UNIVERSITA' DEGLI STUDI DI PADOVA
FACOLTA' DI SCIENZE STATISTICHE

# TESI DI LAUREA TRIENNALE

# VALANGHE: MODELLI PER LA PREVISIONE

Relatore: ch.mo prof. Stuart Coles

Laureando: Michele Cecotti

**Anno Accademico 2005 / 2006**

**Sommario**

## O. Introduzione

*"Lo dobbiamo alle vittime di queste catastrofi, il non lasciare intentata alcuna via per comprendere e prevedere con sempre maggior chiarezza le reazioni della natura agli interventi dell'uomo..."*

Robert Haefeli

PRIMO DIRETTORE DELL'ISTITUTO PER LO STUDIO DELLA NEVE E DELLE VALANGHE

Le valanghe sono masse di neve in caduta, a volte contenenti rocce, terra, ghiaccio, sono il fenomeno più distruttivo e che causa più vittime in montagna, questo anche perché non si hanno sintomi certi che ne rivelino un prossimo verificarsi e che permettano quindi delle misure di prevenzione. L'obiettivo principale di questo studio è poter prevedere tramite degli strumenti statistici se esiste la possibilità che si verifichi una valanga in base alle conoscenze su alcune particolarità del manto nevoso, come lo spessore del manto, la durezza della neve, temperatura del manto, o su specifici eventi metereologici, come l'intensità del vento, la temperatura massima e minima nell'arco delle 24 ore, le condizioni del tempo.

Le previsioni verranno effettuate mediante l'uso di modelli statistici di diversa natura per constatare se esiste una famiglia di modelli che meglio si adatta a questo tipo di variabili ambientali. I dati analizzati in questo studio sono stati presi dal sito internet della divisione meteorologica della regione trentino che mette a disposizione degli utenti un vasto database contenente i rilevamenti annuali di un cospicuo numero di stazioni presenti nel territorio della regione.

In figura 1 si può vedere la distribuzione delle stazioni (puntini gialli) nel territorio del Trentino Alto Adige, visto l'ampio numero delle stazioni presenti e non potendo

materialmente raccogliere tutti i dati di tutti i siti della regione sono stati scelti tre luoghi che non avessero nelle loro rilevazioni un eccessivo numero di valori nulli e che fossero a una certa distanza uno dall'altro.

