108153     0.00281795    -3.838   0.0001    ***
  pMax         0.0110712     0.00506422     2.186   0.0288    **
  lunghezza    0.00477010    0.00140899     3.385   0.0007    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -190.0959   Criterio di Akaike     392.1918
Criterio di Schwarz    419.5579   Hannan-Quinn           402.7655
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 639 (90.4%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 217.345 [0.0000]
```

Figura 7.12. Output di *Gretl* relativo all'*EstrazioneEsatta 12*

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 217.345. Questa soddisfa le seguenti relazioni:

$$G = 217.345 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 13

Si riporta in **Figura 7.13** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 13*.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------
  flag = 1
  const       84.8670        8.88873        9.548   1.33e-021 ***
  materiale   -0.358366      0.0967477     -3.704   0.0002    ***
  anno_posa   -0.0423885     0.00458166    -9.252   2.21e-020 ***
  diametro    -0.00945705    0.00267598    -3.534   0.0004    ***
  pMax         0.0114835     0.00499367     2.300   0.0215    **
  lunghezza    0.00453923    0.00143201     3.170   0.0015    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -191.1771   Criterio di Akaike     394.3541
Criterio di Schwarz    421.7203   Hannan-Quinn           404.9279
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 643 (90.9%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 215.183 [0.0000]
```

Figura 7.13. Output di *Gretl* relativo all'*EstrazioneEsatta 13*

Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 84.8970 - 0.3584x_1 - 0.0424x_2 - 0.0095x_3 + 0.0115x_4 + 0.0045x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(84.8970 - 0.3584x_1 - 0.0424x_2 - 0.0095x_3 + 0.0115x_4 + 0.0045x_5)}}{1 + e^{(84.8970 - 0.3584x_1 - 0.0424x_2 - 0.0095x_3 + 0.0115x_4 + 0.0045x_5)}}$$

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.

In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 215.183. Questa soddisfa le seguenti relazioni:

$$G = 215.183 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 14

Si riporta in **Figura 7.14** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 14*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 74.7916 - 0.4046x_1 - 0.0371x_2 - 0.0095x_3 + 0.0113x_4 + 0.0037x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(74.7916 - 0.4046x_1 - 0.0371x_2 - 0.0095x_3 + 0.0113x_4 + 0.0037x_5)}}{1 + e^{(74.7916 - 0.4046x_1 - 0.0371x_2 - 0.0095x_3 + 0.0113x_4 + 0.0037x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ------------------------------------------------------------
  flag = 1
  const       74.7916        8.17097        9.153   5.52e-020 ***
  materiale   -0.404571      0.0986927     -4.099   4.14e-05  ***
  anno_posa   -0.0371232     0.00423099    -8.774   1.72e-018 ***
  diametro    -0.00949544    0.00289668    -3.278   0.0010    ***
  pMax         0.0113178     0.00501369     2.257   0.0240    **
  lunghezza    0.00373409    0.00132545     2.817   0.0048    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -195.6118   Criterio di Akaike     403.2236
Criterio di Schwarz    430.5898   Hannan-Quinn           413.7973
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 636 (90.0%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 206.314 [0.0000]
```

Figura 7.14. Output di *Gretl* relativo all'*EstrazioneEsatta 14*

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 206.314. Questa soddisfa le seguenti relazioni:

$$G = 206.314 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 15

Si riporta in **Figura 7.15** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 15*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 74.3953 - 0.3501x_1 - 0.0371x_2 - 0.0371x_3 + 0.0086x_4 + 0.0029x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(74.3953 - 0.3501x_1 - 0.0371x_2 - 0.0371x_3 + 0.0086x_4 + 0.0029x_5)}}{1 + e^{(74.3953 - 0.3501x_1 - 0.0371x_2 - 0.0371x_3 + 0.0086x_4 + 0.0029x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ----------------------------------------------------------
  flag = 1
  const       74.3953        7.99881        9.301   1.39e-020 ***
  materiale   -0.350145      0.0917303     -3.817   0.0001    ***
  anno_posa   -0.0371605     0.00412553    -9.007   2.11e-019 ***
  diametro    -0.00855455    0.00259305    -3.299   0.0010    ***
  pMax         0.0118377     0.00490038     2.416   0.0157    **
  lunghezza    0.00294331    0.00114015     2.582   0.0098    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -199.4772   Criterio di Akaike     410.9544
Criterio di Schwarz    438.3205   Hannan-Quinn           421.5281
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 636 (90.0%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 198.583 [0.0000]
```

Figura 7.15. Output di *Gretl* relativo all'*EstrazioneEsatta 15*

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 198.583. Questa soddisfa le seguenti relazioni:

$$G = 198.583 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 16

Si riporta in **Figura 7.16** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 16*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 86.1129 - 0.4015x_1 - 0.0429x_2 - 0.0429x_3 + 0.0013x_4 + 0.0051x_5$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ----------------------------------------------------------
  flag = 1
  const       86.1129        9.26527        9.294   1.48e-020 ***
  materiale   -0.401476      0.101160      -3.969   7.23e-05  ***
  anno_posa   -0.0429159     0.00476921    -8.999   2.29e-019 ***
  diametro    -0.00907285    0.00252073    -3.599   0.0003    ***
  pMax         0.0124544     0.00511639     2.434   0.0149    **
  lunghezza    0.00512557    0.00153812     3.332   0.0009    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -184.4624   Criterio di Akaike     380.9248
Criterio di Schwarz    408.2910   Hannan-Quinn           391.4986
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 646 (91.4%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 228.612 [0.0000]
```

Figura 7.16. Output di *Gretl* relativo all'*EstrazioneEsatta 16*

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(86.1129 - 0.4015x_1 - 0.0429x_2 - 0.0429x_3 + 0.0013x_4 + 0.0051x_5)}}{1 + e^{(86.1129 - 0.4015x_1 - 0.0429x_2 - 0.0429x_3 + 0.0013x_4 + 0.0051x_5)}}$$

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;

- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 228.612. Questa soddisfa le seguenti relazioni:

$$G = 228.612 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 17

Si riporta in **Figura 7.17** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 17*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 92.0854 - 0.2391x_1 - 0.0466x_2 - 0.0073x_3 + 0.0127x_4 + 0.0055x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(92.0854 - 0.2391x_1 - 0.0466x_2 - 0.0073x_3 + 0.0127x_4 + 0.0055x_5)}}{1 + e^{(92.0854 - 0.2391x_1 - 0.0466x_2 - 0.0073x_3 + 0.0127x_4 + 0.0055x_5)}}$$

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;

- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ---------------------------------------------------------
  flag = 1
  const       92.0854        9.36094        9.837   7.78e-023 ***
  materiale   -0.239077      0.0848923     -2.816   0.0049    ***
  anno_posa   -0.0466093     0.00479743    -9.715   2.59e-022 ***
  diametro    -0.00726914    0.00240839    -3.018   0.0025    ***
  pMax         0.0126916     0.00487729     2.602   0.0093    ***
  lunghezza    0.00546669    0.00147824     3.698   0.0002    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -188.9689   Criterio di Akaike     389.9378
Criterio di Schwarz    417.3040   Hannan-Quinn           400.5116
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 648 (91.7%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 219.599 [0.0000]
```

Figura 7.17. Output di *Gretl* relativo all'*EstrazioneEsatta 17*

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 219.599. Questa soddisfa le seguenti relazioni:

$$G = 219.599 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 18

Si riporta in **Figura 7.18** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 18*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 87.1700 - 0.4272x_1 - 0.0435x_2 - 0.0074x_3 + 0.0129x_4 + 0.0049x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(87.1700 - 0.4272x_1 - 0.0435x_2 - 0.0074x_3 + 0.0129x_4 + 0.0049x_5)}}{1 + e^{(87.1700 - 0.4272x_1 - 0.0435x_2 - 0.0074x_3 + 0.0129x_4 + 0.0049x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------
  flag = 1
  const       87.1700        9.48084        9.194   3.77e-020 ***
  materiale   -0.427195      0.106688      -4.004   6.22e-05  ***
  anno_posa   -0.0434697     0.00489431    -8.882   6.59e-019 ***
  diametro    -0.00736146    0.00230053    -3.200   0.0014    ***
  pMax         0.0128538     0.00511449     2.513   0.0120    **
  lunghezza    0.00486247    0.00152441     3.190   0.0014    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -185.0215   Criterio di Akaike     382.0431
Criterio di Schwarz    409.4093   Hannan-Quinn           392.6168
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 645 (91.2%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 227.494 [0.0000]
```

Figura 7.18. Output di *Gretl* relativo all'*EstrazioneEsatta 18*

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;
- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 227.494. Questa soddisfa le seguenti relazioni:

$$G = 227.494 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 19

Si riporta in **Figura 7.19** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 19*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 84.3071 - 0.3781x_1 - 0.0422x_2 - 0.0080x_3 + 0.0132x_4 + 0.052x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(84.3071 - 0.3781x_1 - 0.0422x_2 - 0.0080x_3 + 0.0132x_4 + 0.052x_5)}}{1 + e^{(84.3071 - 0.3781x_1 - 0.0422x_2 - 0.0080x_3 + 0.0132x_4 + 0.052x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ----------------------------------------------------------
  flag = 1
  const       84.3071        9.05652        9.309   1.29e-020 ***
  materiale   -0.378053      0.0996132     -3.795   0.0001    ***
  anno_posa   -0.0421876     0.00467522    -9.024   1.82e-019 ***
  diametro    -0.00798633    0.00236597    -3.376   0.0007    ***
  pMax         0.0132210     0.00507110     2.607   0.0091    ***
  lunghezza    0.00519595    0.00154078     3.372   0.0007    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -188.8162   Criterio di Akaike     389.6323
Criterio di Schwarz    416.9985   Hannan-Quinn           400.2061
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 646 (91.4%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 219.905 [0.0000]
```

Figura 7.19. Output di *Gretl* relativo all'*EstrazioneEsatta 19*

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;

- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 219.905. Questa soddisfa le seguenti relazioni:

$$G = 219.905 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

EstrazioneEsatta 20

Si riporta in **Figura 7.20** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta 20*.
Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 84.1626 - 0.4396x_1 - 0.0418x_2 - 0.0091x_3 + 0.0127x_4 + 0.0044$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{e^{(84.1626 - 0.4396x_1 - 0.0418x_2 - 0.0091x_3 + 0.0127x_4 + 0.0044)}}{1 + e^{(84.1626 - 0.4396x_1 - 0.0418x_2 - 0.0091x_3 + 0.0127x_4 + 0.0044)}}$$

Si traggono le seguenti conclusioni:

- il segno negativo di β_1 indica che all'aumentare del numero assegnato a ciascun materiale, la probabilità di rottura diminuisce;

- il segno negativo di β_2 indica che all'aumentare dell'anno di posa, la probabilità di rottura diminuisce. Condotte più giovani hanno una minore probabilità di rottura;
- il segno negativo di β_3 indica che all'aumentare del diametro della condotta, la probabilità di rottura diminuisce;
- il segno positivo di β_4 indica che all'aumentare del carico in condotta, la probabilità di rottura aumenta;
- il segno positivo di β_5 indica che all'aumentare della lunghezza della condotta, la probabilità di rottura aumenta.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-707
Variabile dipendente: flag
Errori standard basati sull'Hessiana

              coefficiente   errore std.      z       p-value
  ----------------------------------------------------------
  flag = 1
  const        84.1626        9.11744        9.231   2.68e-020 ***
  materiale    -0.439581      0.103969      -4.228   2.36e-05  ***
  anno_posa    -0.0418092     0.00471143    -8.874   7.06e-019 ***
  diametro     -0.00906245    0.00269771    -3.359   0.0008    ***
  pMax          0.0126994     0.00511069     2.485   0.0130    **
  lunghezza     0.00443291    0.00141173     3.140   0.0017    ***

Media var. dipendente  0.149929   SQM var. dipendente    0.357255
Log-verosimiglianza   -187.2870   Criterio di Akaike     386.5740
Criterio di Schwarz    413.9402   Hannan-Quinn           397.1478
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 645 (91.2%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 222.963 [0.0000]
```

Figura 7.20. Output di *Gretl* relativo all'*EstrazioneEsatta 20*

I valori del *test Z* relativi ad ogni variabile, una volta fissato il livello di significatività α pari a 0.05, mostrano che tutti i parametri risultano significativi.
In aggiunta, il software fornisce il valore della statistica *Chi-quadro*, riportata come *Chi-quadro (5)*. Nello specifico, il valore della statistica ammonta a 222.963. Questa soddisfa le seguenti relazioni:

$$G = 222.963 > \chi^2_{\frac{\alpha}{2}} = 12.38$$

$$p - value = 0.0000 < 0.005$$

È possibile affermare, quindi, che il modello completo aggiunge informazioni rispetto al modello con la sola intercetta e quindi le stime vengono implementate, una volta prese in considerazione tutte le variabili indipendenti.

Una volta esaminati i risultati delle venti estrazione, è stato possibile riscontrare che:

- il carico massimo "*pMax"* risulta sempre meno rilevante degli altri parametri, ma in un solo caso è non significativo. Questo è provato dai valori del *p-value* associati in ogni estrazione a questa variabile e a quelli della variabile *test Z*;
- in tutte le estrazioni, le variabili "*materiale*", "*anno di posa*", "*diametro*" e "l*unghezza*" risultano significative nel modello;
- la probabilità di rottura diminuisce all'aumentare del numero assegnato ad ogni materiale, dell'anno di posa e del diametro;
- la probabilità di rottura aumenta all'aumentare del carico massimo e della lunghezza della condotta.

7.2. Verifica dei modelli

Nel paragrafo precedente sono state definite le variabili del modello, distinguendole tra variabili significative e non significative.
Per verificare la bontà di questi modelli nel prevedere una rottura, sono stati adottati i seguenti passi:

1. Sono state considerate le venti estrazioni esaminate nel paragrafo precedente.
2. Da ognuna di queste estrazioni è stato estratto in maniera casuale l'80% degli elementi, cioè 566 condotte.
3. Sono state calcolate le nuove stime dei parametri di regressione logistica per ogni estrazione composta dall'80% degli elementi dell'*EsattaRidotta* generica, denominata *EstrazioneEsatta80%*. Queste stime sono state effettuate prendendo in considerazione tutte le variabili indipendenti citate in precedenza e, nel caso in cui una variabile sia risultata non significativa, sono stati nuovamente stimati i parametri eliminando la variabile citata.
4. Per ogni estrazione, il modello ottenuto considerando l'80% degli elementi è stato successivamente applicato al restante 20% dell'estrazione. Ognuna di queste sotto-estrazioni, composte da 141 condotte, è stata denominata *EstrazioneEsatta20%.*
5. L'applicazione del punto *4* consente di ottenere le probabilità di rottura stimate per ogni condotta facente parte del 20% delle estrazioni. Tali valori sono stati riportati in dei grafici a bolle, nei quali le ordinate rappresentano la probabilità stimata di rottura e le ascisse il dato sulla reale rottura o non rottura della condotta (flag pari a 1 o 0);
6. Infine, per valutare la capacità del modello di predire una rottura, è stato creato un grafico a bolle per *ogni EstrazioneEsatta20%* e sono state comparate le probabilità stimate per le condotte effettivamente rotte e

quelle non rotte. Un buon modello, per essere tale, dovrebbe stimare una maggiore probabilità di rottura per le condotte con flag pari a 1.

Si riporta di seguito un esempio per chiarire la procedura appena esposta.

Esempio 1.

Si prenda in esame l'estrazione **EstrazioneEsatta 2**, composta da 707 righe rappresentanti 106 rotture e 501 condotte non rotte. A partire da questi elementi, è stato estratto l'80% degli elementi. Si ricava un nuovo campione di 566 condotte, denominato **EstrazioneEsatta80%-2** e contraddistinto da un numero casuale di rotture.

A partire da questo campione, sono stati stimati i nuovi parametri della regressione, prendendo in considerazione tutte le variabili indipendenti (**Figura 7.21**).

Figura 7.21. Output di *Gretl* relativo al campione costituito dall'80% degli elementi inclusi nell'*EstrazioneEsatta 2 (tutte le variabili indipendenti incluse)*

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------
  flag = 1
  const       72.1634        8.61006        8.381   5.23e-017 ***
  materiale   -0.394264      0.103616      -3.805   0.0001    ***
  anno_posa   -0.0358535     0.00443192    -8.090   5.97e-016 ***
  diametro    -0.00981212    0.00296700    -3.307   0.0009    ***
  pMax         0.00906978    0.00555466     1.633   0.1025
  lunghezza    0.00556431    0.00154822     3.594   0.0003    ***

Media var. dipendente  0.150177   SQM var. dipendente    0.357561
Log-verosimiglianza   -160.8649   Criterio di Akaike     333.7299
Criterio di Schwarz    359.7614   Hannan-Quinn           343.8897
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 504 (89.0%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 157.124 [0.0000]
```

Data la poca rilevanza della variabile relativa al carico massimo, sono stati nuovamente stimati i parametri del modello con le sole variabili "*materiale*", "*anno di posa*", "*diametro*" e "l*unghezza*" (**Figura 7.22**).

Il modello di regressione logistica multivariata assume la seguente forma:

$$logit[P(Y=1)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5$$

in cui β_0 è la costante del modello, β_1, β_2, β_3 e β_5 rappresentano l'effetto delle variabili indipendenti sulla probabilità di rottura della condotta.

In relazione ai risultati ottenuti, il modello di regressione stimato assume la seguente forma:

$$logit = 72.2308 - 0.4181x_1 - 0.0354x_2 - 0.0091x_3 + 0.0054x_4$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ---------------------------------------------------------
  flag = 1
  const       72.2308        8.60774        8.391   4.80e-017 ***
  materiale   -0.418082      0.0992013     -4.214   2.50e-05  ***
  anno_posa   -0.0354813     0.00440910    -8.047   8.47e-016 ***
  diametro    -0.0106138     0.00298541    -3.555   0.0004    ***
  lunghezza    0.00543760    0.00154751     3.514   0.0004    ***

Media var. dipendente  0.150177   SQM var. dipendente    0.357561
Log-verosimiglianza   -162.1689   Criterio di Akaike     334.3377
Criterio di Schwarz    356.0307   Hannan-Quinn           342.8043
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 503 (88.9%)
Test del rapporto di verosimiglianza: Chi-quadro(4) = 154.516 [0.0000]
```

Figura 7.22. Output di *Gretl* relativo al campione costituito dall'80% degli elementi inclusi nell'*EstrazioneEsatta 2* (variabile *"pMax"* esclusa)

Si calcola quindi la corrispondente probabilità di rottura per la generica condotta facente parte dell'**EstrazioneEsatta20%-2** come:

$$\hat{\pi}(x) = \frac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_5x_5)}}{1 + e^{((\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3++\beta_5x_5)}}$$

$$\hat{\pi}(x) = \frac{e^{(72.2308-0.4181x_1-0.0354x_2-0.0091x_3+0.0054x_5)}}{1 + e^{(72.2308-0.4181x_1-0.0354x_2-0.0091x_3+0.0054x_5)}}$$

Attraverso questa espressione è possibile calcolare la probabilità di rottura di ogni condotta presente nella sotto-estrazione. Considerando, ad esempio, la condotta con *flag* pari a *1*, caratterizzata dal materiale *Ghisa Grigia* (*numero 6*), diametro *100 millimetri*, posata nel *1967* e caratterizzata da una lunghezza di *30 metri*, la sua probabilità stimata di rottura risulta:

$$\hat{\pi}(x) = \frac{e^{(72.2308-0.4181\cdot6-0.0354\cdot1967-0.0091\cdot100+0.0054\cdot30)}}{1 + e^{(72.2308-0.4181\cdot6-0.0354\cdot1967-0.0091\cdot100+0.0054\cdot30)}} = 34.12\%$$

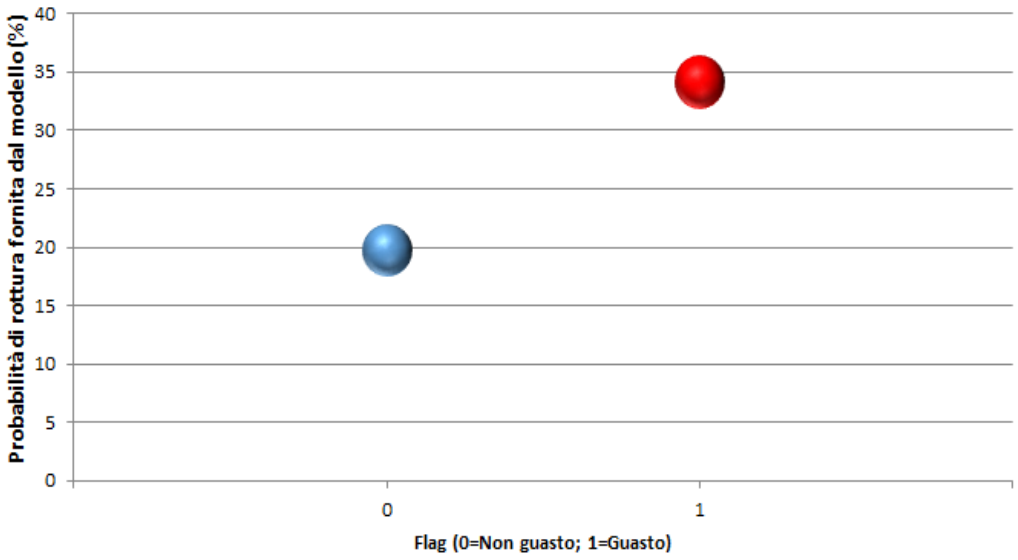
Si prende in considerazione una seconda condotta facente parte dello stesso campione, caratterizzata da *flag* pari a *0* (nessuna rottura nel periodo di

osservazione dal 2006 al 2016), in materiale *Ghisa Grigia* (numero *6*), posato nel 2006, caratterizzato da un diametro di 6*0 millimetri* ed una lunghezza totale di *80 metri*. La sua probabilità stimata di rottura risulta:

$$\hat{\pi}(x) = \frac{e^{(72.2308 - 0.4181 \cdot 6 - 0.0354 \cdot 2006 - 0.0091 \cdot 60 + 0.0054 \cdot 80)}}{1 + e^{(72.2308 - 0.4181 \cdot 6 - 0.0354 \cdot 2006 - 0.0091 \cdot 60 + 0.0054 \cdot 80)}} = 19.72\%$$

Si riportano i risultati nel grafico in **Figura 7.23**.

Figura 7.23. Confronto tra la probabilità di rottura delle due condotte e il dato relativo alla rottura reale



Il grafico mostra che la probabilità stimata per la condotta non rotta (bolla blu) è minore della probabilità stimata per la condotta realmente rotta (bolla rossa). Questa procedura è stata adottata per tutte le condotte dell'estrazione in esame. Per una maggiore chiarezza espositiva dei risultati, sono state calcolate le occorrenze di ogni coppia di punti *probabilità-flag*, cioè il numero di volte cui una determinata probabilità è associata ad un determinato flag. Le bolle che si ricavano hanno una dimensione proporzionale al numero di occorrenze e fanno riferimento alle coppie *probabilità-flag*. Una volta adottata questa metodologia a tutte le condotte facenti parte dell'*EstrazioneEsatta20%-2*, si ottiene il grafico in **Figura 7.24**.

Le bolle di colore grigio e giallo rappresentano rispettivamente i valori medi delle probabilità stimate per le condotte con flag pari a 0 e 1. Inoltre, confermano che in media le probabilità di rottura di condotte realmente guaste sono più alte delle probabilità di rottura delle condotte non guaste.

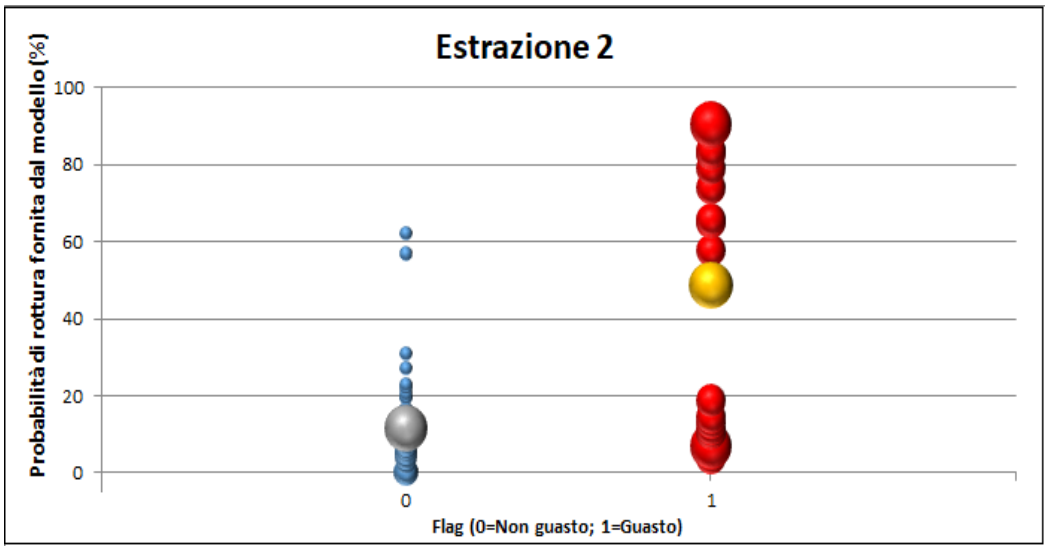


Figura 7.24. Confronto tra la probabilità di rottura e il dato relativo alla rottura reale delle condotte facenti parte dell'*EstrazioneEsatta 2*

La verifica esposta nel precedente esempio è stata eseguita per tutte le estrazioni. Di seguito si riportano le stime di *Gretl* tarate su ogni *EstrazioneEsatta80%* e le verifiche di significatività dei singoli parametri. Nel caso in cui, per una determinata sotto-estrazione, alcuni parametri risultino non significativi, le stime saranno effettuate nuovamente escludendo i parametri in questione. La definizione delle variabili segue la nomenclatura utilizzata nei precedenti casi e riportata nuovamente qui di seguito:

- la variabile risposta dipendente *Y*, condizionata alla rottura o non rottura della generica condotta (rispettivamente flag pari a 1 o 0);
- la variabile indipendente x_1 "*materiale*";
- la variabile indipendente x_2 *"anno di posa"*;
- la variabile indipendente x_3 *"diametro"*;
- la variabile indipendente x_4 "*pMax*", relativa al carico massimo in condotta;
- la variabile indipendente x_5 *"lunghezza*".

L'applicazione di questi modelli sulle 141 condotte facenti parte di ogni *EstrazioneEsatta20%* è riportata a seguito delle verifiche ed è rappresentata da grafici a bolle.

EstrazioneEsatta80%-1

Si riporta in **Figura 7.25** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-1*.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------------
  flag = 1
  const       86.9938        9.97076        8.725   2.66e-018 ***
  materiale   -0.453738      0.122557      -3.702   0.0002    ***
  anno_posa   -0.0432473     0.00516647    -8.371   5.73e-017 ***
  diametro    -0.0108433     0.00332625    -3.260   0.0011    ***
  pMax         0.0165598     0.00584230     2.834   0.0046    ***
  lunghezza    0.00540583    0.00213542     2.532   0.0114    **

Media var. dipendente  0.153710   SQM var. dipendente    0.360990
Log-verosimiglianza   -141.8696   Criterio di Akaike     295.7393
Criterio di Schwarz    321.7708   Hannan-Quinn           305.8992
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 521 (92.0%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 201.992 [0.0000]
```

Figura 7.25. Output di *Gretl* relativo all'*EstrazioneEsatta80%-1*

Tramite la valutazione del *Test* Z, tutte le variabili risultano significative. Di conseguenza, l'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-1* risulta:

$$\hat{\pi}(x) = \frac{e^{(86.9938 - 0.453738x_1 - 0.0432473x_2 - 0.0108433x_3 + 0.0165598x_4 + 0.00540583x_5)}}{1 + e^{(86.9938 - 0.453738x_1 - 0.0432473x_2 - 0.0108433x_3 + 0.0165598x_4 + 0.00540583x_5)}}$$

EstrazioneEsatta80%-2

Per i risultati relativi a questa estrazione, si rimanda all'**Esempio 1**.

EstrazioneEsatta80%-3

Si riporta in **Figura 7.26** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-3*. Tramite la valutazione del *Test* Z, tutte le variabili risultano significative. Di conseguenza, l'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-3* risulta:

$$\hat{\pi}(x) = \frac{e^{(79.853 - 0.362241x_1 - 0.0400231x_2 - 0.00771224x_3 + 0.0141659x_4 + 0.00741502x_5)}}{1 + e^{(79.853 - 0.362241x_1 - 0.0400231x_2 - 0.00771224x_3 + 0.0141659x_4 + 0.00741502x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ------------------------------------------------------------
  flag = 1
  const       79.7853        9.44179        8.450   2.91e-017 ***
  materiale   -0.362241      0.110372      -3.282   0.0010    ***
  anno_posa   -0.0400231     0.00488388    -8.195   2.51e-016 ***
  diametro    -0.00771224    0.00239460    -3.221   0.0013    ***
  pMax         0.0141659     0.00528260     2.682   0.0073    ***
  lunghezza    0.00741502    0.00220213     3.367   0.0008    ***

Media var. dipendente  0.162544   SQM var. dipendente    0.369276
Log-verosimiglianza   -160.2811   Criterio di Akaike     332.5623
Criterio di Schwarz    358.5938   Hannan-Quinn           342.7221
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 512 (90.5%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 181.893 [0.0000]
```

Figura 7.26. Output di *Gretl* relativo all'*EstrazioneEsatta80%-3*

EstrazioneEsatta80%-4

Si riporta in **Figura 7.27** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-4*.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ------------------------------------------------------------
  flag = 1
  const       82.4200        9.68106        8.514   1.69e-017 ***
  materiale   -0.320212      0.0985413     -3.250   0.0012    ***
  anno_posa   -0.0412673     0.00497719    -8.291   1.12e-016 ***
  diametro    -0.00814069    0.00281433    -2.893   0.0038    ***
  pMax         0.00820797    0.00554051     1.481   0.1385
  lunghezza    0.00485161    0.00148676     3.263   0.0011    ***

Media var. dipendente  0.151943   SQM var. dipendente    0.359284
Log-verosimiglianza   -160.1386   Criterio di Akaike     332.2772
Criterio di Schwarz    358.3088   Hannan-Quinn           342.4371
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 511 (90.3%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 162.029 [0.0000]
```

Figura 7.27. Output di *Gretl* relativo all'*EstrazioneEsatta80%-4*

Tramite la valutazione del *Test* Z, la variabile *"pMax"* risulta non signficativa. Di conseguenza, le stime dei parametri sono state nuovamente elaborate in assenza di tale variabile.
L'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-4* risulta:

$$\hat{\pi}(x) = \frac{e^{(81.7965 - 0.352773x_1 - 0.0405458x_2 - 0.00896584x_3 + 0.00471428x_5)}}{1 + e^{(81.7965 - 0.352773x_1 - 0.0405458x_2 - 0.00896584x_3 + 0.00471428x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ----------------------------------------------------------
  flag = 1
  const       81.7965        9.63882        8.486   2.14e-017 ***
  materiale   -0.352773      0.0940978     -3.749   0.0002    ***
  anno_posa   -0.0405458     0.00492350    -8.235   1.79e-016 ***
  diametro    -0.00896584    0.00282140    -3.178   0.0015    ***
  lunghezza    0.00471428    0.00148352     3.178   0.0015    ***

Media var. dipendente  0.151943   SQM var. dipendente    0.359284
Log-verosimiglianza   -161.2128   Criterio di Akaike     332.4256
Criterio di Schwarz    354.1186   Hannan-Quinn           340.8922
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 511 (90.3%)
Test del rapporto di verosimiglianza: Chi-quadro(4) = 159.88 [0.0000]
```

Figura 7.27. Output di *Gretl* relativo all'*EstrazioneEsatta80%-4* senza la variabile *pMax*

EstrazioneEsatta80%-5

Si riporta in **Figura 7.28** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-5*.
Tramite la valutazione del *Test* Z, la variabile *"pMax"* risulta non signficativa. Di conseguenza, le stime dei parametri sono state nuovamente elaborate in assenza di tale variabile.
L'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-5* risulta:

$$\hat{\pi}(x) = \frac{e^{(81.7982 - 0.444489x_1 - 0.0403139x_2 - 0.00790365x_3 + 0.00520250x_5)}}{1 + e^{(81.7982 - 0.444489x_1 - 0.0403139x_2 - 0.00790365x_3 + 0.00520250x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------
  flag = 1
  const       81.5369        9.95727        8.189   2.64e-016 ***
  materiale   -0.434807      0.108326      -4.014   5.97e-05  ***
  anno_posa   -0.0404652     0.00511473    -7.912   2.54e-015 ***
  diametro    -0.00742468    0.00248465    -2.988   0.0028    ***
  pMax         0.00732365    0.00715516     1.024   0.3060
  lunghezza    0.00532992    0.00174445     3.055   0.0022    ***

Media var. dipendente  0.144876   SQM var. dipendente    0.352287
Log-verosimiglianza   -154.3425   Criterio di Akaike     320.6849
Criterio di Schwarz    346.7165   Hannan-Quinn           330.8448
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 515 (91.0%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 159.643 [0.0000]
```

Figura 7.28. Output di *Gretl* relativo all'*EstrazioneEsatta80%-5*

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------
  flag = 1
  const       81.7982        9.98265        8.194   2.53e-016 ***
  materiale   -0.444489      0.106802      -4.162   3.16e-05  ***
  anno_posa   -0.0403139     0.00511626    -7.880   3.29e-015 ***
  diametro    -0.00790365    0.00252156    -3.134   0.0017    ***
  lunghezza    0.00520250    0.00174641     2.979   0.0029    ***

Media var. dipendente  0.144876   SQM var. dipendente    0.352287
Log-verosimiglianza   -154.8075   Criterio di Akaike     319.6150
Criterio di Schwarz    341.3080   Hannan-Quinn           328.0815
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 514 (90.8%)
Test del rapporto di verosimiglianza: Chi-quadro(4) = 158.713 [0.0000]
```

Figura 7.29. Output di *Gretl* relativo all'*EstrazioneEsatta80%-5* senza la variabile *pMax*

EstrazioneEsatta80%-6

Si riporta in **Figura 7.30** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-6*. Tramite la valutazione del *Test* Z, la variabile *"pMax"* risulta non signficativa. Di conseguenza, le stime dei parametri sono state nuovamente elaborate in assenza di tale variabile.

L'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-6* risulta:

$$\hat{\pi}(x) = \frac{e^{(75.1496 - 0.448137x_1 - 0.0368691x_2 - 0.0103892x_3 + 0.00418549x_5)}}{1 + e^{(75.1496 - 0.448137x_1 - 0.0368691x_2 - 0.0103892x_3 + 0.00418549x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ---------------------------------------------------------
  flag = 1
  const       81.5369        9.95727        8.189   2.64e-016 ***
  materiale   -0.434807      0.108326      -4.014   5.97e-05  ***
  anno_posa   -0.0404652     0.00511473    -7.912   2.54e-015 ***
  diametro    -0.00742468    0.00248465    -2.988   0.0028    ***
  pMax         0.00732365    0.00715516     1.024   0.3060
  lunghezza    0.00532992    0.00174445     3.055   0.0022    ***

Media var. dipendente  0.144876   SQM var. dipendente    0.352287
Log-verosimiglianza   -154.3425   Criterio di Akaike     320.6849
Criterio di Schwarz    346.7165   Hannan-Quinn           330.8448
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 515 (91.0%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 159.643 [0.0000]
```

Figura 7.30. Output di *Gretl* relativo all'*EstrazioneEsatta80%-6*

```
Modello 3: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ---------------------------------------------------------
  flag = 1
  const       75.1496        9.10143        8.257   1.50e-016 ***
  materiale   -0.448137      0.111386      -4.023   5.74e-05  ***
  anno_posa   -0.0368691     0.00469269    -7.857   3.94e-015 ***
  diametro    -0.0103892     0.00323430    -3.212   0.0013    ***
  lunghezza    0.00418549    0.00143625     2.914   0.0036    ***

Media var. dipendente  0.136042   SQM var. dipendente    0.343137
Log-verosimiglianza   -146.8848   Criterio di Akaike     303.7695
Criterio di Schwarz    325.4625   Hannan-Quinn           312.2361
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 509 (89.9%)
Test del rapporto di verosimiglianza: Chi-quadro(4) = 156.442 [0.0000]
```

Figura 7.31. Output di *Gretl* relativo all'*EstrazioneEsatta80%-6* senza la variabile *pMax*

EstrazioneEsatta80%-7

Si riporta in **Figura 7.32** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-7*.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ---------------------------------------------------------
  flag = 1
  const       81.5369        9.95727        8.189   2.64e-016 ***
  materiale   -0.434807      0.108326      -4.014   5.97e-05  ***
  anno_posa   -0.0404652     0.00511473    -7.912   2.54e-015 ***
  diametro    -0.00742468    0.00248465    -2.988   0.0028    ***
  pMax         0.00732365    0.00715516     1.024   0.3060
  lunghezza    0.00532992    0.00174445     3.055   0.0022    ***

Media var. dipendente  0.144876   SQM var. dipendente    0.352287
Log-verosimiglianza   -154.3425   Criterio di Akaike     320.6849
Criterio di Schwarz    346.7165   Hannan-Quinn           330.8448
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 515 (91.0%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 159.643 [0.0000]
```

Figura 7.32. Output di *Gretl* relativo all'*EstrazioneEsatta80%-7*

Tramite la valutazione del *Test* Z, la variabile *"pMax"* risulta non signficativa. Di conseguenza, le stime dei parametri sono state nuovamente elaborate in assenza di tale variabile.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ---------------------------------------------------------
  flag = 1
  const       84.2628        10.0133        8.415   3.93e-017 ***
  materiale   -0.360400       0.0949421    -3.796   0.0001    ***
  anno_posa   -0.0418167      0.00511487   -8.176   2.95e-016 ***
  diametro    -0.00744576     0.00255457   -2.915   0.0036    ***
  lunghezza    0.00373733     0.00140167    2.666   0.0077    ***

Media var. dipendente  0.148410   SQM var. dipendente    0.355820
Log-verosimiglianza   -157.5586   Criterio di Akaike     325.1171
Criterio di Schwarz    346.8101   Hannan-Quinn           333.5837
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 514 (90.8%)
Test del rapporto di verosimiglianza: Chi-quadro(4) = 160.256 [0.0000]
```

Figura 7.33. Output di *Gretl* relativo all'*EstrazioneEsatta80%-7* senza la variabile *pMax*

L'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-7* risulta:

$$\hat{\pi}(x) = \frac{e^{(84.2628 - 00.0360400x_1 - 0.0418167x_2 - 0.0418167x_3 + 0.00373733x_5)}}{1 + e^{(84.2628 - 00.0360400x_1 - 0.0418167x_2 - 0.0418167x_3 + 0.00373733x_5)}}$$

EstrazioneEsatta80%-8

Si riporta in **Figura 7.34** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-8*.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ----------------------------------------------------------
  flag = 1
  const       86.2794        9.55824        9.027    1.77e-019 ***
  materiale   -0.355472       0.107835      -3.296    0.0010    ***
  anno_posa   -0.0430341      0.00492386    -8.740    2.33e-018 ***
  diametro    -0.00932502     0.00335957    -2.776    0.0055    ***
  pMax         0.00656625     0.00662479     0.9912   0.3216
  lunghezza    0.00392992     0.00170174     2.309    0.0209    **

Media var. dipendente  0.144876   SQM var. dipendente    0.352287
Log-verosimiglianza   -147.5067   Criterio di Akaike     307.0133
Criterio di Schwarz    333.0449   Hannan-Quinn           317.1732
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 516 (91.2%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 173.315 [0.0000]
```

Figura 7.34. Output di *Gretl* relativo all'*EstrazioneEsatta80%-8*

Tramite la valutazione del *Test* Z, la variabile *"pMax"* risulta non signficativa. Di conseguenza, le stime dei parametri sono state nuovamente elaborate in assenza di tale variabile.

L'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-8* risulta:

$$\hat{\pi}(x) = \frac{e^{(86.3141 - 0.381098x_1 - 0.042722x_2 - 0.0100277x_3 + 0.00379580x_5)}}{1 + e^{(86.3141 - 0.381098x_1 - 0.042722x_2 - 0.0100277x_3 + 0.00379580x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ---------------------------------------------------------
  flag = 1
  const       86.3141        9.54436        9.043   1.52e-019 ***
  materiale   -0.381098      0.103442      -3.684   0.0002    ***
  anno_posa   -0.0427222     0.00489878    -8.721   2.76e-018 ***
  diametro    -0.0100277     0.00335848    -2.986   0.0028    ***
  lunghezza    0.00379580    0.00170668     2.224   0.0261    **

Media var. dipendente  0.144876   SQM var. dipendente    0.352287
Log-verosimiglianza   -148.0368   Criterio di Akaike     306.0736
Criterio di Schwarz    327.7665   Hannan-Quinn           314.5401
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 513 (90.6%)
Test del rapporto di verosimiglianza: Chi-quadro(4) = 172.255 [0.0000]
```

Figura 7.35. Output di *Gretl* relativo all'*EstrazioneEsatta80%-8* senza la variabile *pMax*

EstrazioneEsatta80%-9

Si riporta in **Figura 7.36** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-9*.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ---------------------------------------------------------
  flag = 1
  const       75.3056        9.19390        8.191   2.59e-016 ***
  materiale   -0.286059      0.107322      -2.665   0.0077    ***
  anno_posa   -0.0374655     0.00473032    -7.920   2.37e-015 ***
  diametro    -0.0110099     0.00318647    -3.455   0.0005    ***
  pMax         0.00660735    0.00533226     1.239   0.2153
  lunghezza    0.00192985    0.00123139     1.567   0.1171

Media var. dipendente  0.143110   SQM var. dipendente    0.350494
Log-verosimiglianza   -159.3142   Criterio di Akaike     330.6283
Criterio di Schwarz    356.6599   Hannan-Quinn           340.7882
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 512 (90.5%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 146.135 [0.0000]
```

Figura 7.36. Output di *Gretl* relativo all'*EstrazioneEsatta80%-9*

Tramite la valutazione del *Test* Z, le variabili *"pMax"* e "lunghezza"risultano non signficative. Di conseguenza, le stime dei parametri sono state nuovamente elaborate in assenza di tali variabili.

L'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-9* risulta:

$$\hat{\pi}(x) = \frac{e^{(74.5667 - 0.319501x_1 - 0.0366722x_2 - 0.0117354x_3)}}{1 + e^{(74.5667 - 0.319501x_1 - 0.0366722x_2 - 0.0117354x_3)}}$$

```
Modello 2: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ----------------------------------------------------------
  flag = 1
  const        74.5667       9.14394        8.155   3.50e-016 ***
  materiale    -0.319501     0.101752      -3.140   0.0017    ***
  anno_posa    -0.0366722    0.00467697    -7.841   4.47e-015 ***
  diametro     -0.0117354    0.00317330    -3.698   0.0002    ***

Media var. dipendente  0.143110   SQM var. dipendente    0.350494
Log-verosimiglianza   -160.9712   Criterio di Akaike     329.9425
Criterio di Schwarz    347.2969   Hannan-Quinn           336.7158
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 510 (90.1%)
Test del rapporto di verosimiglianza: Chi-quadro(3) = 142.821 [0.0000]
```

Figura 7.37. Output di *Gretl* relativo all'*EstrazioneEsatta80%-9* senza le variabili *pMax* e *lunghezza*

EstrazioneEsatta80%-10

Si riporta in **Figura 7.38** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-10*. Tramite la valutazione del *Test* Z, tutte le variabili risultano significative. Di conseguenza, l'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-10* risulta:

$$\hat{\pi}(x) = \frac{e^{(74.2903 - 0.308517x_1 - 0.0374027x_2 - 0.00686481x_3 + 0.0130607x_4 + 0.00454991x_5)}}{1 + e^{(74.2903 - 0.308517x_1 - 0.0374027x_2 - 0.00686481x_3 + 0.0130607x_4 + 0.00454991x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ---------------------------------------------------------
  flag = 1
  const       74.2903        9.10216        8.162   3.30e-016 ***
  materiale   -0.308517      0.100531      -3.069   0.0021    ***
  anno_posa   -0.0374027     0.00469703    -7.963   1.68e-015 ***
  diametro    -0.00686481    0.00247146    -2.778   0.0055    ***
  pMax         0.0130607     0.00507840     2.572   0.0101    **
  lunghezza    0.00454991    0.00158408     2.872   0.0041    ***

Media var. dipendente  0.151943   SQM var. dipendente    0.359284
Log-verosimiglianza   -164.3916   Criterio di Akaike     340.7832
Criterio di Schwarz    366.8148   Hannan-Quinn           350.9431
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 510 (90.1%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 153.523 [0.0000]
```

Figura 7.38. Output di *Gretl* relativo all'*EstrazioneEsatta80%-10*

EstrazioneEsatta80%-11

Si riporta in **Figura 7.39** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-11*.

Figura 7.39. Output di *Gretl* relativo all'*EstrazioneEsatta80%-11*

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ---------------------------------------------------------
  flag = 1
  const       75.3789        8.85463        8.513   1.70e-017 ***
  materiale   -0.354178      0.107369      -3.299   0.0010    ***
  anno_posa   -0.0379059     0.00457676    -8.282   1.21e-016 ***
  diametro    -0.00663344    0.00238267    -2.784   0.0054    ***
  pMax         0.0131466     0.00519233     2.532   0.0113    **
  lunghezza    0.00505713    0.00154737     3.268   0.0011    ***

Media var. dipendente  0.148410   SQM var. dipendente    0.355820
Log-verosimiglianza   -158.3447   Criterio di Akaike     328.6893
Criterio di Schwarz    354.7209   Hannan-Quinn           338.8492
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 509 (89.9%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 158.684 [0.0000]
```

Tramite la valutazione del *Test* Z, tutte le variabili risultano significative. Di conseguenza, l'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-11* risulta:

$$\hat{\pi}(x) = \frac{e^{(75.3789 - 0.0354178x_1 - 0.0379059x_2 - 0.00663344x_3 + 0.0131466x_4 + 0.00505713x_5)}}{1 + e^{(75.3789 - 0.0354178x_1 - 0.0379059x_2 - 0.00663344x_3 + 0.0131466x_4 + 0.00505713x_5)}}$$

EstrazioneEsatta80%-12

Si riporta in **Figura 7.40** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-12*.

Figura 7.40. Output di *Gretl* relativo all'*EstrazioneEsatta80%-12*

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ----------------------------------------------------------
  flag = 1
  const       73.9914        9.23755        8.010   1.15e-015 ***
  materiale   -0.493583      0.123162      -4.008   6.13e-05  ***
  anno_posa   -0.0360855     0.00475964    -7.582   3.41e-014 ***
  diametro    -0.0132503     0.00338743    -3.912   9.17e-05  ***
  pMax         0.00865783    0.00518256     1.671   0.0948    *
  lunghezza    0.00439740    0.00143010     3.075   0.0021    ***

Media var. dipendente  0.157244   SQM var. dipendente    0.364352
Log-verosimiglianza   -161.1114   Criterio di Akaike     334.2228
Criterio di Schwarz    360.2544   Hannan-Quinn           344.3827
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 505 (89.2%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 170.278 [0.0000]
```

Tramite la valutazione del *Test* Z, la variabile *"pMax"* risulta non signficativa. Di conseguenza, le stime dei parametri sono state nuovamente elaborate in assenza di tale variabile.