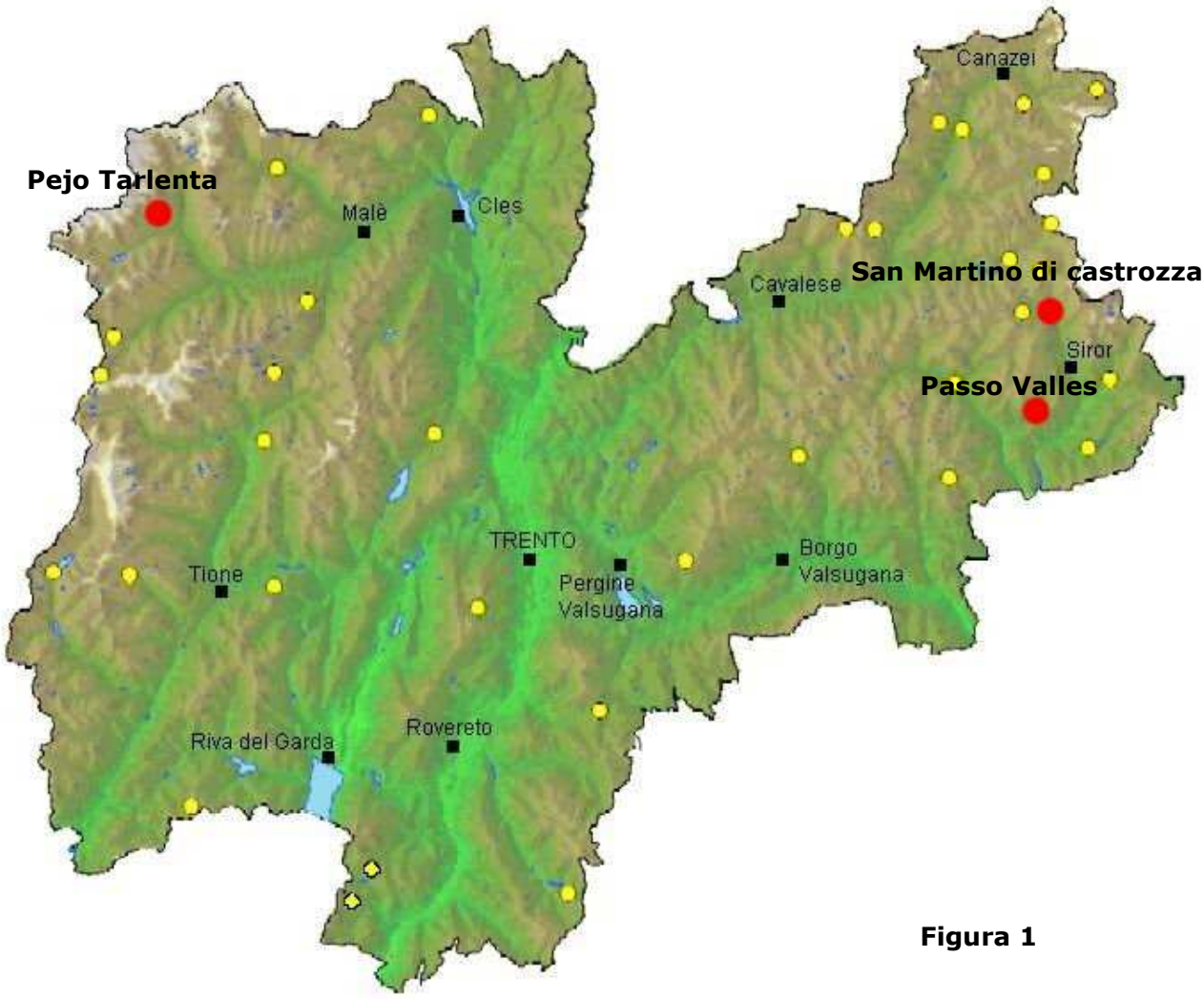



**Figura 1**

*Metodo di raccolta dei dati*

L'analisi del manto nevoso viene effettuata con una "sonda a martello", con l'esame a vista degli strati e con le misure delle temperature e delle densità. I dati ricavati da queste indagini vengono poi opportunamente riportati su una tabella.

**PROFILO PENETROMETRICO:**

Si realizza tramite una sonda formata da un insieme di tubi sormontabili, che misurano ciascuno un metro di lunghezza ed un chilogrammo di peso; il primo termina con una punta a forma di cono con l'angolo sommitale di 60° e la superficie resistente alla penetrazione di 12 cmq.

Si aggiunge un'asta graduata, nella parte superiore, lungo la quale si fa correre un peso mobile di un chilogrammo. Si affonda questa sonda nella neve, facendo cadere il peso con battute ripetute.

Questo metodo di misura dà un'idea abbastanza precisa della diversa resistenza dei vari strati che costituiscono il manto nevoso, a condizione che il sondaggio sia effettuato in aree opportunamente delimitate e ubicate in siti rappresentativi.

**PROFILO STRATIGRAFICO:**

Si realizza tramite un procedimento che, a partire dall'esame del manto nevoso, consente di ottenere preziose informazioni sulle caratteristiche dei diversi strati di neve al suolo. Il sondaggio con la sonda a martello ed il profilo sono complementari: se l'esame con la sonda fornisce dati unicamente quantitativi, il profilo stratigrafico dà maggiori informazioni sulla qualità della neve.

Le osservazioni fatte riguardano:

- la misura della temperatura: si inseriscono nel manto nevoso dei termometri, uno ogni 10 cm, se la coltre è di un metro di spessore, altrimenti ogni 10 cm fino alla profondità di 50 cm e quindi ogni 20 cm fino al suolo;
- individuazione degli strati ed annotazione delle differenze di altezza di questi;

- esame dei differenti strati , per ciascun strato, si devono valutare:
  - valore del contenuto di acqua allo stato liquido: si confeziona con una mano guantata una palla di neve e si valuta con una osservazione il grado di umidità della palla stessa;
  - forma dei grani: si prelevano dei campioni di neve e si esaminano con una lente, valutando l'entità dei processi di metamorfismo a cui la neve è soggetta;
  - durezza: si calcola con il test della mano, valutando empiricamente la maggior o meno facilità di infilare nel manto nevoso un pugno, quattro dita, un dito, una matita o una lama di coltello;
  - dimensione dei grani: si stimano osservando i cristalli di neve con la lente di ingrandimento dopo averli posati su una piastrina millimetrata.