L'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-12* risulta:

$$\hat{\pi}(x) = \frac{e^{(74.3494 - 0.517437x_1 - 0.0358619x_2 - 0.0142502x_3 + 0.00420661x_5)}}{1 + e^{(74.3494 - 0.517437x_1 - 0.0358619x_2 - 0.0142502x_3 + 0.00420661x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------------
  flag = 1
  const       74.3494        9.23036        8.055   7.96e-016 ***
  materiale   -0.517437      0.117700      -4.396   1.10e-05  ***
  anno_posa   -0.0358619     0.00473456    -7.574   3.61e-014 ***
  diametro    -0.0142502     0.00337578    -4.221   2.43e-05  ***
  lunghezza    0.00420661    0.00141299     2.977   0.0029    ***

Media var. dipendente  0.157244   SQM var. dipendente    0.364352
Log-verosimiglianza   -162.4265   Criterio di Akaike     334.8529
Criterio di Schwarz    356.5459   Hannan-Quinn           343.3195
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 507 (89.6%)
Test del rapporto di verosimiglianza: Chi-quadro(4) = 167.648 [0.0000]
```

Figura 7.41. Output di *Gretl* relativo all'*EstrazioneEsatta80%-12* senza la variabile *pMax*

EstrazioneEsatta80%-13

Si riporta in **Figura 7.42** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-13*.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------------
  flag = 1
  const       85.2853        10.1082        8.437   3.25e-017 ***
  materiale   -0.311803       0.104962     -2.971   0.0030    ***
  anno_posa   -0.0427204      0.00519090   -8.230   1.87e-016 ***
  diametro    -0.00901842     0.00280423   -3.216   0.0013    ***
  pMax         0.0121005      0.00527780    2.293   0.0219    **
  lunghezza    0.00348845     0.00159822    2.183   0.0291    **

Media var. dipendente  0.153710   SQM var. dipendente    0.360990
Log-verosimiglianza   -162.4133   Criterio di Akaike     336.8266
Criterio di Schwarz    362.8582   Hannan-Quinn           346.9865
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 514 (90.8%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 160.905 [0.0000]
```

Figura 7.42. Output di *Gretl* relativo all'*EstrazioneEsatta80%-13*

Tramite la valutazione del *Test* Z, tutte le variabili risultano significative. Di conseguenza, l'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-13* risulta:

$$\hat{\pi}(x) = \frac{e^{(85.2853 - 0.311803x_1 - 0.0427204x_2 - 0.00901842x_3 + 0.0121005x_4 + 0.00348845x_5)}}{1 + e^{(85.2853 - 0.311803x_1 - 0.0427204x_2 - 0.00901842x_3 + 0.0121005x_4 + 0.00348845x_5}}$$

EstrazioneEsatta80%-14

Si riporta in **Figura 7.43** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-14*.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ---------------------------------------------------------------
  flag = 1
  const       63.7717        9.00185        7.084   1.40e-012 ***
  materiale   -0.440937      0.115050      -3.833   0.0001    ***
  anno_posa   -0.0312074     0.00466967    -6.683   2.34e-011 ***
  diametro    -0.0129758     0.00383710    -3.382   0.0007    ***
  pMax         0.00888665    0.00520053     1.709   0.0875    *
  lunghezza    0.00437032    0.00151051     2.893   0.0038    ***

Media var. dipendente  0.143110   SQM var. dipendente    0.350494
Log-verosimiglianza   -158.3503   Criterio di Akaike     328.7007
Criterio di Schwarz    354.7322   Hannan-Quinn           338.8605
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 501 (88.5%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 148.063 [0.0000]
```

Figura 7.43. Output di *Gretl* relativo all'*EstrazioneEsatta80%-14*

Tramite la valutazione del *Test* Z, la variabile *"pMax"* risulta non signficativa. Di conseguenza, le stime dei parametri sono state nuovamente elaborate in assenza di tale variabile.
L'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-14* risulta:

$$\hat{\pi}(x) = \frac{e^{(64.3972 - 0.454151x_1 - 0.0311377x_2 - 0.0141086x_3 + 0.00416819x_5)}}{1 + e^{(64.3972 - 0.454151x_1 - 0.0311377x_2 - 0.0141086x_3 + 0.00416819x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ---------------------------------------------------------
  flag = 1
  const       64.3972        8.97529        7.175   7.23e-013 ***
  materiale   -0.454151      0.110810      -4.098   4.16e-05  ***
  anno_posa   -0.0311377     0.00464178    -6.708   1.97e-011 ***
  diametro    -0.0141086     0.00382998    -3.684   0.0002    ***
  lunghezza    0.00416819    0.00149848     2.782   0.0054    ***

Media var. dipendente  0.143110   SQM var. dipendente    0.350494
Log-verosimiglianza   -159.7547   Criterio di Akaike     329.5095
Criterio di Schwarz    351.2024   Hannan-Quinn           337.9760
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 501 (88.5%)
Test del rapporto di verosimiglianza: Chi-quadro(4) = 145.254 [0.0000]
```

Figura 7.44. Output di *Gretl* relativo all'*EstrazioneEsatta80%-14* senza la variabile *pMax*

EstrazioneEsatta80%-15

Si riporta in **Figura 7.45** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-15*.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ---------------------------------------------------------
  flag = 1
  const       75.3564        8.87763        8.488   2.10e-017 ***
  materiale   -0.281149      0.100719      -2.791   0.0052    ***
  anno_posa   -0.0380974     0.00458725    -8.305   9.98e-017 ***
  diametro    -0.00628685    0.00242313    -2.595   0.0095    ***
  pMax         0.0140326     0.00562317     2.495   0.0126    **
  lunghezza    0.00222306    0.00131170     1.695   0.0901    *

Media var. dipendente  0.143110   SQM var. dipendente    0.350494
Log-verosimiglianza   -157.0223   Criterio di Akaike     326.0447
Criterio di Schwarz    352.0762   Hannan-Quinn           336.2046
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 513 (90.6%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 150.719 [0.0000]
```

Figura 7.45 Output di *Gretl* relativo all'*EstrazioneEsatta80%-15*

Tramite la valutazione del *Test* Z, la variabile *"lunghezza"* risulta non signficativa. Di conseguenza, le stime dei parametri sono state nuovamente elaborate in assenza di tale

variabile. L'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-15* risulta:

$$\hat{\pi}(x) = \frac{e^{(74.7327 - 0.269216x_1 - 0.0377527 - 0.00609642x_3 + 0.0139428x_4)}}{1 + e^{(74.7327 - 0.269216x_1 - 0.0377527 - 0.00609642x_3 + 0.0139428x_4)}}$$

Figura 7.45 Output di *Gretl* relativo all'*EstrazioneEsatta80%-15*

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.     z       p-value
  ---------------------------------------------------------
  flag = 1
  const       75.3564        8.87763        8.488   2.10e-017 ***
  materiale   -0.281149      0.100719      -2.791   0.0052    ***
  anno_posa   -0.0380974     0.00458725    -8.305   9.98e-017 ***
  diametro    -0.00628685    0.00242313    -2.595   0.0095    ***
  pMax         0.0140326     0.00562317     2.495   0.0126    **
  lunghezza    0.00222306    0.00131170     1.695   0.0901    *

Media var. dipendente  0.143110   SQM var. dipendente    0.350494
Log-verosimiglianza   -157.0223   Criterio di Akaike     326.0447
Criterio di Schwarz    352.0762   Hannan-Quinn           336.2046
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 513 (90.6%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 150.719 [0.0000]
```

EstrazioneEsatta80%-16

Si riporta in **Figura 7.47** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-16*. Tramite la valutazione del *Test* Z, tutte le variabili risultano significative. Di conseguenza, l'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-16* risulta:

$$\hat{\pi}(x) = \frac{e^{(93.7207 - 0.376575x_1 - 0.0467038x_2 - 0.0121561x_3 + 0.0125492x_4 + 0.00473270x_5)}}{1 + e^{(93.7207 - 0.376575x_1 - 0.0467038x_2 - 0.0121561x_3 + 0.0125492x_4 + 0.00473270x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------------
  flag = 1
  const       93.7207        10.8653        8.626   6.37e-018 ***
  materiale   -0.376575       0.116125     -3.243   0.0012    ***
  anno_posa   -0.0467038      0.00557031   -8.384   5.10e-017 ***
  diametro    -0.0121561      0.00377194   -3.223   0.0013    ***
  pMax         0.0125492      0.00540160    2.323   0.0202    **
  lunghezza    0.00473270     0.00177742    2.663   0.0078    ***

Media var. dipendente  0.137809   SQM var. dipendente    0.345005
Log-verosimiglianza   -132.9678   Criterio di Akaike     277.9356
Criterio di Schwarz    303.9671   Hannan-Quinn           288.0955
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 526 (92.9%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 187.958 [0.0000]
```

Figura 7.47. Output di *Gretl* relativo all'*EstrazioneEsatta80%-16*

EstrazioneEsatta80%-17

Si riporta in **Figura 7.48** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-17*.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------------
  flag = 1
  const       101.863        11.3866        8.946   3.69e-019 ***
  materiale    -0.261608      0.0989006    -2.645   0.0082    ***
  anno_posa    -0.0515981     0.00583867   -8.837   9.81e-019 ***
  diametro     -0.00599966    0.00229569   -2.613   0.0090    ***
  pMax          0.0156078     0.00544526    2.866   0.0042    ***
  lunghezza     0.00451411    0.00175519    2.572   0.0101    **

Media var. dipendente  0.151943   SQM var. dipendente    0.359284
Log-verosimiglianza   -145.8385   Criterio di Akaike     303.6769
Criterio di Schwarz    329.7085   Hannan-Quinn           313.8368
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 523 (92.4%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 190.629 [0.0000]
```

Figura 7.48. Output di *Gretl* relativo all'*EstrazioneEsatta80%-17*

Tramite la valutazione del *Test* Z, tutte le variabili risultano significative. Di conseguenza, l'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-17* risulta:

$$\hat{\pi}(x) = \frac{e^{(101.863 - 0.261608x_1 - 0.0515981x_2 - 0.0099966x_3 + 0.0156078x_4 + 0.00451411x_5)}}{1 + e^{(101.863 - 0.261608x_1 - 0.0515981x_2 - 0.0099966x_3 + 0.0156078x_4 + 0.00451411x_5)}}$$

EstrazioneEsatta80%-18

Si riporta in **Figura 7.49** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-18*.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------------
  flag = 1
  const       86.1310        10.3499        8.322   8.66e-017 ***
  materiale   -0.417976       0.110597     -3.779   0.0002    ***
  anno_posa   -0.0430585      0.00533831   -8.066   7.27e-016 ***
  diametro    -0.00574886     0.00213445   -2.693   0.0071    ***
  pMax         0.0118658      0.00570631    2.079   0.0376    **
  lunghezza    0.00429282     0.00170796    2.513   0.0120    **

Media var. dipendente  0.146643   SQM var. dipendente    0.354063
Log-verosimiglianza   -146.8311   Criterio di Akaike     305.6622
Criterio di Schwarz    331.6937   Hannan-Quinn           315.8220
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 516 (91.2%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 178.203 [0.0000]
```

Figura 7.49. Output di *Gretl* relativo all'*EstrazioneEsatta80%-18*

Tramite la valutazione del *Test* Z, tutte le variabili risultano significative. Di conseguenza, l'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-18* risulta:

$$\hat{\pi}(x) = \frac{e^{(86.1310 - 0.417976x_1 - 0.0430585x_2 - 0.00574886x_3 + 0.0118658x_4 + 0.00429282x_5)}}{1 + e^{(86.1310 - 0.417976x_1 - 0.0430585x_2 - 0.00574886x_3 + 0.0118658x_4 + 0.00429282x_5)}}$$

EstrazioneEsatta80%-19

Si riporta in **Figura 7.50** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-19*. Tramite la valutazione del *Test* Z, tutte le variabili risultano significative. Di conseguenza, l'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-19* risulta:

$$\hat{\pi}(x) = \frac{e^{(81.2593 - 0.393566x_1 - 0.0407714x_2 - 0.00719067x_3 + 0.0167052x_4 + 0.00547607x_5)}}{1 + e^{(81.2593 - 0.393566x_1 - 0.0407714x_2 - 0.00719067x_3 + 0.0167052x_4 + 0.00547607x_5)}}$$

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------------
  flag = 1
  const       81.2593        9.86684        8.236   1.79e-016 ***
  materiale   -0.393566      0.121828      -3.231   0.0012    ***
  anno_posa   -0.0407714     0.00512018    -7.963   1.68e-015 ***
  diametro    -0.00719067    0.00249367    -2.884   0.0039    ***
  pMax         0.0167052     0.00568926     2.936   0.0033    ***
  lunghezza    0.00547607    0.00180164     3.039   0.0024    ***

Media var. dipendente  0.144876   SQM var. dipendente    0.352287
Log-verosimiglianza   -149.1746   Criterio di Akaike     310.3491
Criterio di Schwarz    336.3807   Hannan-Quinn           320.5090
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 519 (91.7%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 169.979 [0.0000]
```

Figura 7.50. Output di *Gretl* relativo all'*EstrazioneEsatta80%-19*

EstrazioneEsatta80%-20

Si riporta in **Figura 7.51** l'output fornito da *Gretl* nel caso dell'*EstrazioneEsatta80%-20*.

```
Modello 1: Logit multinomiale, usando le osservazioni 1-566
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------------
  flag = 1
  const       87.1647        10.3852        8.393   4.73e-017 ***
  materiale   -0.412477       0.106614     -3.869   0.0001    ***
  anno_posa   -0.0434718      0.00534912   -8.127   4.40e-016 ***
  diametro    -0.00798155     0.00275701   -2.895   0.0038    ***
  pMax         0.0138132      0.00586051    2.357   0.0184    **
  lunghezza    0.00372441     0.00167970    2.217   0.0266    **

Media var. dipendente  0.153710   SQM var. dipendente    0.360990
Log-verosimiglianza   -150.7426   Criterio di Akaike     313.4853
Criterio di Schwarz    339.5168   Hannan-Quinn           323.6452
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 516 (91.2%)
Test del rapporto di verosimiglianza: Chi-quadro(5) = 184.246 [0.0000]
```

Figura 7.51. Output di *Gretl* relativo all'*EstrazioneEsatta80%-20*

Tramite la valutazione del *Test* Z, tutte le variabili risultano significative. Di conseguenza, l'equazione utilizzata per stimare la probabilità di rottura delle condotte facenti parte dell'*EstrazioneEsatta20%-20* risulta:

$$\hat{\pi}(x) = \frac{e^{(87.1647 - 0.412477x_1 - 0.0434718x_2 - 0.00798155x_3 + 0.0138132x_4 + 0.00372441x_5)}}{1 + e^{(87.1647 - 0.412477x_1 - 0.0434718x_2 - 0.00798155x_3 + 0.0138132x_4 + 0.00372441x_5)}}$$

Si riportano, infine, i grafici a bolle relativi all'applicazione dei modelli appena stimati. Nello specifico, è stata calcolata la probabilità di rottura per tutte le condotte facenti parte di ogni *EstrazioneEsatta20%* e le probabilità medie di rottura per le condotte con flag pari a 0 e 1, definite rispettivamente da bolle grigie e gialle.
Inoltre, in **Tabella 7.1** sono riassunte tali probabilità medie di rottura. È molto importante notare che, per ogni estrazione, la probabilità media di rottura riferita alle condotte non guaste è sempre inferiore alla probabilità media di rottura di quelle che sono andate realmente incontro ad uno o più guasti nel periodo di osservazione.

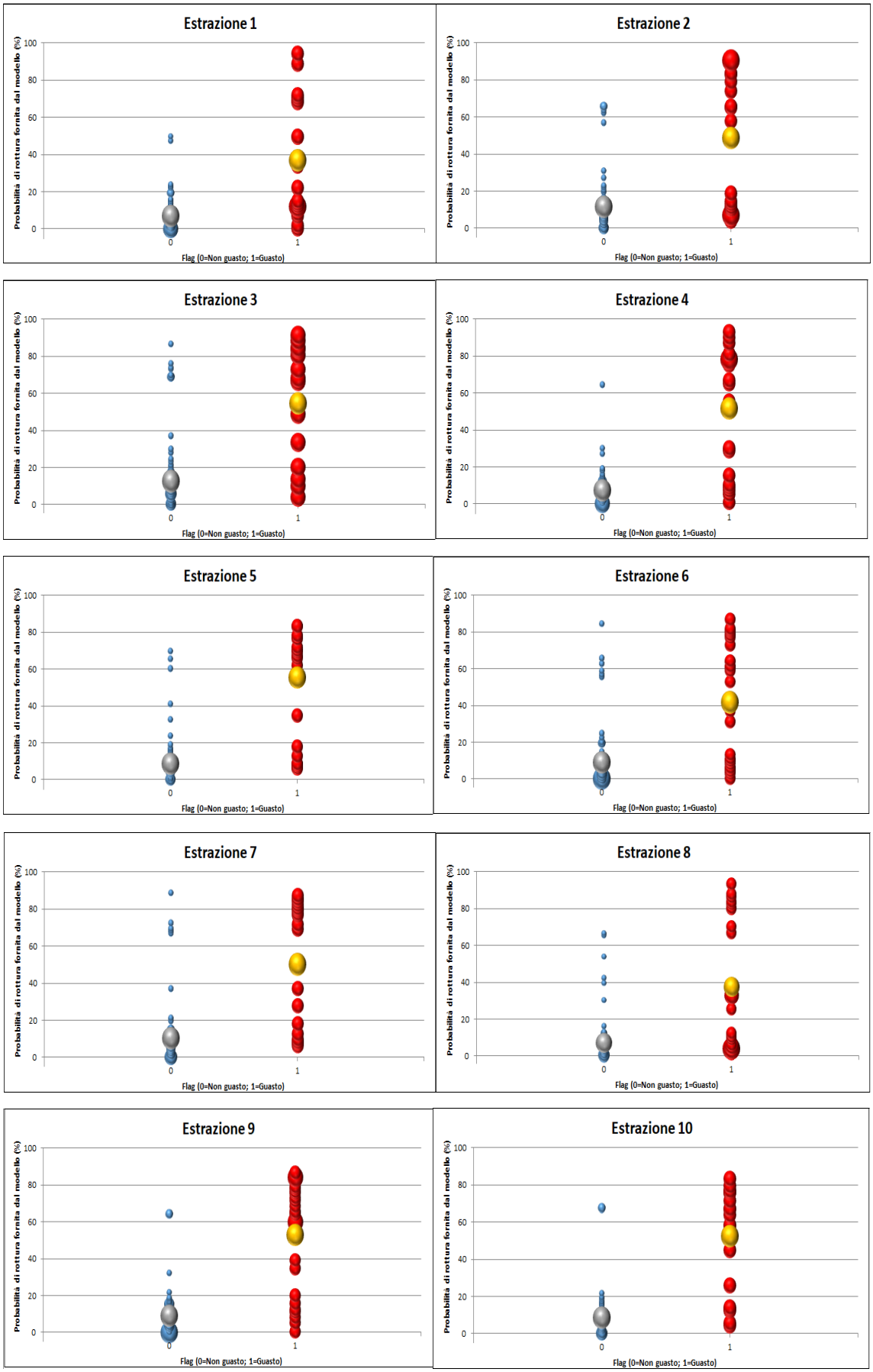


Figura 7.52. Grafici a bolle relativi alle prime 10 *EstrazioneEsatta20%*

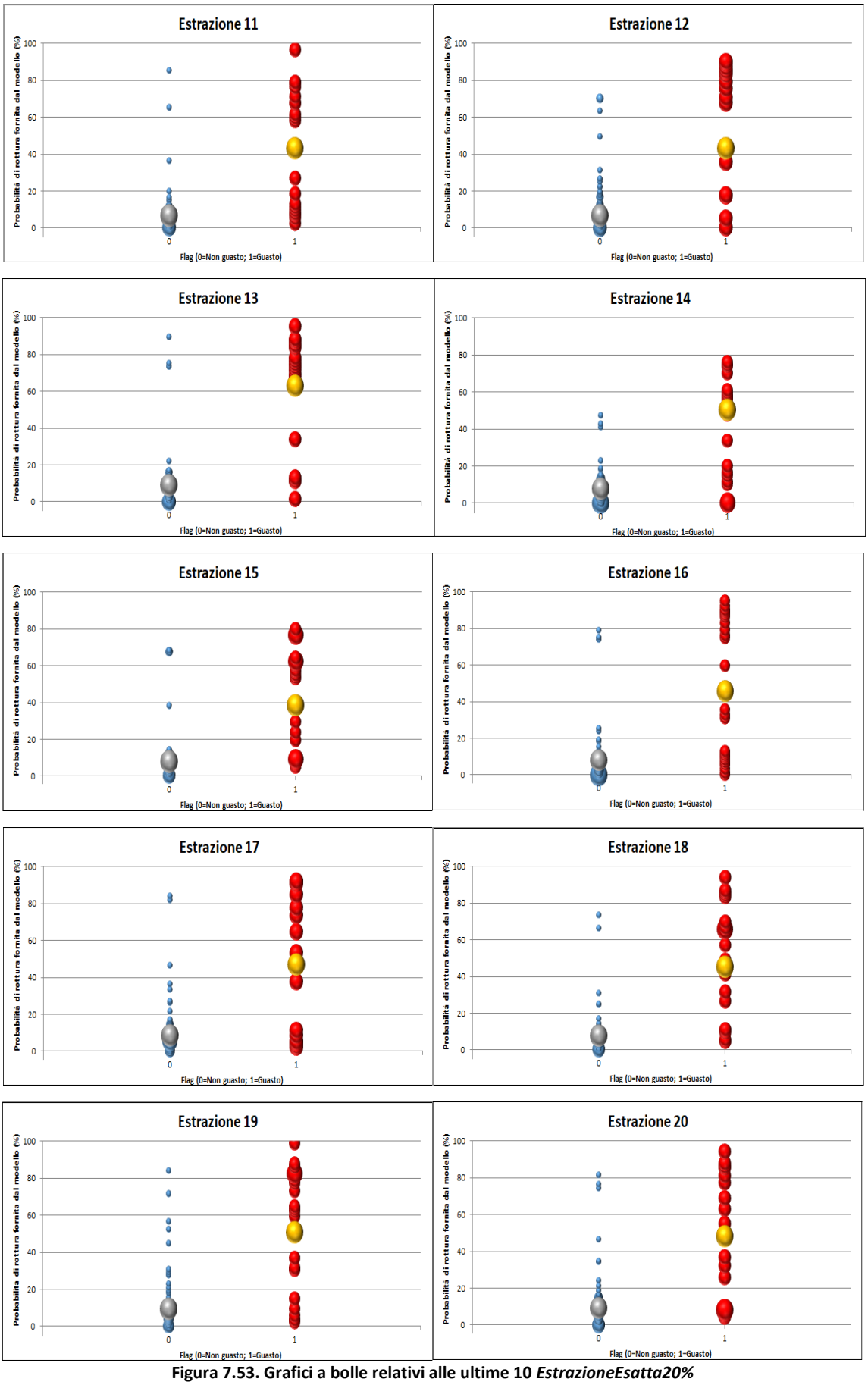


Figura 7.53. Grafici a bolle relativi alle ultime 10 *EstrazioneEsatta20%*

Tabella 7.1. Valori medi delle probabilità stimate per le condotte con flag pari a 0 e 1 di ogni *EstrazioneEsatta20%*

Estrazione	Probabilità media stimata Flag=0 (%)	Probabilità media stimata Flag=1 (%)
1	6.95	36.83
2	11.42	48.63
3	12.59	54.64
4	7.49	51.60
5	8.70	55.51
6	8.95	41.70
7	10.33	50.10
8	7.23	37.53
9	9.10	52.82
10	8.72	52.51
11	6.73	43.07
12	9.96	62.49
13	8.88	62.87
14	7.78	50.21
15	7.98	38.42
16	7.89	45.63
17	8.59	46.93
18	7.61	45.21
19	9.48	50.78
20	9.26	48.35
Media	**8.78**	**48.79**

Dai grafici a bolle in **Figura 7.52-53** è possibile osservare che i modelli stimano delle probabilità di rottura in media più elevate per le condotte realmente rotte nel periodo di osservazione, confrontati con i valori che caratterizzano le condotte non guaste.
Si riporta, infine, in **Figura 7.54** il grafico a bolle relativo alla totalità dei dati caratterizzanti ogni *EstrazioneEsatta20%*. A causa della grande dispersività dei dati e per una migliore resa grafica, le percentuali sono state suddivise in classi di larghezza 1%. Anche in questo caso si osserva che la stima della percentuale media totale relativa alle condotte rotte è maggiore di quella relativa alle condotte mai guastate nel periodo di osservazione (valori riportati anche in **Tabella 7.1**).
Confrontando tali probabilità medie con quelle ricavate nel lavoro precedente, dove ogni condotta era suddivisa in tratti e la lunghezza non rappresentava una variabile indipendente del modello, i valori risultano differenti:

- nel primo caso studio, la probabilità media associata ai tratti non rotti ammontava a 2.21%, mentre in questo caso è pari a 8.78%;
- nel primo caso studio la probabilità media associata ai tratti rotti ammontava a 5.93%, mentre in questo caso è pari a 48.79%.

Da questi valori si deduce che, in generale, il modello in esame associa delle probabilità più alte di rottura alle condotte, ma sembra associare valori molto più elevati a quelle realmente rotte, se si confrontano i risultati ottenuti nei due lavori di tesi. Il modello così composto, in definitiva, riconosce in modo più determinante le condotte realmente rotte.

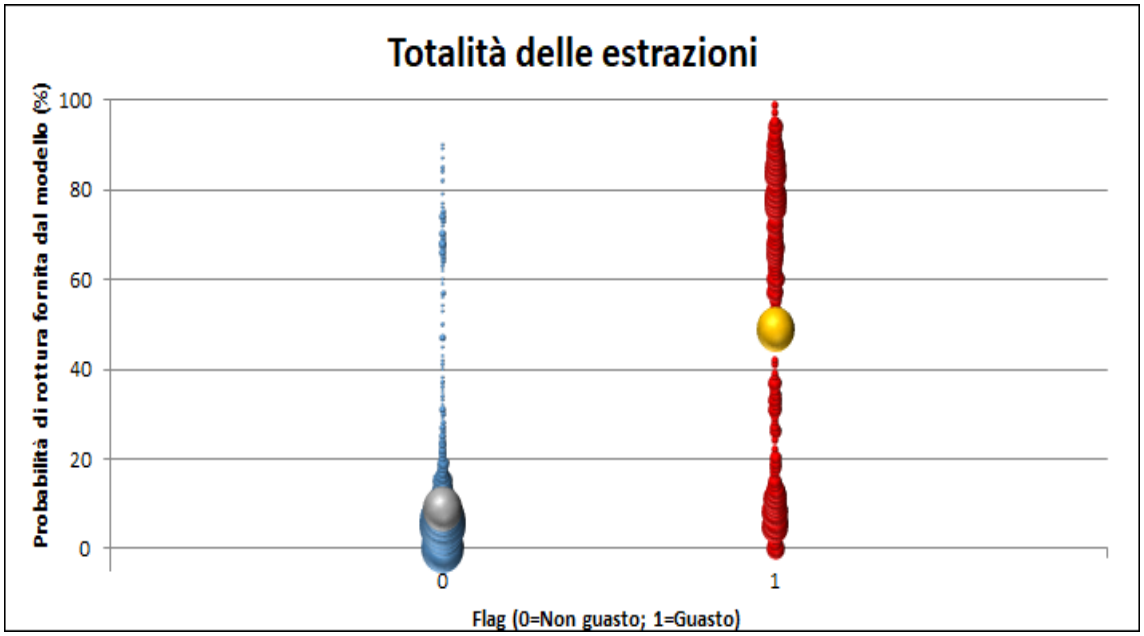


Figura 7.54. Grafico a bolle relativo alla totalità dei dati costituenti ogni *EstrazioneEsatta20%*

7.3. Valore di soglia (*cut-off*) della probabilità

Una volta ottenute le probabilità di rottura per tutti gli elementi delle estrazioni, è di fondamentale importanza definire un valore di probabilità di soglia al di là del quale una condotta può considerarsi suscettibile di rottura nei 10 anni successivi al periodo di osservazione.

La definizione di tale valore rappresenta il limite di separazione tra la positività e la negatività del test e influenza in maniera determinante la probabilità di ottenere dei falsi negativi e falsi positivi. Per esempio, definendo una probabilità di soglia molto bassa, si ottengono numerosi falsi positivi, in quanto si considerano guaste condotte in realtà integre e caratterizzate da valori di probabilità più elevati di quello di *cut-off*. D'altro canto, un valore di soglia molto alto comporta una valutazione errata in termini di falsi negativi; infatti, vengono considerate integre condotte che in realtà hanno riportato dei guasti nel periodo di osservazione. Tali elementi presentano probabilità di rottura inferiori a quelle di soglia, nonostante si sia verificato un guasto. In conclusione, la scelta del valore di *cut-off* deve risultare un valore di compromesso che permette di ottimizzare e ridurre al minimo i falsi negativi e i falsi positivi. Per tale ragione, sono stati seguiti i seguenti passi:

1. È stato tarato un nuovo modello di regressione logistica globale sull'intera tabella *Esatta.*

2. Tale modello è stato applicato a tutte le condotte della tabella *Esatta,* per valutarne la relativa probabilità di rottura.
3. Sono stati ipotizzati diversi valori di probabilità di soglia.
4. È stato valutato il numero di condotte con flag pari a 0 e probabilità stimata di rottura maggiore o uguale al valore di soglia; tali elementi risultano dei falsi positivi (riportati come *FP*), in quanto suscettibili a rottura, secondo il modello, ma integri nel periodo di osservazione 2006-2016.
5. È stato valutato il numero di condotte con flag pari a 1 e probabilità stimata di rottura inferiore al valore di soglia; tali elementi risultano dei falsi negativi (riportati come *FN*), poiché sono considerati dal modello come non suscettibili di rottura, ma nella realtà hanno presentato una o più rotture nel periodo di osservazione.
6. Si sceglie il valore di soglia ottimo, tale da limitare la presenza di falsi positivi e falsi negativi ed ottimizzare la loro percentuale rispetto al totale.

Si definiscono, in prima istanza, le variabili del modello:

- la variabile risposta dipendente *Y*, condizionata alla rottura o non rottura della generica condotta (rispettivamente flag pari a 1 o 0);
- la variabile indipendente x_1 "*materiale"*;
- la variabile indipendente x_2 *"anno di posa"*;
- la variabile indipendente x_3 *"diametro"*;
- la variabile indipendente x_4 "*pMax*", relativa al carico massimo in condotta. Tale variabile risulta non significativa e non è stata presa in esame;
- la variabile indipendente x_5 *"lunghezza*".

Il modello generale di regressione logistica sull'intero insieme di dati contenuti nella tabella *Esatta* presenta le caratteristiche riportate in **Figura 7.55**. Il modello di regressione logistica multivariata assume la seguente forma:

$$logit = 84.1473 - 0.4424x_1 - 0.0424x_2 - 0.0117x_3 + 0.004x_5$$

La corrispondente probabilità di rottura stimata è calcolata come:

$$\hat{\pi}(x) = \frac{\mathrm{e}^{(84.1473 - 0.4424x_1 - 0.0424x_2 - 0.0117x_3 + 0.004x_5)}}{1 + \mathrm{e}^{(84.1473 - 0.4424x_1 - 0.0424x_2 - 0.0117x_3 + 0.004x_5)}}$$

Si riporta, inoltre, un confronto tra le stime dei parametri di tutte le *EstrazioneEsatta80%* e quelle relative al modello che considera l'intera tabella *Esatta*. Nei 21 modelli stimati, i parametri significativi e di interesse sono caratterizzati da medesimi segni di positività e negatività matematica e presentano ordini di grandezza comparabili.

```
Modello 2: Logit multinomiale, usando le osservazioni 1-5855
Variabile dipendente: flag
Errori standard basati sull'Hessiana

             coefficiente   errore std.      z       p-value
  ---------------------------------------------------------------
  flag = 1
  const       84.1473        5.76126        14.61    2.58e-048 ***
  materiale   -0.442421      0.0659391      -6.710   1.95e-011 ***
  anno_posa   -0.0423892     0.00294527    -14.39    5.79e-047 ***
  diametro    -0.0117093     0.00256312     -4.568   4.91e-06  ***
  lunghezza    0.00398938    0.000921348     4.330   1.49e-05  ***

Media var. dipendente  0.018104   SQM var. dipendente    0.133340
Log-verosimiglianza   -377.6133   Criterio di Akaike     765.2267
Criterio di Schwarz    798.6019   Hannan-Quinn           776.8312
Note: SQM = scarto quadratico medio; E.S. = errore standard

Numero dei casi 'previsti correttamente' = 5746 (98.1%)
Test del rapporto di verosimiglianza: Chi-quadro(4) = 305.304 [0.0000]
```

Figura 7.55. Output di *Gretl* relativo al modello stimato sulla tabella *Esatta*

Tabella 7.2. Confronto tra i parametri stimati per ogni *EstrazioneEsatta80%* e per il modello generale

EstrazioneEsatta80%	Costante	x_1	x_2	x_3	x_4	x_5
1	86.9938	-0.4537	-0.4320	-0.0108	0.0166	0.0054
2	72.2308	-0.4180	-0.0355	-0.0106	/	0.0054
3	79.7853	-0.3622	-0.0400	-0.0077	0.0142	0.0074
4	82.4200	-0.3202	-0.0413	-0.0081	-0.0082	0.0049
5	81.7982	-0.4445	-0.0403	-0.0079	/	0.0052
6	75.1496	-0.4481	-0.0369	-0.0104	/	0.0042
7	84.2628	-0.3604	-0.0418	-0.0740	/	0.0014
8	86.3141	-0.3811	-0.0427	-0.0100	/	0.0038
9	74.5667	-0.3195	-0.0367	-0.0117	/	/
10	74.2903	-0.3085	-0.0374	-0.0069	0.0131	0.0045
11	75.3789	-0.3542	-0.0379	-0.0066	0.0131	0.0051
12	74.3494	-0.5174	0.0359	-0.0143	/	0.0042
13	85.2853	-0.3118	-0.0427	-0.0090	0.0121	0.0035
14	64.3972	-0.4542	-0.0311	-0.0141	/	0.0042
15	75.3564	-0.2811	-0.0381	-0.0063	0.0140	0.0022
16	93.7207	-0.3766	-0.0467	-0.0121	0.0126	0.0047
17	101.8630	-0.2616	-0.0516	-0.0060	0.0156	0.0045
18	86.1310	-0.4180	-0.0431	-0.0057	0.0119	0.0043
19	81.2593	-0.3935	-0.0408	-0.0072	0.0167	0.0055
20	87.1647	-0.4124	-0.0435	-0.0080	0.0139	0.0037
Media	**81.1359**	**-0.3799**	**-0.0562**	**-0.0124**	**0.0121**	**0.0044**
Modello generale	84.1473	-0.4424	-0.0424	-0.0117	/	0.0040

Il modello generale è stato successivamente applicato alla tabella *Esatta*, per stimare la probabilità di rottura di ogni condotta.
Una volta fissati arbitrariamente i valori di soglia nell'intervallo di variazione tra il 10% e il 30%, è stato calcolato il numero di falsi positivi, falsi negativi e la loro somma (**Tabella 7.3**). Il valore di soglia ottimale è quello che minimizza la somma totale dei falsi e, come è possibile notare nella tabella sottostante, tale percentuale è pari al 27-28%, con un totale di falsi positivi e falsi negativi pari a 83.

Tabella 7.3. Numero di falsi positivi e falsi negativi al variare del valore di soglia

Cut-off (%)	FP	FN	Totale
10	194	48	242
11	189	48	237
12	183	48	231
13	182	48	230
14	177	49	226
15	177	53	230
16	177	55	232
17	173	55	228
18	152	55	207
19	92	55	147
20	48	57	105
21	43	61	104
22	33	64	97
23	28	67	95
24	20	68	88
25	16	69	85
26	15	69	84
27	13	70	83
28	12	71	83
29	12	72	84
30	11	84	95

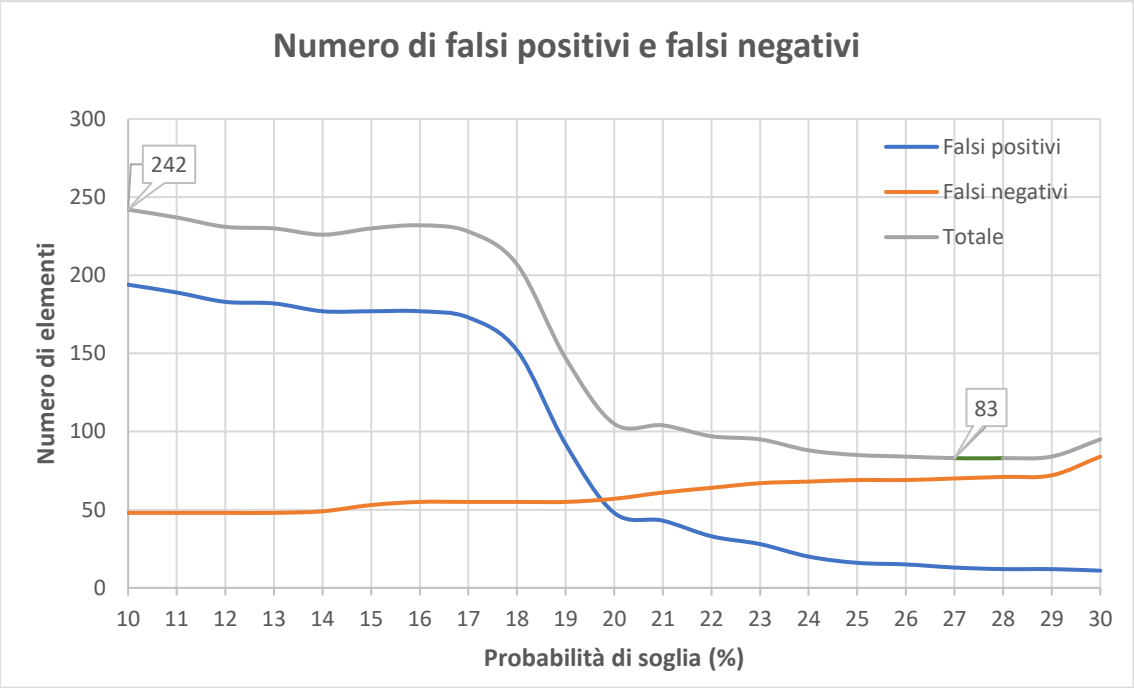


Figura 7.56. Andamento del numero di falsi positivi e negativi al variare della probabilità di soglia

7.5. Conclusioni

In questo capitolo è stato stimato e applicato il modello di regressione logistica su un determinato numero di estrazioni, al fine di valutare dapprima la significatività delle variabili indipendenti quali *"materiale"*, *"anno di posa"*, *"diametro"*, "*pMax"* e *"lunghezza"* e, successivamente, la bontà del modello nel prevedere le rotture delle condotte.
Il primo passo è stato quello di definire 20 estrazioni a partire dalla tabella *Esatta*, costituita da 5855 righe e 12 colonne, tra cui 106 rotture e 5749 condotte mai rotte nell'arco temporale tra il 2006 e il 2016. Queste estrazioni sono contraddistinte da una percentuale di rottura pari a quella della tabella fondamentale *ReteTorino*. Sono stati seguiti i seguenti passi:

- a partire dalla tabella *Esatta* sono state estratte le 106 condotte rotte con flag pari all'unità;
- a partire dalla tabella *Esatta* sono stati estratti in maniera casuale 20 campioni costituiti da 601 condotte con flag pari a 0;
- l'unione tra i 20 campioni di condotte non rotte e le 106 condotte con flag unitario restituisce 20 estrazioni contraddistinte da una percentuale di rottura pari al 15% e denominate *EstrazioneEsatta*.

Per ogni *EstrazioneEsatta* è stato tarato un modello di regressione logistica. Una volta esaminati i risultati di *Gretl* riferiti a ogni *EstrazioneEsatta*, è stato possibile riscontrare che:

- il carico massimo "*pMax"* risulta non significativo in una sotto-estrazione. Nella quasi totalità dei casi, risulta comunque meno significativo delle altre variabili;
- in tutte le estrazioni, le variabili "*materiale*", "*anno di posa*", "*diametro*" e "l*unghezza*" risultano rilevanti nel modello;
- la probabilità di rottura diminuisce all'aumentare del numero assegnato ad ogni materiale, dell'anno di posa e del diametro;
- la probabilità di rottura aumenta all'aumentare del carico massimo e della lunghezza della condotta.

Successivamente, per poter valutare la bontà di previsione dei modelli nei riguardi di una rottura, da ogni *EstrazioneEsatta* è stato estratto l'80% delle condotte. Questa aggiuntiva estrazione genera 20 sotto-estrazioni denominate *EstrazioneEsatta80%* (566 condotte) ed ulteriori sotto-estrazioni denominate *EstrazioneEsatta20%* (141 condotte), costituite dagli elementi esclusi da ogni *EstrazioneEsatta80%*. Nello specifico, per verificare la bontà di questi modelli nel prevedere una rottura, sono stati adottati i seguenti passi:

1. Sono state considerate le venti estrazioni esaminate nel paragrafo precedente (*EstrazioneEsatta)*.
2. Da ognuna di queste estrazioni è stato estratto in maniera casuale l'80% degli elementi, cioè 566 condotte (*EstrazioneEsatta80%)*.
3. Sono state calcolate le nuove stime dei parametri di regressione logistica per ogni *EstrazioneEsatta80%*. Queste stime sono state effettuate prendendo in considerazione tutte le variabili indipendenti citate in precedenza e, nel caso in cui una variabile sia risultata non significativa, sono stati nuovamente stimati i parametri eliminando la variabile citata.
4. Per ogni estrazione, il modello ottenuto considerando l'80% degli elementi è stato successivamente applicato al restante 20% dell'estrazione. Ognuna di queste sotto-estrazioni, composte da 141 condotte, è stata denominata *EstrazioneEsatta20%.*
7. L'applicazione del punto *4* consente di ottenere le probabilità di rottura stimate per le condotte facenti parte di ogni *EstrazioneEsatta20%*.
8. Infine, per valutare la capacità del modello di predire una rottura, è stato creato un grafico a bolle per ogni *EstrazioneEsatta20%* e sono state comparate le probabilità stimate per le condotte effettivamente rotte e quelle non rotte. Un buon modello, per essere tale, dovrebbe stimare una maggiore probabilità di rottura per le condotte con flag pari a 1.

Una volta esaminati i risultati delle venti estrazioni *EstrazioneEsatta80%*, è stato possibile riscontrare che:

- il carico massimo "*pMax"* risulta non significativo in 9 sotto-estrazioni su 20. Nella quasi totalità dei casi, risulta comunque meno significativo delle altre variabili;
- la variabile *"lunghezza"* risulta non significativa in due sotto-estrazioni su 20;
- in tutte le estrazioni, le variabili "*materiale*", "*anno di posa*" e "*diametro*" risultano rilevanti nel modello;
- la probabilità di rottura diminuisce all'aumentare del numero assegnato ad ogni materiale, dell'anno di posa e del diametro;
- la probabilità di rottura aumenta all'aumentare del carico massimo e della lunghezza della condotta.

Successivamente, esaminando i grafici a bolle ottenuti dall'applicazione dei modelli di regressione logistica su ogni *EstrazioneEsatta20%* è possibile affermare che:

- per ogni estrazione, la maggior parte delle condotte con flag pari a 0 presenta probabilità di rottura inferiori rispetto agli stessi valori calcolati per le condotte con flag pari a 1;

- in conclusione, i modelli associano alle condotte con flag pari a 1 una probabilità stimata media di rottura maggiore rispetto alla probabilità media calcolata per le condotte con flag pari a 0. Tali probabilità medie, se confrontate con il caso studio precedente, mostrano che il modello così composto riconosce in maniera migliore le condotte realmente rotte, poiché ne associa probabilità di rottura molto più elevate (48.79% in questo caso e 5.93% nel primo caso studio). La nuova struttura del modello risulta quindi più idonea a rappresentare i dati raccolti da *SMAT*.

Infine, è stato valutato il valore di soglia oltre il quale considerare una condotta suscettibile di rottura. Tale valore è definito a partire dall'analisi del numero di falsi positivi e falsi negativi, al variare dello stesso. Il valore di *cut-off* ottimale si attesta attorno al 27-28% e minimizza il numero di falsi a 83.

Capitolo 8

Applicazione del modello di regressione polinomiale

In questo capitolo verranno descritti i risultati ottenuti dall'applicazione del modello di regressione polinomiale, esposto nel **Capitolo 3**, sulla tabella *Esatta Finale*. Questa contiene tutte le condotte con la totale copertura delle informazioni riguardanti diametro, materiale, anno di posa, carico massimo e lunghezza totale delle stesse. L'applicazione di tale modello consente di definire, a partire da una suddivisione in classi delle condotte, il numero di rotture previste (*NR*), in funzione delle variabili esplicative. La variabile dipendente *NR* sarà quindi funzione delle caratteristiche delle condotte quali:

- diametro;
- materiale;
- anno di posa;
- lunghezza.

A differenza di quanto fatto nell'applicazione del modello di regressione logistica, in questo caso non verrà analizzata l'eventuale influenza del carico massimo sulla variabile indipendente, ma saranno prese in considerazione le sole variabili suggerite nella pubblicazione *"Development of pipe deterioration models for water distribution systems using EPR"* di *Berardi & al.* (*2008*).
Anche l'informazione sul materiale sarà introdotta in maniera differente: infatti, questo non sarà inserito come variabile esplicativa nel modello, ma verrà utilizzato come criterio preliminare di raggruppamento delle condotte. I materiali studiati, in accordo con quanto fatto nel capitolo precedente, saranno l'acciaio, l'eternit, la ghisa grigia e la ghisa sferoidale. A seguito del passaggio dalla tabella *ReteTorino* alla tabella *Esatta,* il PEAD non presenta condotte rotte ed è quindi di scarsa utilità per fini statistici.Anche in questo caso, il punto di partenza per la costruzione del modello è la tabella *Esatta*, costituita da 5855 righe, contraddistinte da 106 guasti nell'arco temporale tra il 2006 e il 2016 e una percentuale di rotture pari all'1.8%. Nei successivi paragrafi verranno seguiti i passi esposti nel **Capitolo 3**:

1. A partire dal database contenente le condotte della rete in esame, complete di informazioni riguardanti il diametro, materiale, anno di posa e lunghezza delle stesse, saranno estratte delle sotto-tabelle riguardanti i singoli materiali. Nel caso in esame, il campione di partenza è la tabella *Esatta* e i materiali analizzati sono la ghisa grigia, ghisa sferoidale, eternit ed acciaio. Le conseguenti sotto-tabelle saranno denominate come *EsattaGhisaGrigia*, *EsattaGhisaSferoidale*, *EsattaEternit* ed *EsattaAcciaio.*
2. Per tutte le sotto-tabelle sarà effettuata una suddivisione in classi: a seguito di tale stratificazione, ogni classe sarà caratterizzata da un diametro e da un'età equivalenti pesati

rispetto alla lunghezza. La suddivisione deve essere effettuata in modo tale che ogni classe contenga un numero comparabile di condotte, per poter assicurare una valenza statistica. Affinché questo sia rispettato, nel caso di reti complesse, sarà necessario studiare la distribuzione di probabilità di diametri ed età delle condotte facenti parte delle sotto-tabelle. Nel caso studio in esame, sono state scelte 3 classi di suddivisione per diametri ed età. Ne seguono 9 classi di condotte per ogni materiale, contraddistinte da diametro ed età equivalenti e dalla somma delle lunghezze e del numero di rotture di ogni condotta.

3. Una volta noti questi dati, tramite il software *Gretl* saranno formulate differenti forme polinomiali e si stimeranno i parametri del modello e i rispettivi coefficienti di determinazione. Ogni forma analizzata sarà contraddistinta da un certo numero di coefficienti polinomiali ed esponenti incogniti che, insieme al *CoD* si posizioneranno in un punto sulla *frontiera di Pareto*. L'analisi dell'*ottimo paretiano* porterà alla scelta dell'equazione ottimale per ogni materiale. Infine, si valuta la significatività dei singoli parametri, come visto nel **Capitolo 2**, nel caso della regressione logistica.

8.1 Suddivisione delle condotte secondo il materiale

Come accennato in precedenza, l'applicazione di tale modello prevede che le condotte vengano suddivise gruppi contraddistinti dallo stesso materiale. Attraverso il software *Python* è stato possibile ottenere le sotto-tabelle *EsattaAcciaio, EsattaEternit*, *EsattaGhisaGrigia* ed *EsattaGhisaSferoidale*, punto di partenza per l'applicazione del modello di regressione polinomiale. Sono così composte:

- la tabella *EsattaAcciaio* presenta 347 condotte di lunghezza complessiva 48.78 chilometri e 5 guasti;
- la tabella *EsattaEternit* presenta 34 condotte di lunghezza complessiva 1.72 chilometri e 10 guasti;
- la tabella *EsattaGhisaGrigia* presenta 305 condotte di lunghezza complessiva 27.75 chilometri e 59 guasti;
- la tabella *EsattaGhisaSferoidale* presenta 5074 condotte di lunghezza complessiva 224.69 chilometri e 32 guasti;

Si riporta in **Tabella 8.1** la lunghezza totale e i guasti caratterizzanti ogni materiale.

Tabella 8.1. Suddivisione della tabella *Esatta* secondo il materiale

Materiale	N° condotte	Lunghezza (km)	N° di guasti
Acciaio	347	48.78	5
Eternit	34	1.72	10
Ghisa grigia	305	27.75	59
Ghisa sferoidale	5074	224.69	32
TOT	5760	306.62	106

La tabella *Esatta* è caratterizzata in modo predominante dal materiale ghisa sferoidale: come già spiegato in precedenza, questo materiale è di recente utilizzo e dispone nella maggior parte dei casi del dato relativo all'anno di posa. Proprio questa informazione è stata il discriminante per l'esclusione di molti elementi, nel passaggio dalla tabella *ReteTorino* alla *Esatta*.

8.2 Suddivisione delle condotte in classi

Una volta ottenute le quattro tabelle, è stato necessario suddividere le condotte in classi di diametri ed età equivalenti. Tali grandezze equivalenti sono definite secondo le espressioni:

$$A_{Classe} = \frac{\sum_{Classe}(L_p \cdot A_p)}{L_{Classe}} \quad (1)$$

$$D_{Classe} = \frac{\sum_{classe}(L_p \cdot D_p)}{L_{Classe}} \quad (2)$$

dove:

- A_p e D_p rappresentano rispettivamente l'età e il diametro di ogni condotta facente parte di una determinata classe;
- L_{Classe} indica la lunghezza totale delle condotte facenti parte di una determinata classe.

Ogni classe contiene un determinato numero di condotte ed è contraddistinta da un'età equivalente A_{Classe} e da un diametro equivalente D_{Classe}, pesati rispetto alla lunghezza totale delle condotte di ogni classe. L_p, invece, è la lunghezza di ogni condotta facente parte di una determinata classe. Pesare i dati rispetto alla lunghezza ha una rilevanza statistica poiché questa indicazione include informazioni non sempre disponibili e correlate alla stessa quali: variabilità dei carichi stradali, variabilità dei valori di carico idraulici e carichi del terreno.
Affinché le classi abbiano lo stesso peso statistico, è necessario che queste contengano un numero comparabile di elementi. Per questo motivo, prima di applicare le **Equazioni 1** e **2** sono state esaminate la densità di probabilità e la funzione di probabilità cumulata dei diametri e delle età di ogni materiale.
Si richiamano passi esposti nel **Capitolo 3**-**Esempio 4** per la definizione delle due funzioni sopra citate.

1. Si calcolano le occorrenze per tutti i valori di diametri ed età.
2. Per ogni valore si calcola la funzione densità di probabilità come il rapporto tra il numero di occorrenze di un valore di diametro/età e le occorrenze totali di tutti i valori di diametri/età. A titolo di esempio, se ad un determinato diametro corrispondono 204 occorrenze e il numero totale di elementi è 2775, la corrispettiva funzione densità di probabilità *f(x)* sarà (204·100)/2775=7.46%.
3. Per ogni valore, si calcola la funzione probabilità cumulata *F(x)* come la sommatoria progressiva delle densità di probabilità *f(x)* di tutti gli elementi precedenti a quello in esame.

4. Il dominio della funzione probabilità cumulata varia tra 0 e 1. Quindi, scelto il numero di classi, si suddivide la *F(x)* in 3 classi, cercando di ottenere degli intervalli di larghezza pari al 33.33%. Si ricavano i corrispondenti intervalli di variazione di diametri ed età mediante i quali suddividere i dati del campione.

Per chiarire al meglio i passaggi effettuati, si faccia riferimento nuovamente all'**Esempio 1** del **Capitolo 3**, di cui si riportano i dati in **Tabella 8.2**.

Esempio 1.
Si supponga di dover esaminare una rete composta da 9 condotte, contraddistinte da un codice identificatore, numero di rotture, età, lunghezza e diametro (**Tabella 8.1)**.

Tabella 8.2. Composizione della rete esempio

idPipe	NR	A_p (anni)	L_p (m)	D_p (mm)
1	1	30	10	63
2	3	30	55	63
3	0	30	35	63
4	0	40	10	75
5	5	40	55	75
6	2	40	5	90
7	2	25	10	100
8	3	25	35	100
9	4	25	15	100

È necessario dapprima calcolare le occorrenze dell'*i-esimo* valore di diametro ed età e, successivamente, calcolare le funzioni densità di probabilità $f_i(x)$ e probabilità cumulata $F_i(x)$ secondo le equazioni:

$$f(x_i) = \frac{N°occorrenze_i}{N°\ occorrenze\ totali} \tag{3}$$

$$F(\mathrm{x}_i) = \sum_{j=1}^{i} f(x_i) \tag{4}$$

Dove il numero di occorrenze si riferisce all'occorrenza del valore *i-esimo* di ogni diametro/età e il numero di occorrenze totali corrisponde al numero di elementi totali.
Nell'esempio in esame, il diametro da 63 millimetri caratterizza 3 condotte, mentre la totalità delle occorrenze è pari a 9. Di conseguenza, la funzione densità di probabilità espressa in termini percentuali vale:

$$f(63) = \frac{3}{9} \cdot 100 = 33.33\%$$

Allo stesso modo, la stessa funzione per i diametri da 75, 90 e 1000 millimetri vale:

$$f(75) = \frac{2}{9} \cdot 100 = 22.22\%$$

$$f(90) = \frac{1}{9} \cdot 100 = 11.11\%$$

$$f(100) = \frac{3}{9} \cdot 100 = 33.33\%$$

Le corrispondenti funzioni di probabilità cumulata sono date dalla sommatoria progressiva delle funzioni densità di probabilità. Ad esempio, per i diametri da 90 e 100 millimetri, valgono:

$$F(90) = f(63) + f(75) + f(90) = 77.77\%$$

Una volta applicate le **Equazioni 3** e **4** a tutti i valori di diametro ed età, si ottengono le T**abelle 8.3-4**.

Tabella 8.3. Funzioni densità di probabilità e probabilità cumulata dei diametri

D_p (mm)	N° occorrenze	f(x) (%)	F(x) (%)
63	3	33.33	33.33
75	2	22.22	55.55
90	1	11.11	66.66
100	3	33.33	100

Tabella 8.4. Funzioni densità di probabilità e probabilità cumulata dei diametri

A_p (anni)	N° occorrenze	f(x) (%)	F(x) (%)
25	3	33.33	33.33
30	3	33.33	66.66
40	3	33.33	100

Si riportano graficamente le funzioni densità di probabilità e probabilità cumulata per i diametri e le età.

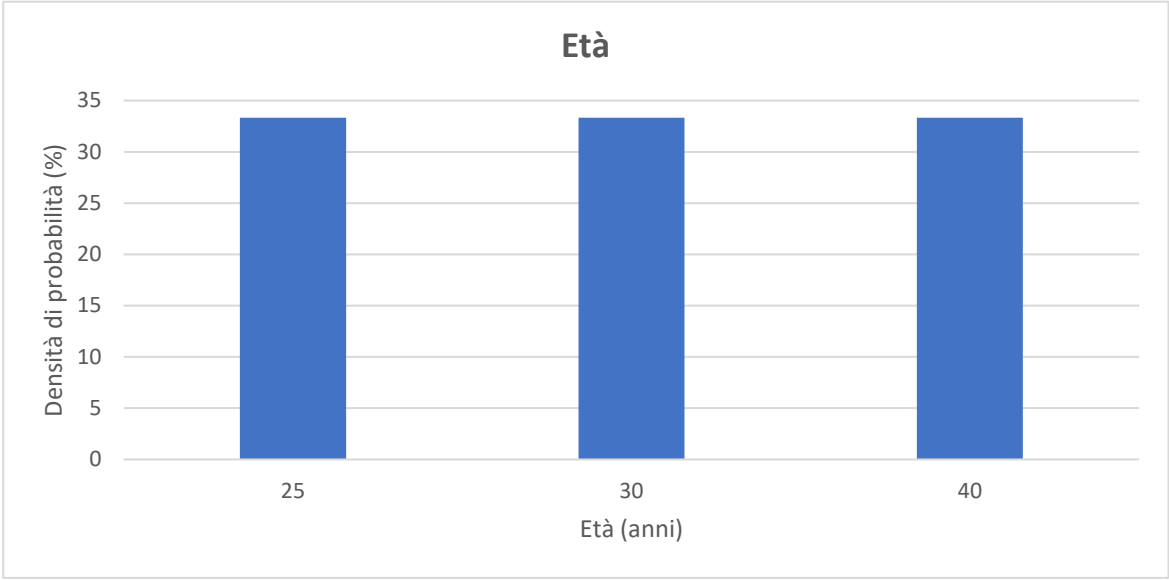


Figura 8.1. Funzione densità di probabilità dei diametri

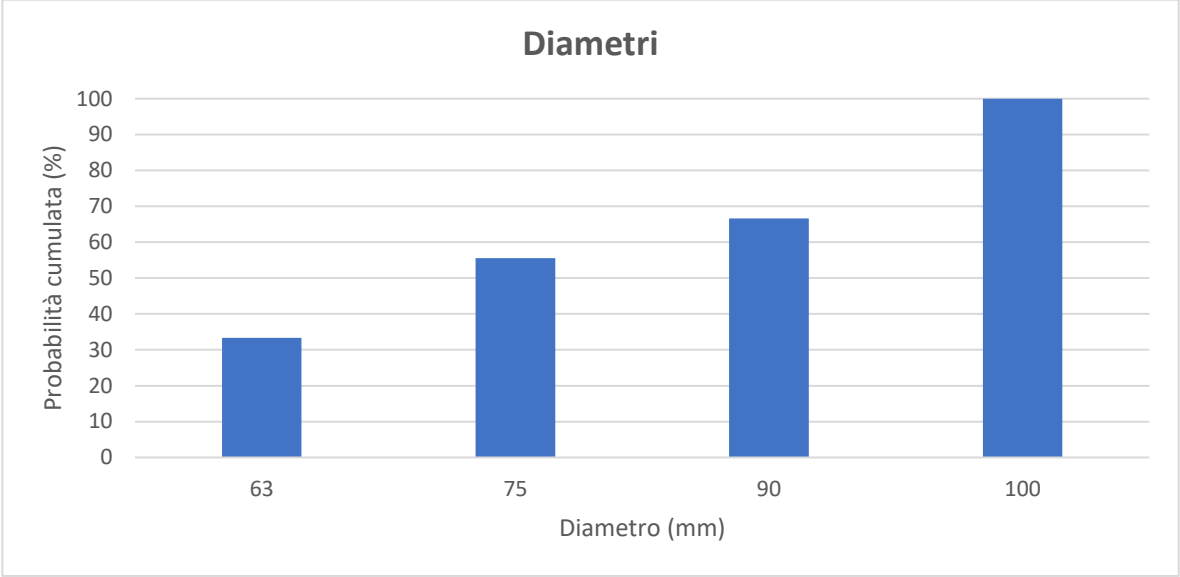


Figura 8.2. Funzione probabilità cumulata dei diametri

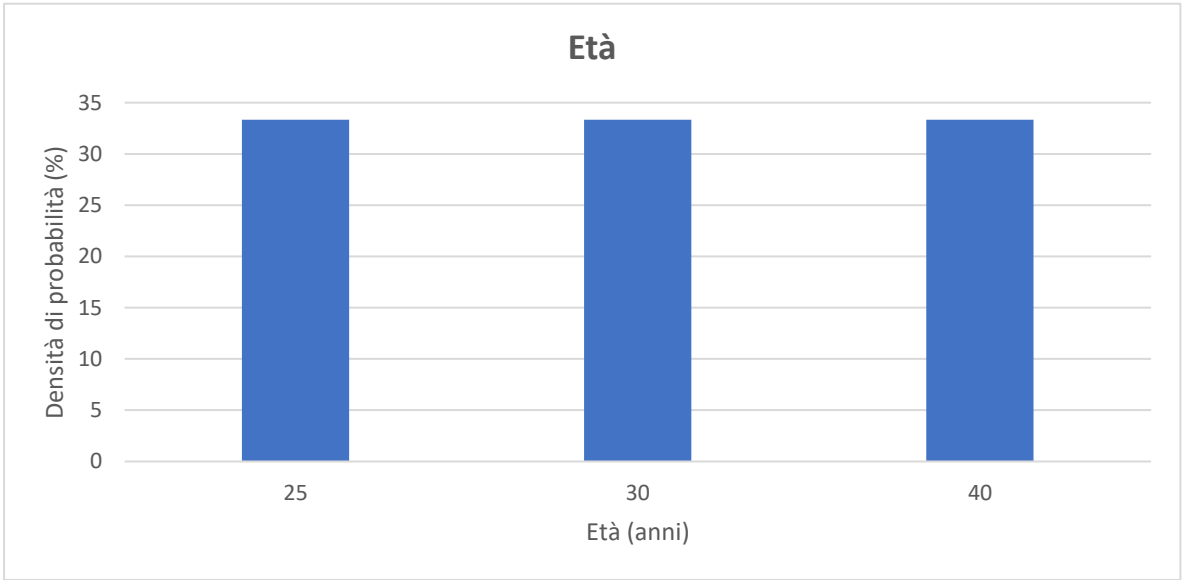


Figura 8.3. Funzione densità di probabilità dell'età

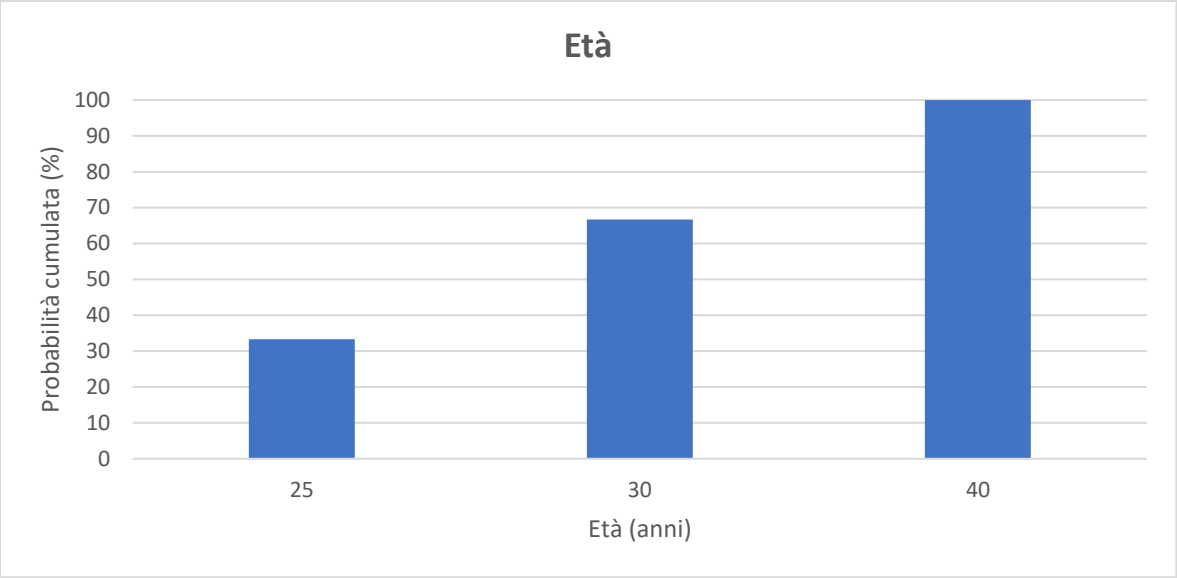


Figura 8.4. Funzione probabilità cumulata dell'età

Raggruppando le condotte comprese tra 0 e 63 millimetri, 64 e 90 millimetri e quelle superiori ai 90 millimetri si otterranno 3 classi composte dallo stesso numero di elementi. Le stesse considerazioni valgono per l'età delle condotte: infatti, raggruppando le condotte caratterizzate da età compresa tra 0 e 25 anni, tra 26 e 30 anni e quelle caratterizzate da età superiore ai 30, sarà possibile ottenere classi composte dallo stesso numero di elementi.
Si ottengono le seguenti classi di diametri:

- classe 1: comprende gli elementi caratterizzati da diametri compresi tra 0 e 63 millimetri;
- classe 2: comprende gli elementi caratterizzati da diametri compresi tra 64 e 90 millimetri;
- classe 3: comprende gli elementi caratterizzati da diametri superiori al 90 millimetri.

Si ottengono le seguenti classi di età:

- classe 1: comprende gli elementi caratterizzati da età compresa tra 0 e25 anni;
- classe 2: comprende gli elementi caratterizzati da età compresa tra 26 e 30 anni;
- classe 3: comprende gli elementi caratterizzati da età superiore ai 30 anni.

L'intersezione tra diametri ed età restituisce 9 gruppi di condotte, i cui range di variazione sono riportati in **Tabella 8.5**. La classe 1-1 fa riferimento alla classe 1 dei diametri e alla classe 1 delle età, cioè agli elementi caratterizzati da diametri compresi tra 0 e 63 millimetri ed età tra 0 e 25 anni. Tale notazione è valida per le restanti 8 classi.
È necessario, quindi, applicare le **Equazioni 1** e **2** ad ogni gruppo ottenuto, per calcolare il diametro e l'età rappresentativa. Per quanto riguarda la lunghezza della classe e il numero di rotture, queste informazioni sono calcolate come la somma dei singoli dati che caratterizzano ogni condotta della classe.

Si ottengono i risultati riportati in **Tabella 8.5**: le 9 classi di condotte con le rispettive informazioni saranno la base sulla quale stimare i diversi modelli di regressione polinomiale. È importante notare che solo 3 classi su 9 presentano elementi non nulli, a causa della non omogenea distribuzione di diametri ed età e la conseguente mancanza di condotte ricadenti in un determinato e contemporaneo range di diametri ed età.

Tabella 8.5. Range di raggruppamento delle 9 classi

Classe	D_p (mm)	A_p (anni)
1-1	0-63	0-25
1-2	0-63	26-30
1-3	0-63	31-40
2-1	64-90	0-25
2-2	64-90	26-30
2-3	64-90	31-40
3-1	91-100	0-25
3-2	91-100	26-30
3-3	91-100	31-40

Tabella 8.6. Classi di condotte e relative informazioni

Classe	D_p (mm)	A_p (anni)	L_p (m)	NR
1-1	0	0	0	0
1-2	63	30	100	4
1-3	0	0	0	0
2-1	0	0	0	0
2-2	0	0	0	0
2-3	76	40	70	7
3-1	100	25	100	9
3-2	0	0	0	0
3-3	0	0	0	0

Si riportano di seguito le funzioni densità di probabilità (*f(x)*) e probabilità cumulata (*F(x))* per i diametri e le età di tutti i materiali.

Acciaio

In **Figura 8.5** sono riportati i diagrammi relativi alle funzioni densità di probabilità e probabilità cumulata dei diametri.

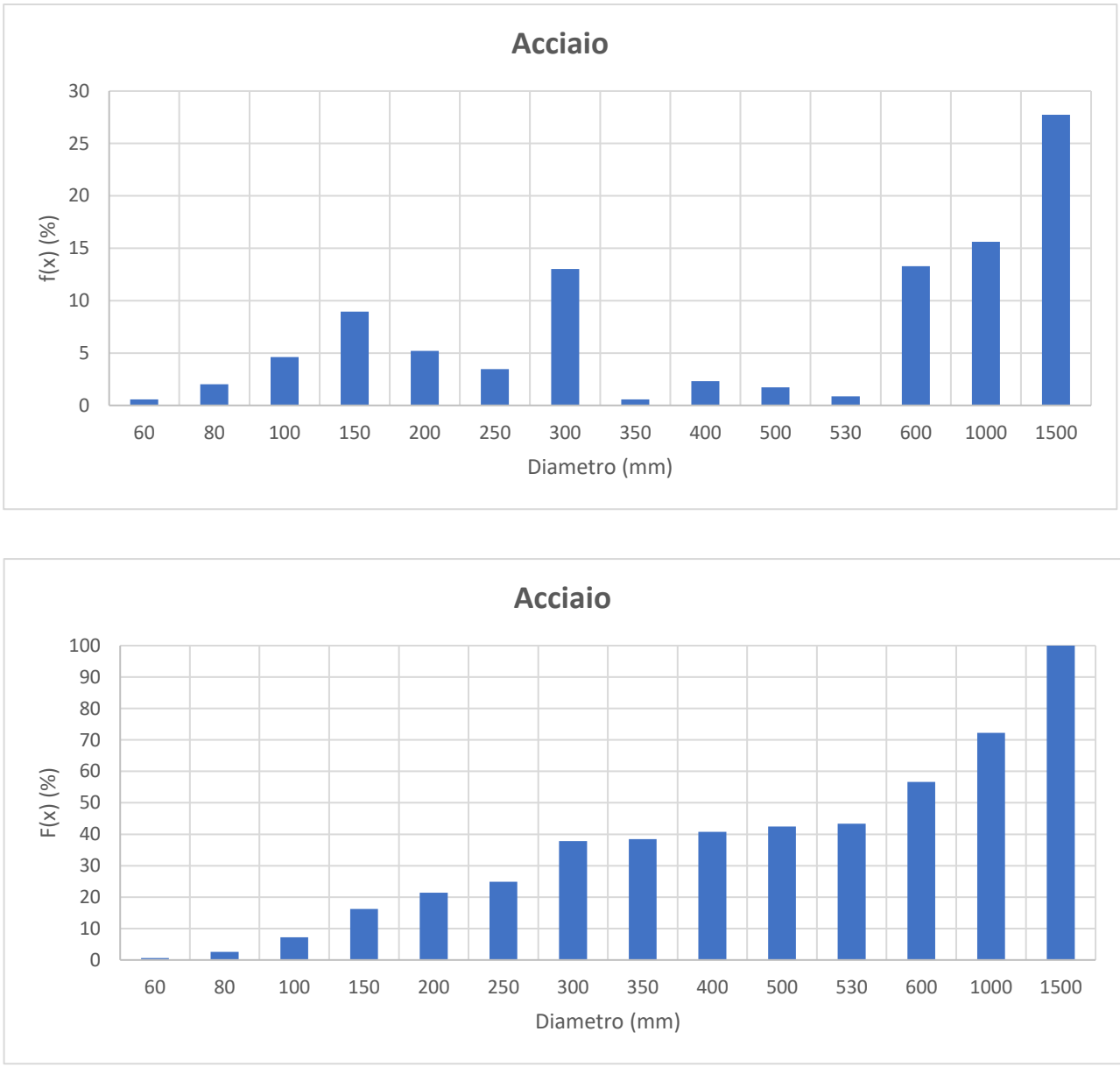


Figura 8.5. Funzioni densità di probabilità e probabilità cumulata per i diametri in acciaio

Le occorrenze maggiori si trovano in corrispondenza dei diametri da 1500 millimetri.
Si riportano di seguito le stesse grandezze calcolate per l'età delle condotte: le occorrenze più numerose, in questo caso, si trovano in corrispondenza dei 63 anni.

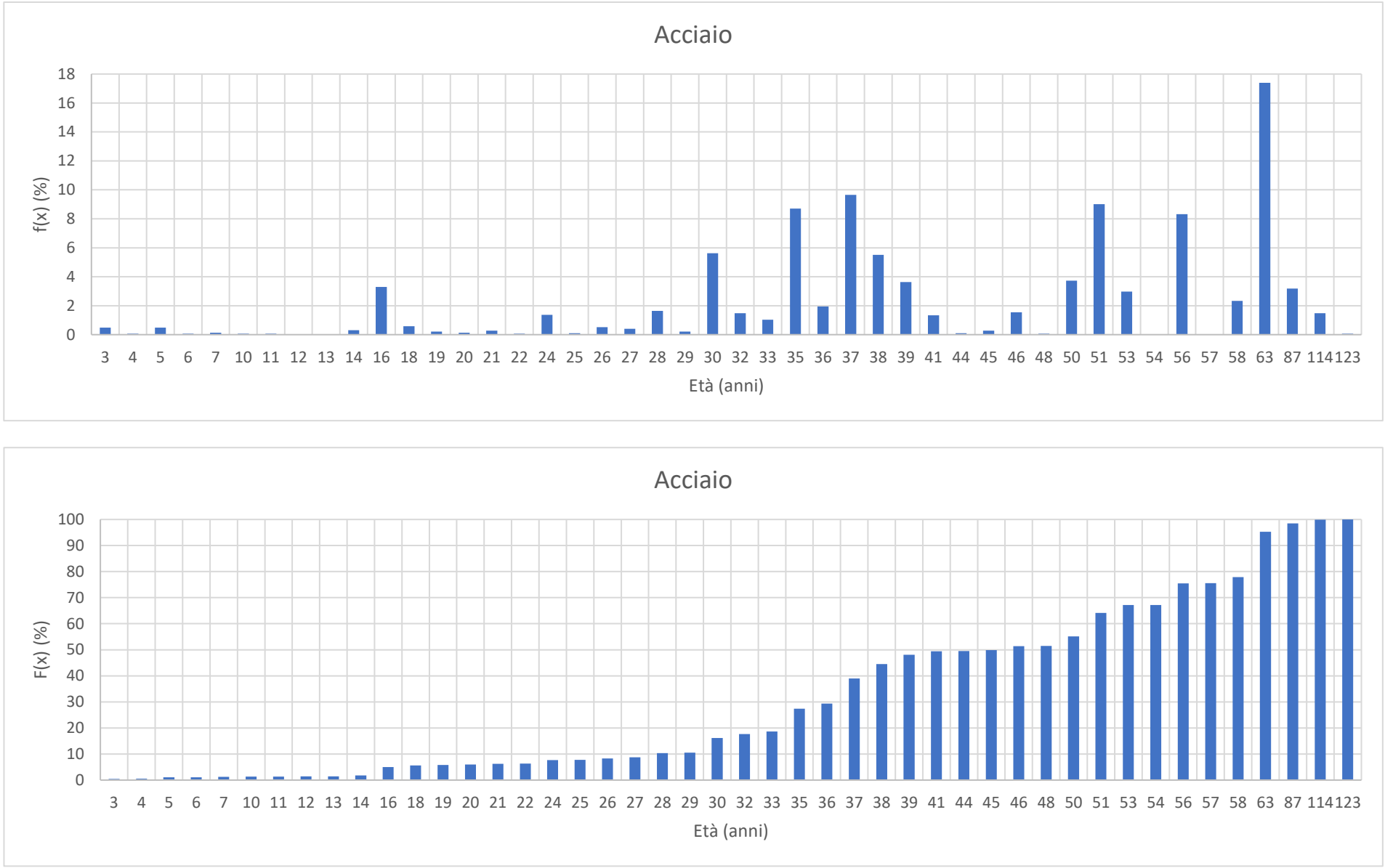
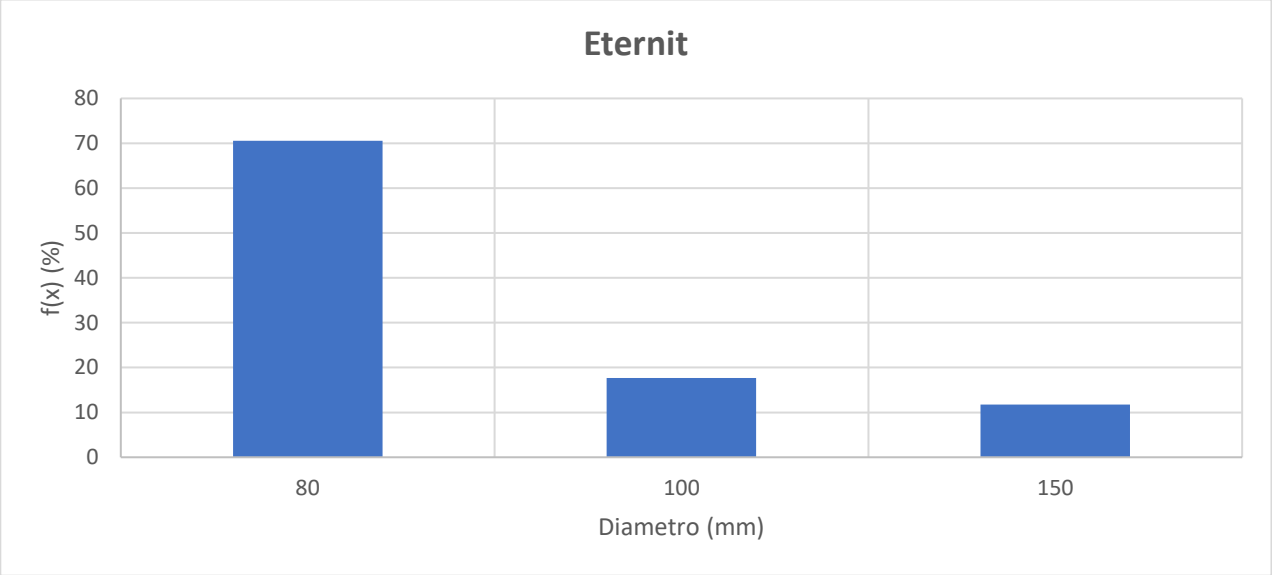


Figura 8.6. Funzioni densità di probabilità e probabilità cumulata per l'età delle condotte in acciaio

Eternit

In **Figura 8.7** sono riportati i diagrammi relativi alle funzioni densità di probabilità e probabilità cumulata dei diametri.



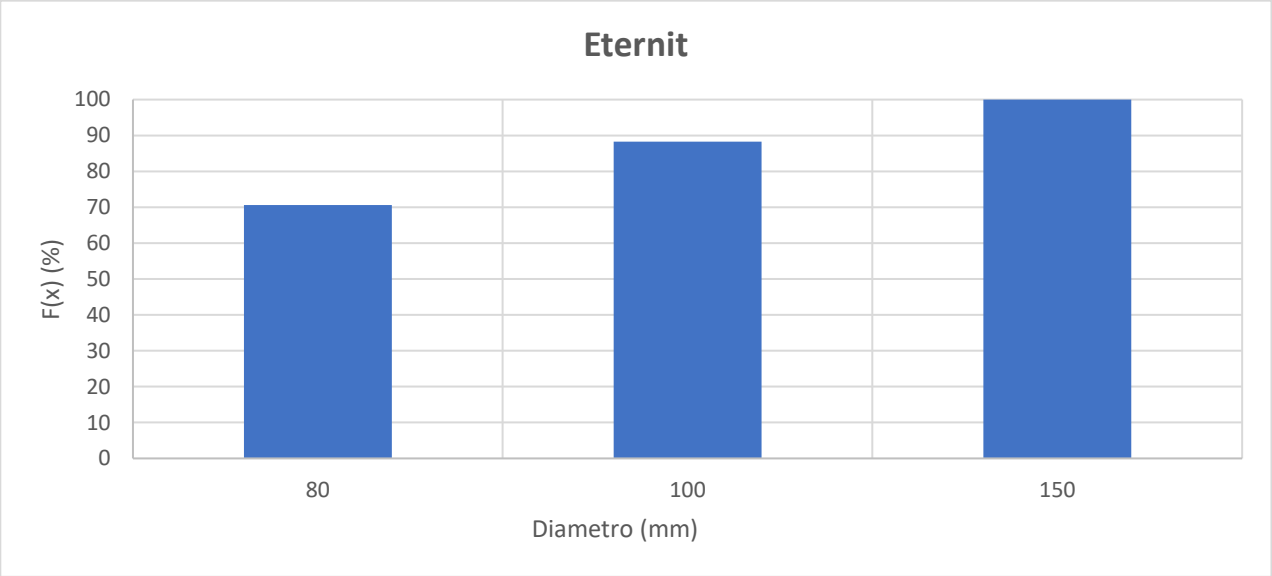
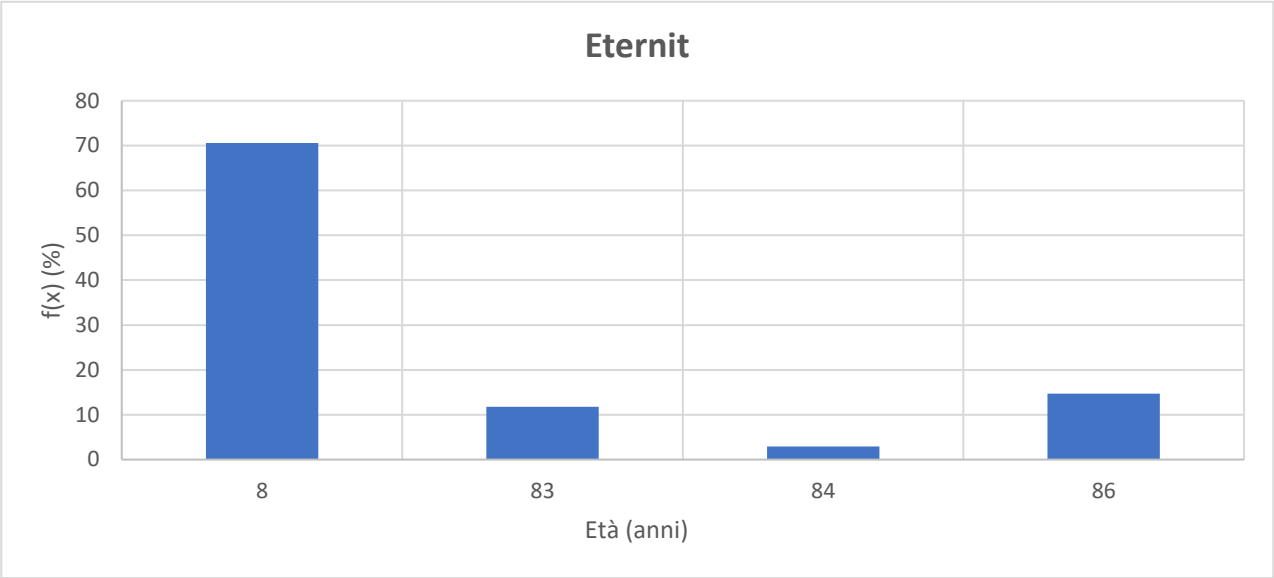


Figura 8.7. Funzioni densità di probabilità e probabilità cumulata per i diametri in eternit

Le occorrenze maggiori si trovano in corrispondenza dei diametri da 80 millimetri.
Si riportano di seguito le stesse grandezze calcolate per l'età delle condotte. Le occorrenze più numerose, in questo caso, si trovano in corrispondenza di 8 anni.



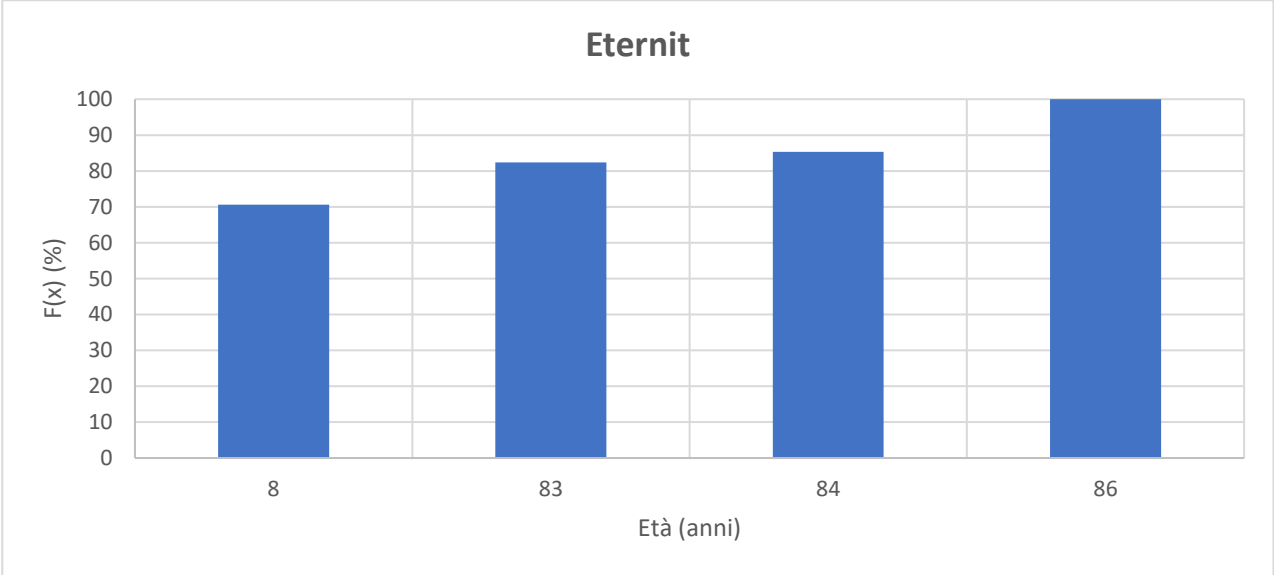
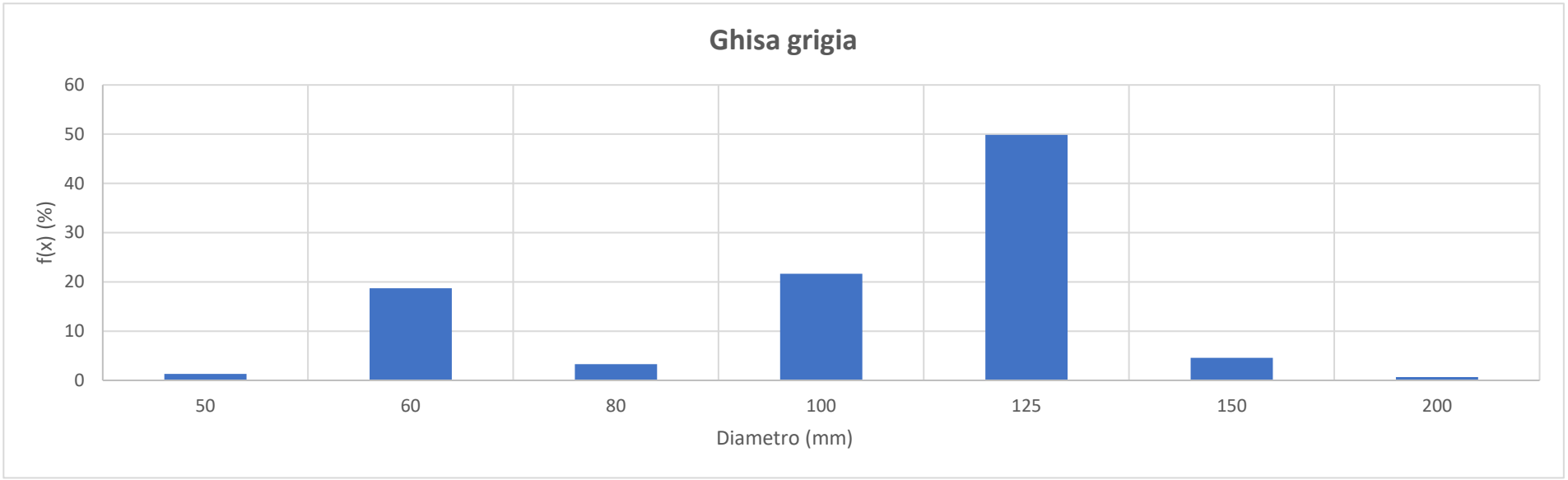


Figura 8.8. Funzioni densità di probabilità e probabilità cumulata per l'età delle condotte in eternit

Ghisa grigia

In **Figura 8.9** sono riportati i diagrammi relativi alle funzioni densità di probabilità e probabilità cumulata dei diametri.
I diametri maggiori si trovano in corrispondenza dei 125 millimetri, mentre l'età più ricorrente è quella relativa ai 91 anni.



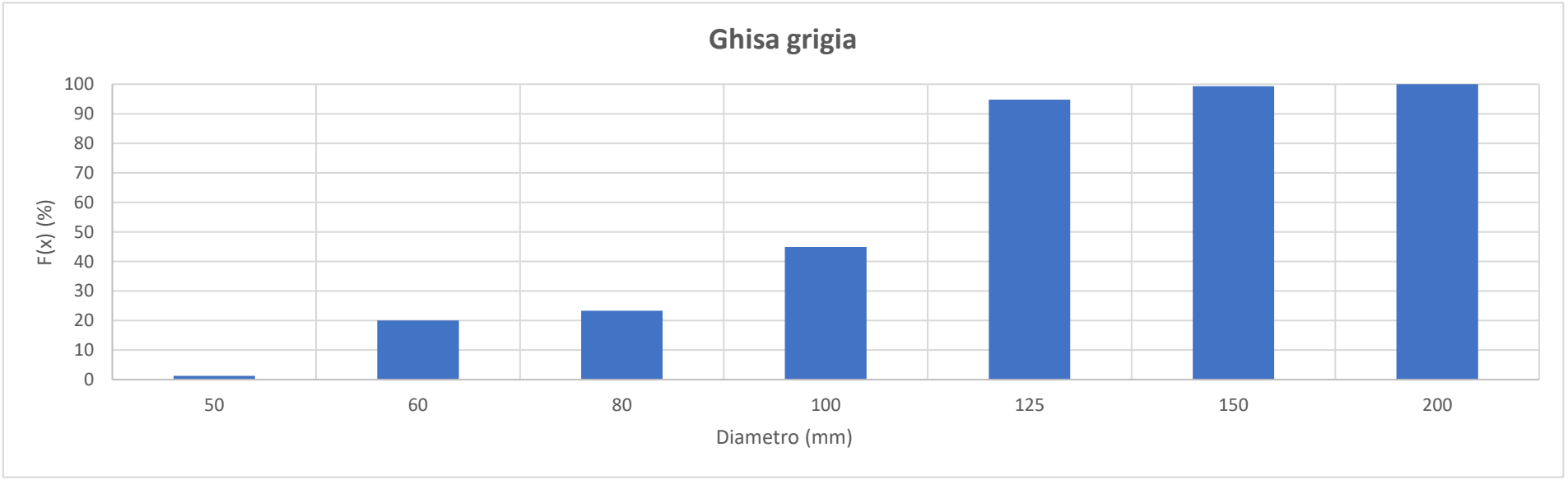
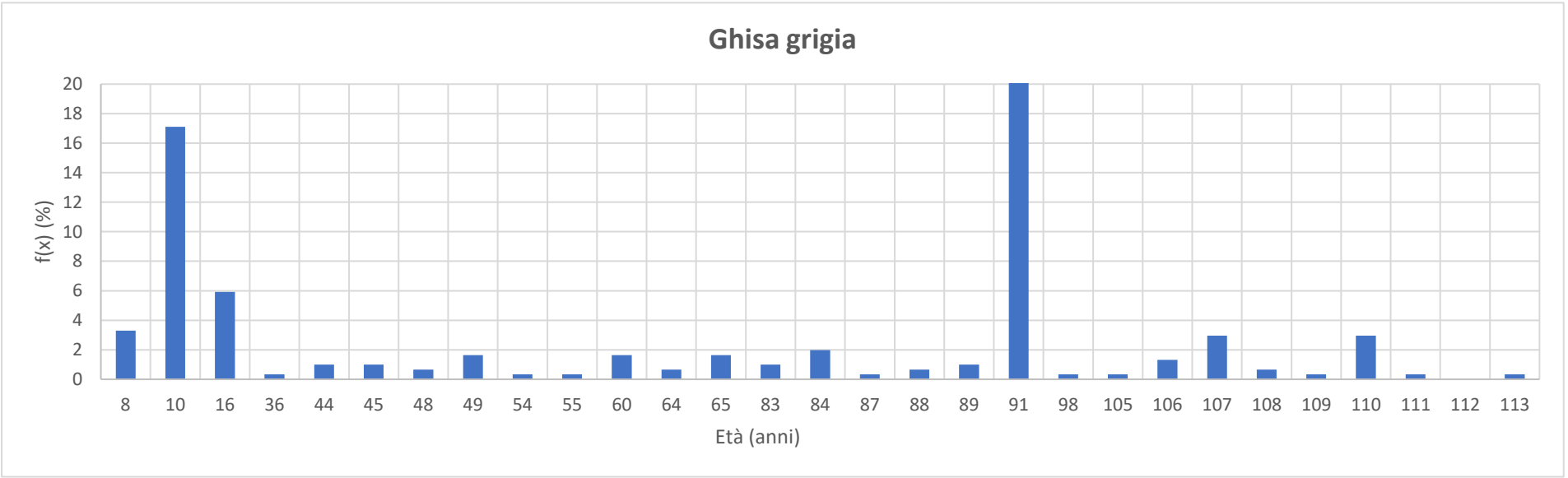


Figura 8.9. Funzioni densità di probabilità e probabilità cumulata per i diametri in ghisa grigia



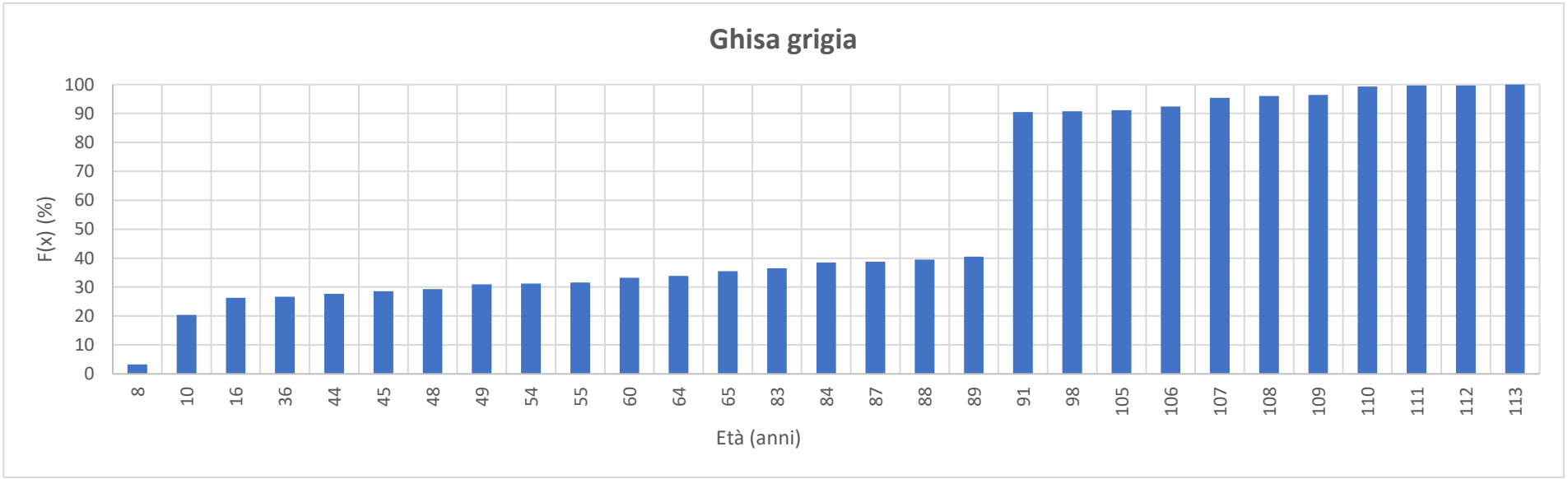
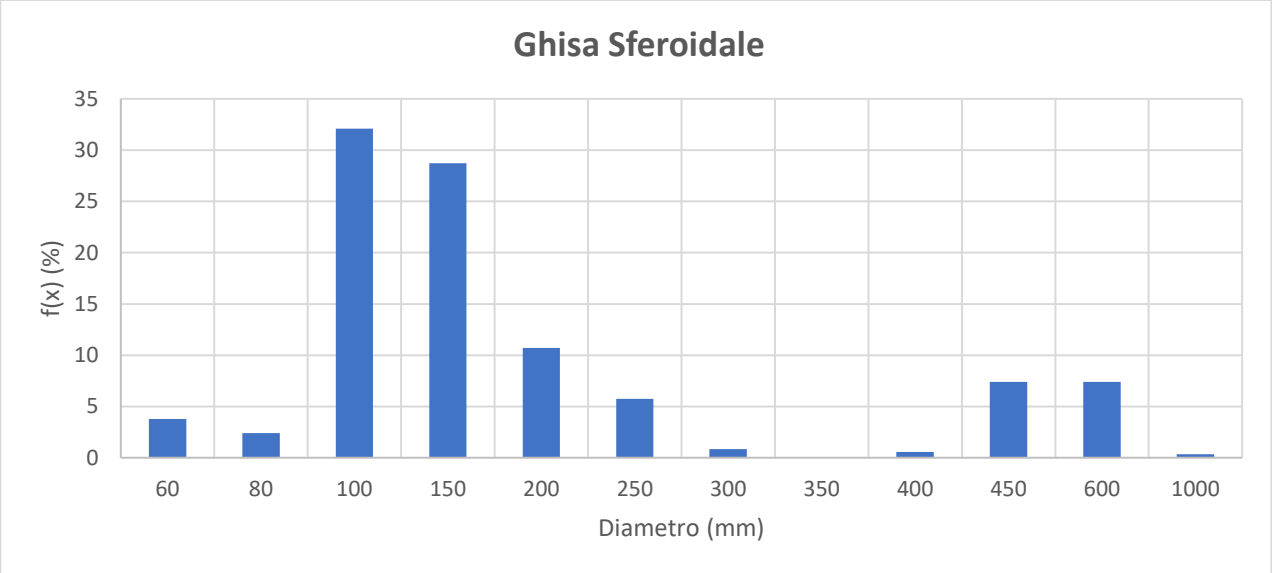


Figura 8.10. Funzioni densità di probabilità e probabilità cumulata per l'età delle condotte in ghisa grigia

Ghisa sferoidale

In **Figura 8.11** sono riportati i diagrammi relativi alle funzioni densità di probabilità e probabilità cumulata dei diametri.
I diametri maggiori si trovano in corrispondenza dei 100 e 150 millimetri.



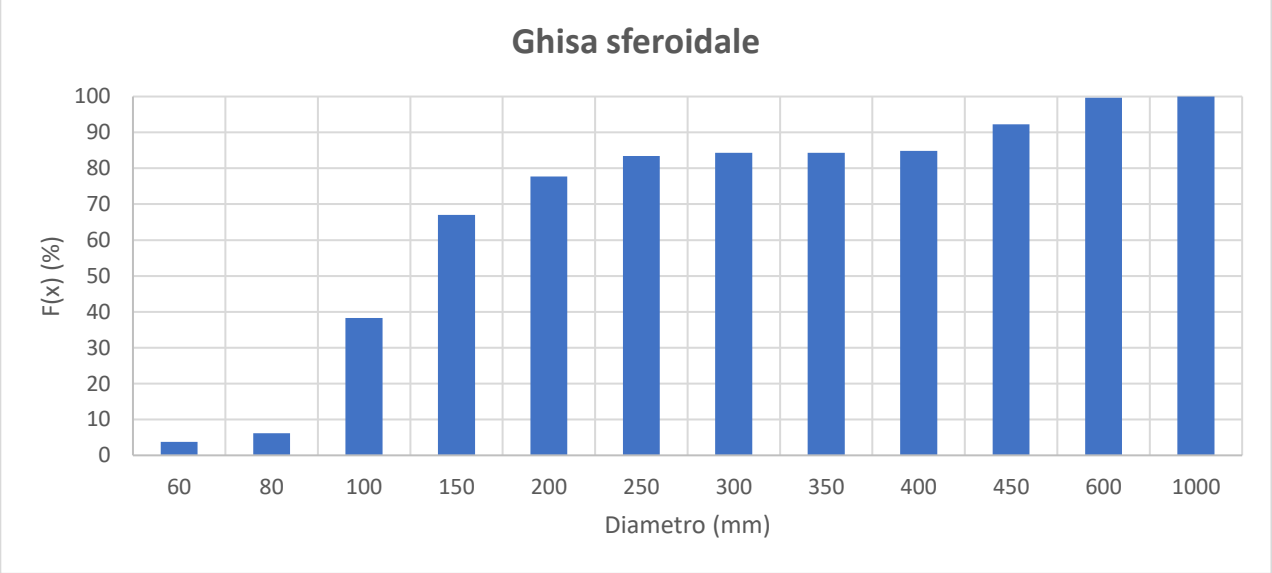


Figura 8.11. Funzioni densità di probabilità e probabilità cumulata per i diametri in ghisa sferoidale

Si riportano di seguito le stesse grandezze calcolate per l'età delle condotte. Le occorrenze maggiori di età si trovano in corrispondenza dei 18 anni, a riprova che la ghisa sferoidale sia un materiale di recente utilizzo.

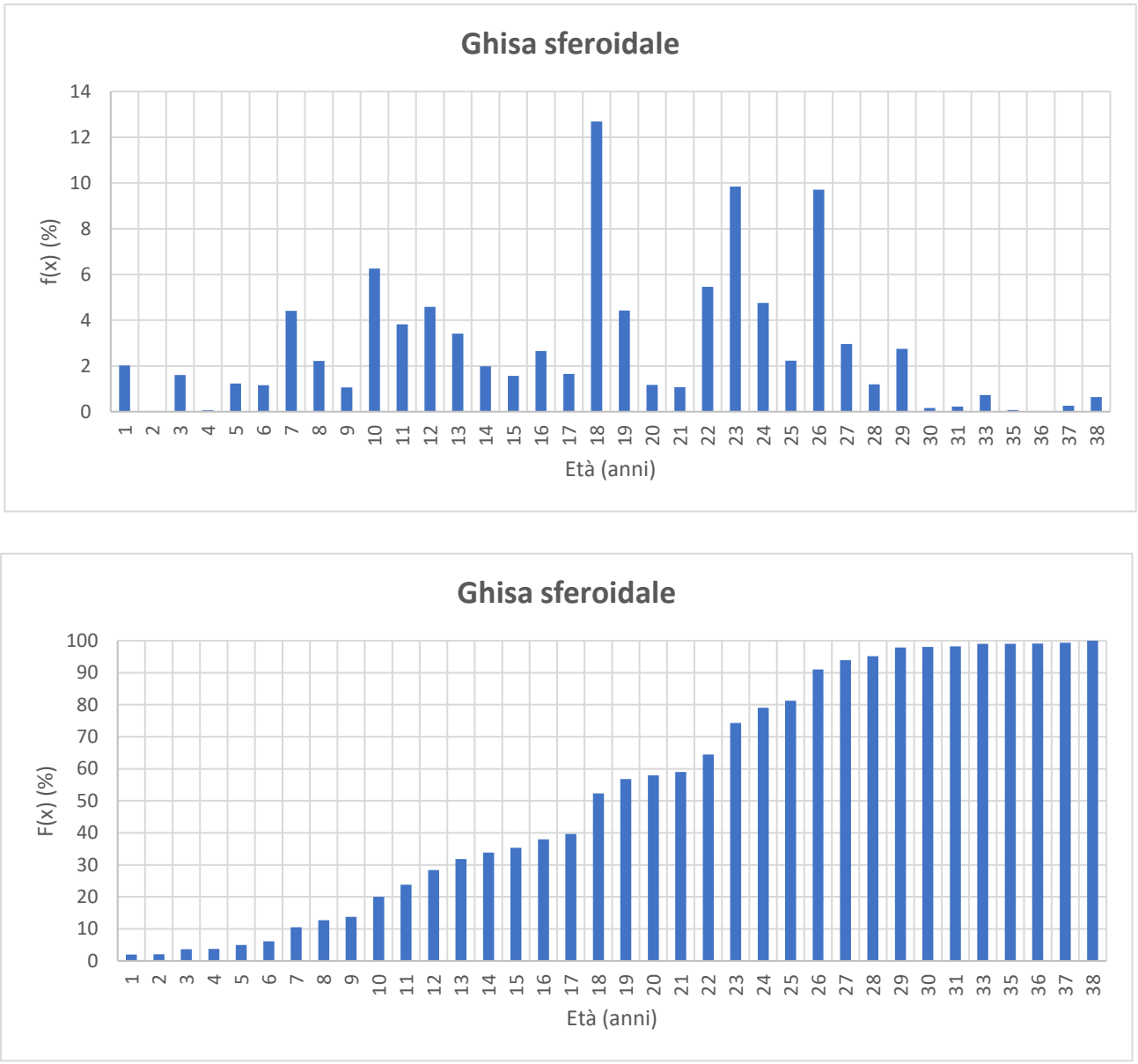


Figura 8.12. Funzioni densità di probabilità e probabilità cumulata per l'età delle condotte in ghisa grigia

Una volta esaminate le distribuzioni di diametri ed età e fissato il numero di classi, le corrispondenti funzioni di probabilità cumulata sono state suddivise in tre intervalli. Affinché le classi abbiamo al loro interno un numero di elementi comparabili, tali intervalli devono essere caratterizzati da una larghezza prossima al 33%.

Si ricavano le classi riportate qui di seguito che fanno riferimento ai materiali presi singolarmente. È possibile notare che tutti i materiali ad eccezione della ghisa sferoidale presentano classi vuote e questo è dovuto ad una limitata grandezza del campione di dati ed alla non uniforme distribuzione dei valori di diametri ed età. Inoltre, è importate sottolineare che un maggior numero di classi non vuote porta a stime più affidabili dei parametri del modello.

I dati riportati saranno le informazioni sulle quali i differenti modelli di regressione polinomiale saranno tarati.

Tabella 8.7. Classi ricavate per l'acciaio

Acciaio					
Classe		**D_{classe} (mm)**	**$Età_{classe}$ (anni)**	**L_{tot} (m)**	**NR**
1	1	215.54	20.53	5990	3
	2	92	50.73	110	2
	3	0	0	0	0
2	1	961.34	23.68	2690	0
	2	0	0	0	0
	3	0	0	0	0
3	1	1500	16	8030	0
	2	0	0	0	0
	3	0	0	0	0

Tabella 8.8. Classi ricavate per l'eternit

Eternit					
Classe		**D_{classe} (mm)**	**$Età_{classe}$ (anni)**	**L_{tot} (m)**	**NR**
1	1	80	8	860	0
	2	0	0	0	0
	3	0	0	0	0
2	1	0	0	0	0
	2	0	0	0	0
	3	100	85.76	330	6
3	1	0	0	0	0
	2	150	83	610	4
	3	0	0	0	0

Tabella 8.9. Classi ricavate per la ghisa grigia

Ghisa Grigia					
Classe		**D_{classe} (mm)**	**$Età_{classe}$ (anni)**	**L_{tot} (m)**	**NR**
1	1	77.36	18.60	4690	14
	2	87.43	80.49	1360	20
	3	110.31	107.78	1450	16
2	1	0	0	0	0
	2	0	0	0	0
	3	90.55	65.92	6640	0
3	1	150	11.73	2800	1
	2	0	0	0	0
	3	170.41	110.27	490	8

Tabella 8.10. Classi ricavate per la ghisa sferoidale

Ghisa sferoidale					
Classe		**D_{classe} (mm)**	**$Età_{classe}$ (anni)**	**L_{tot} (m)**	**NR**
1	1	92.47	9.11	47170	8
	2	95.84	19.03	36370	8
	3	93.55	25.66	38290	2
2	1	150	8.37	20450	3
	2	150	18.64	25130	3
	3	150	25.14	47170	3
3	1	318.77	10.44	23100	3
	2	499.43	18.3	27130	1
	3	244.57	26.26	24860	1

8.3 Stima dei parametri del modello

Una volta calcolate le 9 classi di elementi per ogni materiale, è necessario formulare differenti modelli polinomiali, in modo tale da ricercare la forma che meglio si adatta al campione di dati. Ogni modello sarà contraddistinto da una differente complessità di espressione, quantificabile attraverso il numero di parametri X_h, di esponenti incogniti e di monomi A_j che lo compongono.
Il modello polinomiale adottato per definire la dipendenza tra il numero di rotture per una determinata classe di condotte e le caratteristiche delle stesse è contraddistinto dalla seguente formulazione:

$$Y = a_0 + \sum_{j=1}^{m} a_j X_1^{ES(j,1)} \cdot\ldots\cdot (X_k)^{ES(j,k)} \cdot f\left(X_1^{ES(j,k+1)} \cdot\ldots\cdot (X_k)^{ES(j,2k)}\right) \qquad (5)$$

dove:

- *Y* è la variabile indipendente, denominata anche *NR,* ed indica il numero di rotture;
- X_k è la generica variabile esplicativa (ad esempio diametro, materiale, anno di posa e lunghezza);
- *ES* è la matrice degli esponenti incogniti;
- a_j è il generico coefficiente polinomiale da stimare;
- *m* è il numero di termini polinomiali addizionali alla costante a_0;
- *f* rappresenta una funzione arbitraria che può incrementare la capacità di previsione delle rotture.

I parametri del modello appena definiti sono calcolati a partire da una popolazione campionaria di dimensioni finite e la loro stima avviene attraverso il software statistico *Gretl.* Nello specifico, il numero di coefficienti polinomiali e di esponenti incogniti è imposto e variato arbitrariamente dall'utente, affinché il modello si adatti nel miglior modo possibile al campione.

Inoltre, come ampiamente descritto nel **Capitolo 3**, ogni forma polinomiale sarà caratterizzata da una determinata bontà di adattamento, misurata attraverso il *coefficiente di determinazione* (*CoD*). Infine, ogni modello rappresenterà un punto sulle *Frontiere di Pareto* nei grafici *1-CoD-X_h* e *1-Cod-A_j* e sarà scelta la forma che, in maniera ottimale, potrà descrivere il numero di rotture di ogni classe di materiale.
Questa procedura è stata effettuata per i quattro materiali presi in esame, dei quali si riportano i risultati qui di seguito.

Acciaio

Si riportano in **Tabella 8.11** le forme polinomiali stimate a partire dalle classi riportate in **Tabella 8.7**.

Tabella 8.11. Modelli formulati nel caso dell'acciaio

Acciaio				
Formula	CoD	No. X_h	No.A_j	1-CoD
$NR = -0.002 \cdot D$ +0.0005L	0.42	2	2	0.58
$NR = 5.54E-05 \cdot L$+0.03811 $\cdot A$	0.44	2	2	0.56
$NR = 0.00041 \cdot L - 0.00258 \cdot D + 0.04567 \cdot A$	0.99	3	2	0.01

Nel caso dell'acciaio non è stato possibile formulare numerose forme polinomiali complesse, poiché solo 4 classi su 9 risultano non vuote.
I modelli indicano che il numero di rotture aumenta al diminuire del diametro e all'aumentare della lunghezza e dell'età della condotta.
Tra i modelli formulati, il terzo è caratterizzato da un coefficiente di determinazione pari a 0.99 e i parametri stimati risultano tutti significativi, come è possibile vedere in **Figura 8.13**. Inoltre, si riporta in **Figura 8.14** la sola *frontiera di Pareto* relativa alle coppie di punti *1-CoD-X_h*.
Questo risulta il modello ottimale, contraddistinto da una limitata complessità di espressione (3 variabili esplicative e 2 polinomi) e una buona capacità di adattamento ai dati. In aggiunta, tutti i parametri sono significativi, come attestato dai valori di *p-value*, sensibilmente inferiori alla significatività scelta, pari a 0.05.

```
Risultati di gretl per giova 2018-07-06 09:39, pagina 1

Sono state usate derivate numeriche
Tolleranza = 1.81899e-012
Convergenza raggiunta dopo 13 iterazioni

Modello 4: NLS, usando le osservazioni 1-9
NR = alpha*L+beta*D+gamma*Anni

                stima        errore std.    rapporto t    p-value
  ------------------------------------------------------------------
  alpha       0.000414829    2.97635e-05      13.94      8.50e-06 ***
  beta       -0.00257952     0.000172127     -14.99      5.56e-06 ***
  gamma       0.0456702      0.00300471       15.20      5.12e-06 ***

Media var. dipendente    0.555556   SQM var. dipendente      1.130388
Somma quadr. residui     0.150337   E.S. della regressione   0.158291
R-quadro non centrato    0.985293   R-quadro centrato       -0.007193
Log-verosimiglianza      5.644011   Criterio di Akaike      -5.288022
Criterio di Schwarz     -4.696348   Hannan-Quinn            -6.564852
Note: SQM = scarto quadratico medio; E.S. = errore standard
```

Figura 8.13. Output di *Gretl* relativo al terzo modello formulato

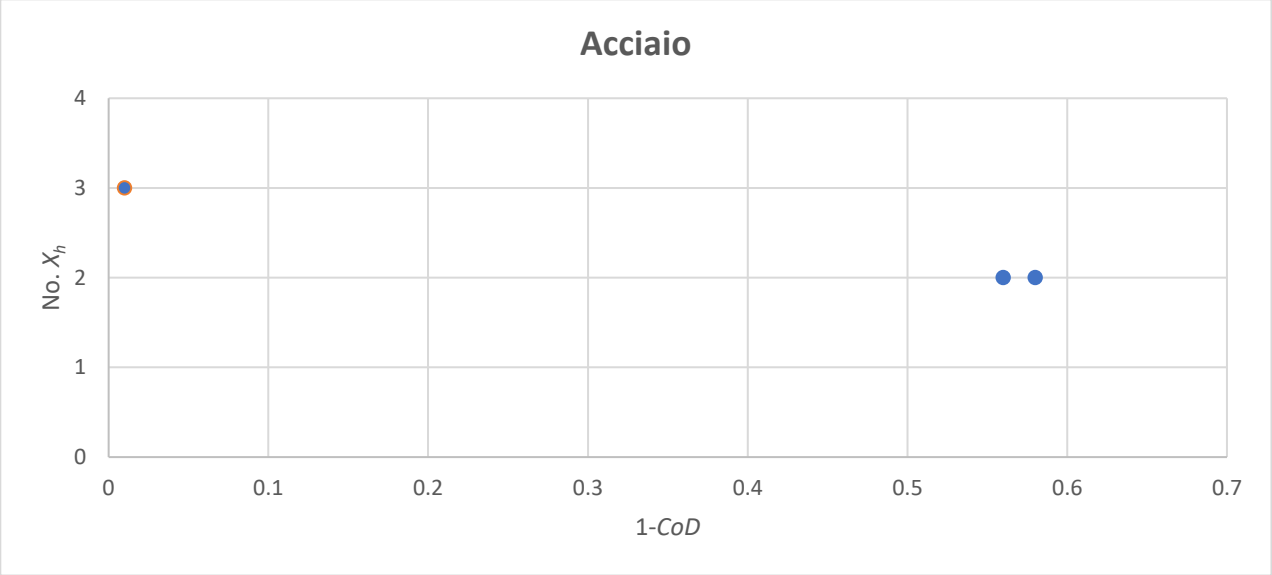


Figura 8.14. *Frontiera di Pareto* relativa alle coppie di punti *1-CoD-X_h*

Eternit

Si riportano in **Tabella 8.12** le forme polinomiali stimate a partire dalle classi riportate in **Tabella 8.8**.

Tabella 8.12. Modelli formulati nel caso dell'eternit

Formula	CoD	No. X_h	No.A_j	1-CoD
$NR = 0.0234442 \cdot D^{1.05713}$	0.6338	1	1	0.3662
$NR = 0.000471943 \cdot D \cdot A - 0.0005337 \cdot (A + L)$	0.83	4	2	0.17
$NR = 0.0653492 \cdot D - 0.00624323 \cdot (A + L)$	0.84	3	2	0.16
$NR = -0.0134497 \cdot D + 0.0795345 \cdot A$	0.98	2	1	0.02

Nel caso dell'eternit solo 3 classi su 9 sono risultate non vuote ed è stato possibile formulare quattro differenti modelli polinomiali. I modelli indicano che il numero di rotture aumenta all'aumentare del diametro (per i primi 3 modelli) e mostra una tendenza opposta nell'ultimo modello. Il numero di rotture diminuisce all'aumentare dell'età e della lunghezza delle condotte, ma mostra un andamento opposto nell'ultimo modello stimato.
Tra i modelli formulati, il quarto è caratterizzato da un coefficiente di determinazione pari a 0.98 e i parametri stimati risultano tutti significativi, come è possibile vedere in **Figura 8.15**. Inoltre, questa formulazione è in accordo con l'analisi del tasso di fallanza in relazione ai diametri, come riscontrato nel **Capitolo 6**.

```
Risultati di gretl per giova 2018-07-06 09:51, pagina 1

Sono state usate derivate numeriche
Tolleranza = 1.81899e-012
Convergenza raggiunta dopo 10 iterazioni

Modello 2: NLS, usando le osservazioni 1-9
NR = alpha*D+beta*Anni

              stima      errore std.   rapporto t   p-value
  ------------------------------------------------------------
  alpha     -0.0134497   0.00435561      -3.088     0.0176   **
  beta       0.0795345   0.00718187      11.07      1.09e-05 ***

Media var. dipendente   1.111111   SQM var. dipendente     2.260777
Somma quadr. residui    0.808957   E.S. della regressione  0.339949
R-quadro non centrato   0.980216   R-quadro centrato      -0.020223
Log-verosimiglianza    -1.928893   Criterio di Akaike      7.857786
Criterio di Schwarz     8.252235   Hannan-Quinn            7.006566
Note: SQM = scarto quadratico medio; E.S. = errore standard

GNR: R-squared = 1.3726e-016, max |t| = 1.2888e-008
La convergenza sembra ragionevolmente completa
```

Figura 8.15. Output di *Gretl* relativo al quarto modello formulato

Si riporta in **Figura 8.16** la sola *frontiera di Pareto* relativa alle coppie di punti *1-CoD-*X_h. È possibile notare che il secondo ed il terzo modello, nonostante la maggiore complessità di formulazione, si adattano in modo minore alle classi, se confrontati con la capacità del quarto modello. In definitiva, quest'ultimo risulta quello ottimale poiché, a differenza dei 3 modelli precedenti, è contraddistinto da parametri significativi, come attestato dai valori di *p-value*, inferiori alla significatività scelta e pari a 0.05.

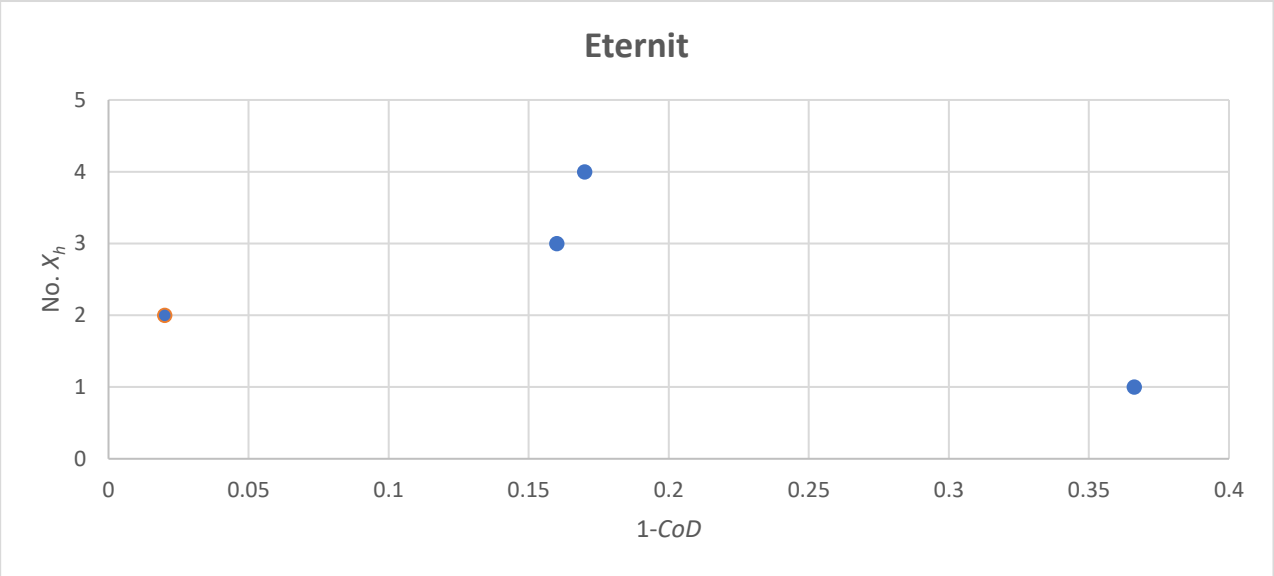


Figura 8.16. ***Frontiera di Pareto*** **relativa alle coppie di punti** ***1-CoD-X_h***

Ghisa grigia

Si riportano in **Tabella 8.13** le forme polinomiali stimate a partire dalle classi riportate in **Tabella 8.9**.

Tabella 8.13. Modelli formulati nel caso della ghisa grigia

Ghisa grigia				
Formula	CoD	No. X_h	No.A_j	1-CoD
$NR = 0.07039 \cdot D$	0.01	1	1	0.99
$NR = D^{0.46323}$	0.21	1	1	0.79
$NR = -2.50839 \cdot 10^{-5} \cdot D \cdot L + 3.56927 \cdot \ln(D)$	0.33	3	2	0.67
$NR = 2.84593 \cdot 10^{-7} \cdot D \cdot L \cdot A$	0.43	3	1	0.57
$NR = -1.09444 \cdot 10^{-7} \cdot D^{2.2287} \cdot L + 2.51136 \cdot \ln(D) + \ln(L)$	0.75	4	2	0.25

Nel caso della ghisa grigia 6 classi su 9 sono risultate non vuote ed è stato possibile formulare cinque differenti modelli polinomiali. I modelli, nel complesso, non forniscono indicazioni coerenti riguardo le variabili diametro e lunghezza. La variabile età è, invece, contemplata nel quarto modello e, all'aumentare di quest'ultima, aumenta il numero di rotture.
Tra i modelli formulati, il quinto è quello che meglio si adatta alle classi, con un numero di variabili esplicative pari a 4 e 2 polinomi. I parametri del modello, però, risultano tutti non significativi, ad eccezione dell'esponente relativo alla variabile diametro (**Figura 8.17**).

```
Risultati di gretl per giova 2018-07-06 10:08, pagina 1

Sono state usate derivate numeriche
Tolleranza = 1.81899e-012
Convergenza raggiunta dopo 324 iterazioni

Modello 2: NLS, usando le osservazioni 1-6
NR = alpha*D^gamma*L+beta*l_D+l_L

                stima        errore std.    rapporto t    p-value
  ------------------------------------------------------------------
  alpha      -1.09444e-07    3.64671e-07     -0.3001      0.7837
  beta        2.51136        0.821615         3.057       0.0551   *
  gamma       2.22878        0.680154         3.277       0.0465   **

Media var. dipendente    9.833333   SQM var. dipendente      8.207720
Somma quadr. residui     84.05900   E.S. della regressione   5.293361
R-quadro non centrato    0.750443   R-quadro centrato       -0.001140
Log-verosimiglianza     -16.43291   Criterio di Akaike       38.86582
Criterio di Schwarz      38.24110   Hannan-Quinn             36.36501
Note: SQM = scarto quadratico medio; E.S. = errore standard

GNR: R-squared = 7.77666e-015, max |t| = 1.52256e-007
La convergenza sembra ragionevolmente completa
```

Figura 8.17. Output di *Gretl* relativo al quarto modello formulato

Si riporta in **Figura 8.18** la sola *frontiera di Pareto* relativa alle coppie di punti *1-CoD-*X_h.

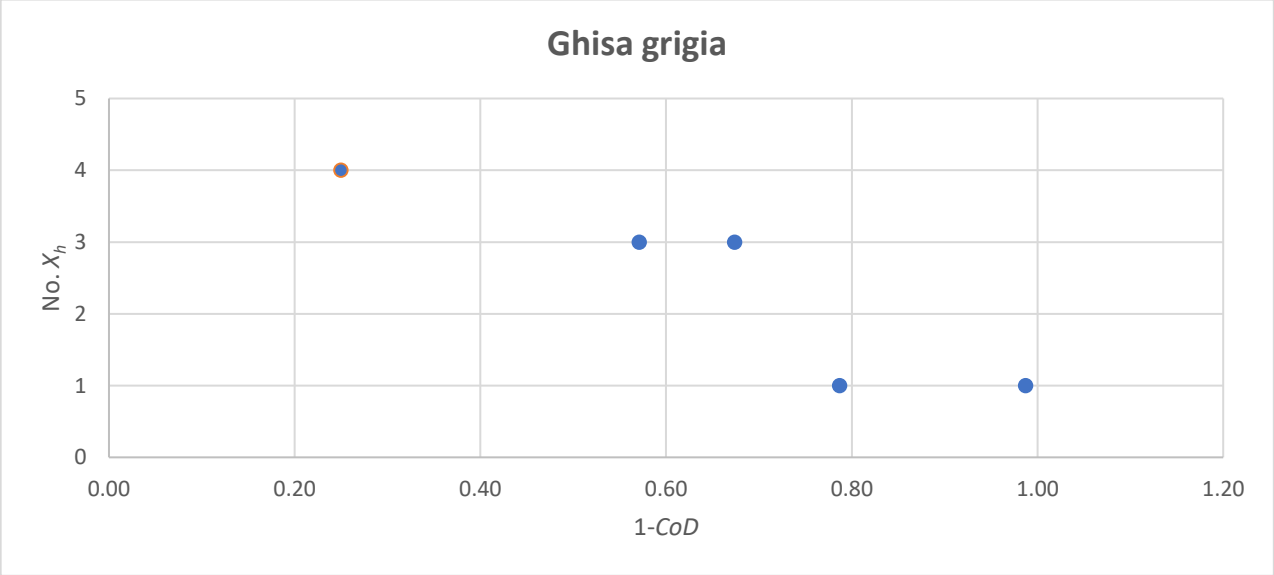


Figura 8.18. *Frontiera di Pareto* relativa alle coppie di punti *1-CoD-*X_h

Ghisa sferoidale

Si riportano in **Tabella 8.14** le forme polinomiali stimate a partire dalle classi riportate in **Tabella 8.10**.

Tabella 8.13. Modelli formulati nel caso della ghisa grigia

Ghisa sferoidale				
Formula	CoD	No. X_h	No.A_j	1-CoD
$NR = 1303.2 \cdot D^{-1.19060}$	0.49	1	1	0.51
$NR = -0.001819 \cdot D + 0.0002062 \cdot L - 0.161397 \cdot A$	0.56	3	2	0.44
$NR = 0.011171 \cdot D + 0.0001453 \cdot L - 0.0009607 \cdot A \cdot L$	0.59	4	2	0.41
$NR = 6.65793 \cdot D^{-0.835051} \cdot L^{0.469912} \cdot \ln(A)^{-1.34436}$	0.67	3	1	0.33

Nel caso della ghisa grigia tutte le classi sono risultate non vuote ed è stato possibile formulare quattro differenti modelli polinomiali. I modelli, nel complesso, non forniscono indicazioni coerenti riguardo le variabili lunghezza e diametro. Invece, all'aumentare dell'età si ritrova sempre una diminuzione del numero di rotture previsto.
Tra i modelli formulati, il quarto è quello che meglio si adatta alle classi, con un numero di variabili esplicative pari a 3 e un singolo polinomio. I parametri del modello, però, risultano tutti non significativi, come dimostrano i valori di *p-value* (**Figura 8.19**).

```
Risultati di gretl per giova 2018-07-06 10:22, pagina 1

Sono state usate derivate numeriche
Tolleranza = 1.81899e-012
Convergenza raggiunta dopo 388 iterazioni

Modello 5: NLS, usando le osservazioni 1-9
NR = alpha*D^beta*L^gamma*l_Anni^delta

               stima      errore std.    rapporto t    p-value
  ------------------------------------------------------------
  alpha        6.65793      67.6337         0.09844     0.9254
  beta        -0.835051      0.680460      -1.227       0.2744
  gamma        0.469912      0.735411       0.6390      0.5510
  delta       -1.34436       0.922001      -1.458       0.2046

Media var. dipendente    3.555556    SQM var. dipendente      2.650996
Somma quadr. residui     18.54582    E.S. della regressione   1.925919
R-quadro non centrato    0.670134    R-quadro centrato       -0.000006
Log-verosimiglianza     -16.02404    Criterio di Akaike       40.04807
Criterio di Schwarz      40.83697    Hannan-Quinn             38.34563
Note: SQM = scarto quadratico medio; E.S. = errore standard

GNR: R-squared = 2.03633e-013, max |t| = 9.73694e-007
La convergenza sembra ragionevolmente completa
```

Figura 8.19. Output di *Gretl* relativo al quarto modello formulato

Si riporta in **Figura 8.20** la sola *frontiera di Pareto* relativa alle coppie di punti *1-CoD-X_h*.

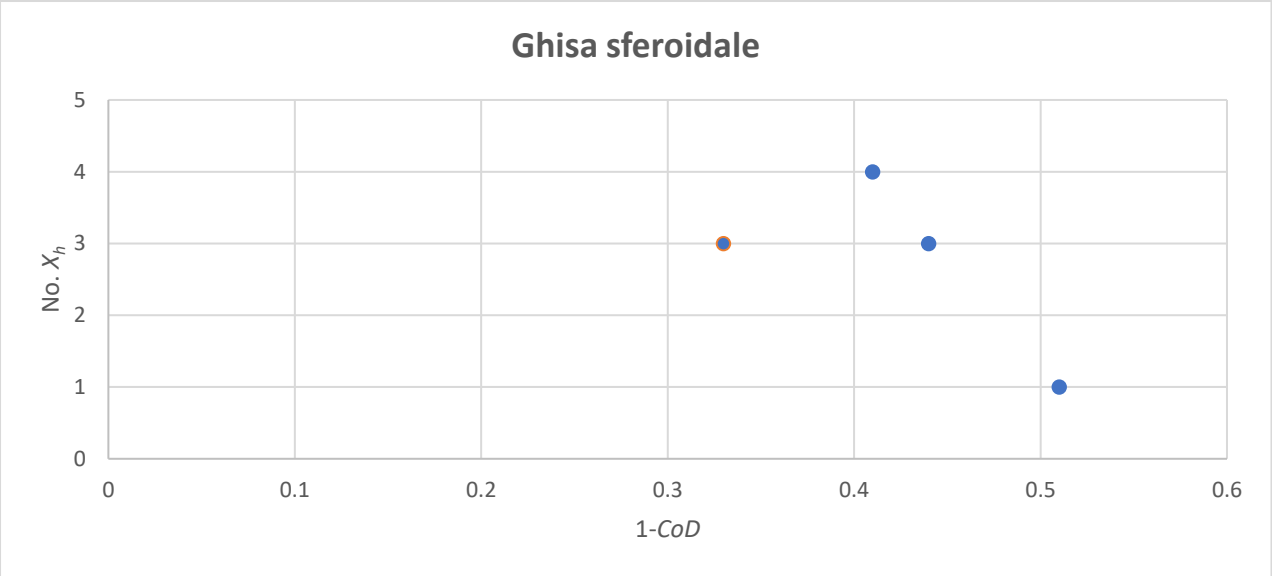


Figura 8.20. ***Frontiera di Pareto*** **relativa alle coppie di punti** ***1-CoD-X_h***

Infine, si riporta in **Tabella 8.14** l'applicazione dei modelli ottimali scelti alle classi del relativo materiale e il conseguente numero di rotture previsto.

Tabella 8.14. Confronto tra le rotture stimate e quelle previste dal modello

		Acciaio		Eternit		Ghisa grigia		Ghisa sferoidale	
Classe		NR	NR previste	NR	NR previste	NR	NR previste	NR	NR previste
1	1	3	2.84	0	-0.44	14	11.07	8	8.22
	2	2	2.12	\	\	20	15.28	8	4.80
	3	0	0.00	\	\	16	15.61	2	4.40
2	1	0	-0.30	\	\	\	\	3	3.91
	2	\	\	\	\	\	\	3	2.80
	3	\	\	6	5.48	0	3.42	3	3.30
3	1	0	0.15	\	\	1	-1.17	3	1.93
	2	\	\	4	4.58	\	\	1	1.07
	3	\	\	\	\	8	14.05	1	1.60

È importante notare che l'affidabilità dei risultati aumenta all'aumentare del numero di classi non vuote, poiché la stima dei parametri può avvenire su un numero maggiore di punti. Le rotture previste dai modelli approssimano in numero quelle osservate nei casi dell'acciaio, dell'eternit e della ghisa sferoidale, mentre mostrano grandi discostamenti per alcune classi della ghisa grigia.
Nonostante questi risultati, il modello polinomiale applicato a questo caso studio non ha portato a risultati soddisfacenti, poiché la natura dei dati e la distribuzione dei valori di diametri ed età non hanno permesso di ottenere 9 classi di elementi per ogni materiale. Di conseguenza, i modelli sono stati tarati facendo affidamento su un numero ridotto di punti e, nella maggior parte dei casi, le variabili indipendenti sono risultate non significative, nonostante dei buoni valori del coefficiente di determinazione. Inoltre, l'influenza delle variabili indipendenti sul numero di rotture non è coerente in tutti i modelli: in alcuni casi, i risultati indicano delle proporzionalità dirette tra il numero di rotture e una specifica variabile, in altri, invece, delle proporzionalità inverse.

La principale causa della difficile applicazione del modello di regressione polinomiale al caso studio potrebbe essere la limitata grandezza del campione di dati. Infatti, il punto di partenza è stata la *Tabella Esatta*, ottenuta dalla forte filtrazione dei dati contenuti nella tabella *ReteTorino*, e contenente le sole condotte complete di informazioni riguardo le variabili indipendenti esaminate.
In definitiva, l'applicazione di tale modello a un caso studio caratterizzato da dati più completi (in particolar modo nel campo dell'anno di posa) potrebbe portare a risultati migliori.

8.4 Conclusioni

In questo capitolo è stato applicato il modello di regressione polinomiale al caso studio in esame. Così come per la regressione logistica, il modello ha fatto affidamento ai dati contenuti nella tabella *Esatta*, contenente tutte le condotte complete di informazioni riguardo le variabili indipendenti esaminate.
Si riportano i passi percorsi per la stima dei differenti modelli di regressione.

1. È stata effettuata una stratificazione preliminare dei dati attraverso una suddivisione secondo materiali. Si sono ottenute le tabelle *EsattaGhisaGrigia*, *EsattaGhisaSferoidale*, *EsattaEternit* ed *EsattaAcciaio.* Su queste sotto-tabelle sono state applicate le formulazioni **1** e **2** per ottenere 9 classi per ogni materiale.
2. Prima di poter applicare le formulazioni **1** e **2**, per ogni sotto-tabella sono state esaminate le funzioni densità di probabilità e probabilità cumulata di diametri ed età (secondo le formulazioni **3** e **4**), in modo da poter ottenere delle classi comparabili in numero di elementi contenuti. La non omogena distribuzioni di valori di queste grandezze ha portato spesso a classi non simili e, in alcuni casi, a classi vuote.
3. Una volta ottenute le 9 classi per ogni materiale, sono stati stimati differenti modelli polinomiali ed è stata scelta la forma che meglio si è adattata al campione di dati. Per effettuare tale scelta, si è fatto affidamento al *coefficiente di determinazione* e all'analisi della *Frontiera di Pareto*.
4. I modelli stimati mostrano in molti casi variabili non significative e risultati contrastanti nell'interpretazione dei contributi delle singole variabili esplicative.

Alla luce dei risultati ottenuti, si è concluso che il modello di regressione polinomiale non è adatto a prevedere il numero di rotture delle classi di elementi presenti nel caso studio poiché:

- il campione di dati di partenza è di limitata estensione, a causa della selezione avvenuta nel passaggio dalla tabella *ReteTorino* alla tabella *Esatta* e la conseguente eliminazione di elementi sprovvisti di informazioni quali diametro, anno di posa, carico massimo e lunghezza;
- la distribuzione dei valori di diametri ed età ha portato a numerose classi vuote e, conseguentemente, un minore numero di dati sui quali tarare i modelli;
- i modelli presentano risultati non coerenti, in quanto l'apporto delle singole variabili indipendenti sul numero di rotture mostra dei trend differenti per le forme polinomiali formulate.

Per concludere, il modello di regressione polinomiale fornisce risultati poco affidabili, è contraddistinto da una limitata bontà di adattamento ai dati di partenza e, quindi, non si presta all'applicazione in questo lavoro di tesi.

Conclusioni

Il lavoro di tesi ha avuto la finalità di esaminare in che modo il numero di rotture osservate nel sistema acquedottistico di Torino tra il periodo 2006-2016 sia stato influenzato dalle caratteristiche delle condotte quali materiale, diametro, anno di posa, carico massimo e lunghezza.
Questa tesi è nata dalla stretta collaborazione tra il *Politecnico di Torino* e la società *SMAT* e si pone a valle del lavoro intrapreso dalla collega *Dottoressa Clara Ghigo* e terminato nel dicembre 2017 ("*Analisi dei guasti nella rete acquedottistica di Torino", Clara Ghigo; relatori: Luca Ridolfi, Fulvio Boano, Marco Scibetta)*. A partire da tale lavoro, si è deciso di analizzare più dettagliatamente l'applicazione e i risultati del modello di regressione logistica e di applicare un secondo modello statistico, quello di regressione polinomiale (*Berardi_2008*).
Successivamente, è stata presentata la composizione della rete idrica di Torino, oggetto di studio di questo lavoro, ed è stato analizzato il database messo a disposizione da *SMAT.* Una volta selezionati i campi di interesse statistico ed integrati i dati mancanti riguardo la posizione ed il carico massimo nelle condotte, sono state ricavate tre tabelle fondamentali: la tabella *Wwvcondottetorino*, contenente le sole informazioni riguardanti i guasti di tipo "*fuga condotta*" che hanno avuto luogo tra il 2006 e il 2016 nel comune di Torino, la tabella *ReteTorino*, contenente tutte le condotte della rete con relativa informazione di rottura/non rottura, e la tabella *Esatta*, contenente le sole condotte della precedente tabella caratterizzate da una piena copertura dei campi di interesse statistico. Dall'analisi della composizione della rete idrica di Torino si ricava che le condotte della rete sono caratterizzate principalmente dai materiali ghisa grigia e ghisa sferoidale, lunghezze inferiori ai 300 metri, diametri inferiori ai 100 millimetri ed informazioni riguardanti l'anno di posa che fanno riferimento prevalentemente al ventennio 1990-2010.
Dall'analisi dei dati riguardanti le rotture di tipo *"fuga condotta"* che hanno avuto luogo nel comune di Torino (contenuti nella tabella *Wwvcondottetorino*) negli anni 2006-2016, una volta integrate le informazioni riguardanti la pressione e l'anno di posa, si è ricavato che:

- i diametri caratterizzati dal maggior numero di rotture risultano quelli tra i 100 e i 150 millimetri. Normalizzando i guasti rispetto ai chilometri di condotte, la classe più vulnerabile risulta quella tra i 13 e i 40 millimetri;
- il materiale che presenta più guasti è la ghisa grigia, ma il polietilene e l'eternit presentano un tasso di fallanza maggiore;
- la ghisa grigia sembra essere più vulnerabile a basse temperature, al contrario di eternit, PEAD e acciaio che presentano maggiori guasti nei mesi estivi.

Dall'analisi dei dati ottenuti dall'intersezione tra la tabella delle rotture (*Wwvcondottetorino)* ed il database cartografico contenente tutte le condotte della rete, si è ricavato che:

- il materiale con più alto tasso di fallanza è l'eternit;
- il tasso di fallanza aumenta al diminuire del diametro;
- l'informazione relativa all'anno di posa è spesso mancante e la sua analisi fornisce indicazioni poco utili;
- il tasso di fallanza non sembra seguire sempre un determinato trend al variare della lunghezza delle condotte;
- il tasso di fallanza non sembra seguire sempre un determinato trend al variare del carico massimo.

Tali dati sono contenuti nella tabella *ReteTorino* (26998 condotte) ma presentano numerosi elementi caratterizzati da campi sprovvisti di copertura come *materiale*, *anno di posa*, *diametro*, *carico massimo*, e *lunghezza.* Questi risultano quindi inadatti per l'applicazione di un modello statistico e si è deciso di eliminare tutti gli elementi carenti di queste informazioni. Si è ottenuta così la tabella *Esatta* (5855 condotte). L'analisi di questa tabella, nonostante la rimozione di molti elementi, conferma i risultati ottenuti per la tabella *ReteTorino*.
Successivamente, è stato applicato il modello di regressione logistica multivariata alla tabella *Esatta*, dopo aver definito la variabile dipendente *rottura* (*flag*) e le variabili indipendenti *materiale*, *anno di posa*, *diametro*, *carico massimo*, e *lunghezza.* Dall'analisi delle 20 estrazioni (*EstrazioneEsatta)* caratterizzate dal 15% di rotture si è riscontrato che:

- in tutte le estrazioni, le variabili "*materiale*", "*anno di posa*", "*diametro*" e "l*unghezza*" risultano rilevanti nel modello;
- il carico massimo risulta non significativo in una sotto-estrazione e nella quasi totalità dei casi è meno significativo degli altri parametri;
- la probabilità di rottura diminuisce all'aumentare del numero assegnato ai materiali, dell'anno di posa e del diametro; la probabilità diminuisce all'aumentare del carico massimo e della lunghezza della condotta;

Esaminata la significatività dei parametri e la capacità del modello di adattarsi al campione di dati, a partire da ogni *EstrazioneEsatta* è stato estratto in maniera casuale l'80% degli elementi. In questa fase di *fitting* sono state definite le nuove sotto-estrazioni, denominate *EstrazioneEsatta80%*, ed è stato nuovamente applicato il modello di regressione logistica. Nel caso di non significatività di un parametro, la stima è stata effettuata una seconda volta in assenza di tale parametro. I risultati di questa fase hanno portato alle seguenti conclusioni:

- in tutte le estrazioni, le variabili "*materiale*", "*anno di posa*", "*diametro*" e "l*unghezza*" risultano significative nel modello;
- la variabile *"lunghezza"* risulta non significativa in due sotto-estrazioni su 20;
- il carico massimo "*pMax"* risulta non significativo in 9 sotto-estrazioni su 20. Nella quasi totalità dei casi, risulta comunque meno significativo delle altre variabili;
- la probabilità di rottura diminuisce all'aumentare del numero assegnato ai materiali, dell'anno di posa e del diametro; la probabilità diminuisce all'aumentare del carico massimo e della lunghezza della condotta.

Ne è seguita la fase di *testing*, nella quale i modelli precedentemente stimati sono stati applicati al 20% rimanente di ogni estrazione, denominata *EstrazioneEsatta20%*. Una volta valutata la probabilità di rottura di ogni condotta facente parte di queste estrazioni, i risultati sono stati riportati in grafici a bolle. L'analisi di questi grafici ha mostrato che per tutte le *EstrazioneEsatta20%* i modelli associano alle condotte realmente rotte una probabilità media di rottura maggiore della probabilità media calcolata per le condotte non rotte. La probabilità media per le condotte non rotte ammonta a 8.78%, nettamente inferiore a quella stimata per le condotte rotte e pari a 48.79%. Tali probabilità risultano superiori a quelle ricavate nel caso studio precedente, soprattutto nel caso delle condotte realmente rotte: infatti, nel precedente lavoro di tesi la probabilità media ammontava a 5.93%, valore ben inferiore a quello ricavato in questo caso. Si può dedurre che il modello così composto riconosce in maniera più efficiente le condotte rotte, poiché ne associa probabilità più elevate.
Infine, è stato valutato il valore di soglia (*cut-off*) oltre il quale una condotta caratterizzata da una determinata probabilità stimata dal modello possa considerarsi suscettibile di rottura. Una volta stimato un modello generale sull'insieme dei dati contenuti nella tabella Esatta, è stata analizzata la variabilità dei falsi positivi e dei falsi negativi in relazione al valore di soglia; si è ricavato che il valore ottimale di soglia ammonta al 27-28%. Infatti, nell'intervallo di variazione tra il 10% ed il 30%, la somma totale dei falsi positivi e dei falsi negativi presenta un massimo in corrispondenza del 10%, con 242 elementi totali, ed un minimo in corrispondenza delle probabilità 27-28%, con 83 elementi totali. Nello specifico, in corrispondenza della probabilità del 27% si sono ottenuti 13 falsi positivi e 70 falsi negativi. In corrispondenza della probabilità del 28%, invece, si sono ottenuti 12 falsi positivi e 71 falsi negativi.
In conclusione, il modello di regressione logistica che prende in esame le singole condotte e considera la lunghezza come una variabile indipendente si presta in maniera migliore a rappresentare il campione di dati messo a disposizione da *SMAT* e indica che tutte le condotte con una probabilità di rottura stimata superiore al 27-28% andranno probabilmente incontro a rottura nei 10 anni successivi al periodo di osservazione.

Infine, è stato tarato il modello di regressione polinomiale sul campione di dati contenuti nella tabella *Esatta*. A seguito di una preliminare stratificazione secondo il materiale,

sono state ottenute le sotto-tabelle correlate, per ognuna delle quali sono state calcolate 9 classi. Tale suddivisione non ha potuto prescindere dall'analisi delle funzioni densità di probabilità e probabilità cumulata di diametri ed età, affinché ogni classe potesse contenere un numero comparabile di elementi. Nonostante questo, la non omogenea distribuzione di queste due grandezze ha portato in alcuni casi a classi vuote e, conseguentemente, i modelli polinomiali sono stati tarati su un numero di punti inferiore a 9. Infatti, per ogni materiale, a partire dalle 9 classi, sono state ipotizzate differenti forme di modelli polinomiali ed è stato scelto quello ottimale, mediante considerazioni sul coefficiente di determinazione e sul numero e complessità dei polinomi che compongono l'espressione. Si è concluso che tale modello, in generale, non è adatto a prevedere il numero di rotture per le differenti classi di condotte, poiché il campione di dati di partenza è di estensione limitata. L'assenza di una sufficiente quantità di dati ha portato alla necessità di tarare i modelli su un numero di punti inadeguato per una corretta stima dei parametri. Per tale motivo, inoltre, ogni variabile indipendente mostra delle relazioni con la variabile dipendente (il numero di rotture) che variano a seconda delle differenti forme polinomiali.

Per concludere, il modello di regressione logistica multivariata è quello che meglio si adatta ai dati a disposizione per il caso studio *SMAT* e potrebbe essere in grado di identificare le classi di condotte più vulnerabili, permettendo un'efficiente programmazione degli interventi di manutenzione e sostituzione delle parti della rete. Tale modello risulta implementato dalle modifiche apportate, se si confrontano i risultati ottenuti con quelli del caso studio precedente.

Bibliografia

Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining [2012], "Introduction to linear regression analysis", quinta edizione, John Wiley & Sons.

Alvisi S., Franchini M. [2008], "Comparative analysis of two probabilistic pipe breakage models apllied to a real water distribution system".

Yong Wang, Tarek Zayed, M., Osama Moselhi [2008], "Prediction Models for Annual Break Rates of Water Mains".

Berardi L., Giustolisi O., Kapelan Z., Savic D.A. [2008], "Development of pipe deterioration models for water distribution systems using EPR".

Ghigo C. [2017], "Analisi dei guasti nella rete acquedottistica di Torino".

APPENDICE A

Si riporta di seguito una breve descrizione delle tabelle fondamentali presentate all'interno della tesi, in ordine alfabetico.

Esatta: tabella contenente le sole condotte complete di informazioni riguardo le variabili di rilevanza statistica: materiale, diametro, anno di posa, carico massimo e lunghezza. È ottenuta dalla tabella *ReteTorino*.

EsattaAcciaio, EsattaEternit, EsattaGhisaGrigia, EsattaGhisaSferoidale: tabelle ottenute a partire dalla tabella *Esatta*, una volta classificate le condotte secondo il materiale.

EstrazioneEsatta: tabelle ottenute dall'estrazione delle 106 condotte rotte e di 601 condotte non rotte (estratte in maniera casuale per 20 volte) contenute nella tabella *Esatta*. Tali tabelle presentano la percentuale di rotture caratteristica della tabella *ReteTorino.*

EstrazioneEsatta80%: tabelle ottenute dall'estrazione dell'80% (20 estrazioni casuali di 566 condotte) delle condotte dalla tabella *Esatta*.

EstrazioneEsatta20%: tabelle contenenti il restante 20% (20 estrazioni casuali di 141 condotte) delle condotte della tabella *Esatta*.

ReteTorino: tabella rappresentativa dell'intera rete di Torino contenente le informazioni riguardo: materiale, diametro, anno di posa, carico massimo, carico medio, intervallo di carico, posizione geografica e lunghezza delle condotte.

ReteTorinoAcciaio, ReteTorinoEternit, ReteTorinoGhisaGrigia, ReteTorinoGhisaSferoidale: tabelle ottenute a partire dalla tabella *ReteTorino*, una volta classificate le condotte secondo il materiale.

Vie: tabella contenente i dati riguardanti l'indirizzo degli interventi eseguiti all'interno della rete di distribuzione gestita da SMAT. È stata ricavata dal database *Maximo*.

Workorder: tabella contenente le caratteristiche generali degli interventi eseguiti dal 2006 al 2016 all'interno della rete di distribuzione gestita da SMAT. È stata ricavata dal database *Maximo*.

Woserviceaddress: tabella contenente i dati che permettono la georeferenziazione degli interventi all'interno della rete di distribuzione gestita da SMAT.

Wwvcondottetorino: tabella contenente gli interventi riguardanti i guasti di tipo *"fuga condotta"* verificatisi all'interno della rete del comune di Torino, nel periodo di osservazione 2006-2016