- misura del peso specifico della neve appartenente ad ogni strato: si preleva una carota di neve con un apposito cilindro metallico e la si pesa con un dinamometro

## I. Raccolta dati

Nel sito della regione trentino i dati erano suddivisi per stazione e per stagione invernale, erano in forma tabulare, ad ogni riga corrispondeva una rilevazione quasi sempre in giorni

diversi e ad ogni colonna corrispondeva una delle variabili elencate di seguito.

DATA – data in cui è stata fatta la rilevazione.

ORA - ora in cui è stata fatta la rilevazione.

STAZ - stazione in cui è stata fatta la rilevazione

WW - condizioni del tempo al momento del rilievo.

N - nuvolosità espressa in ottavi di copertura del cielo.

V - visibilità orizzontale.

VQ1 - attività del vento nelle ultime 24h.

VQ2 - attività del vento nelle ultime 24h.

Ta - temperatura dell'aria al momento del rilievo

TMin – Minima temperatura nelle ultime 24h.

TMax – Massima temperatura nelle ultime 24h.

HS - altezza totale del manto nevoso al suolo(cm)

HN - altezza della neve caduta nelle ultime 24h.

FI - densità della neve fresca.

TH1 - temperatura della neve a 10cm di profondità.

TH3 - temperatura della neve a 30cm di profondità.

PR - penetrazione della sonda nella neve.

S - strato superficiale della neve.

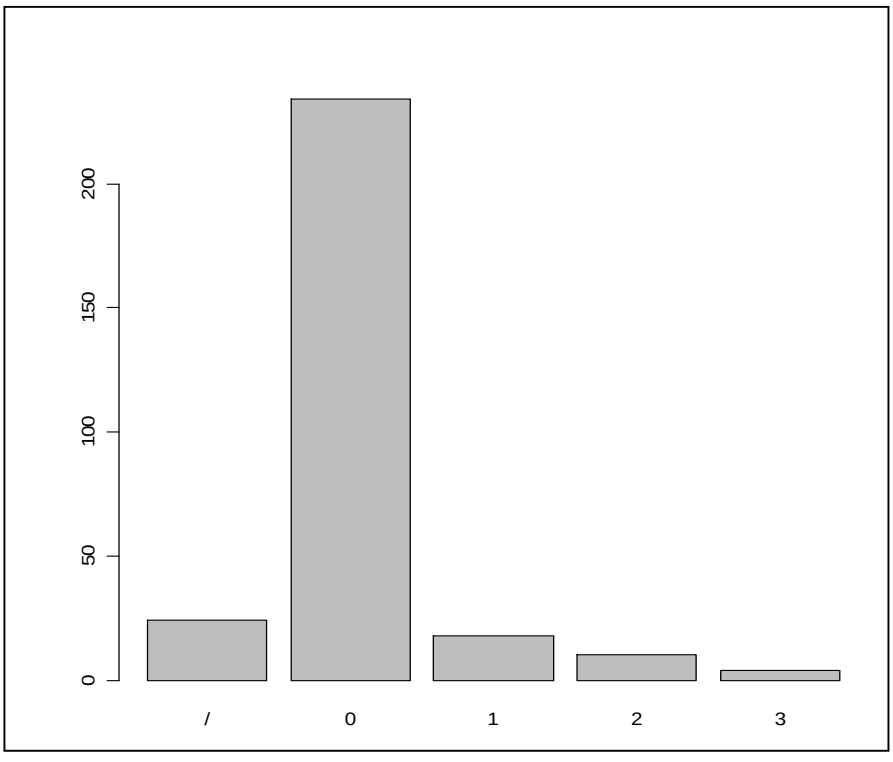
B - Variabile non indicata nella documentazione

L1 - mole delle valanghe osservate.

E' stata scelta la stagione da dicembre 2005 ad aprile 2006, in quanto aveva le rilevazioni più recenti. La forma tabulare in cui si trovavano i dati in internet non permetteva di poterli

inserire in R, programma che sarebbe stato usato per gli studi.
La tabella che conteneva i dati quindi è stata trasferita prima in un foglio excel e poi in un file .dat .
Nella prima osservazione dei dati è stata individuata la variabile risposta, quella su cui ci si sarebbe basati per costituire il modello e la sua previsione.
La variabile in questione è L1 che, in relazione alle altre variabili, indica se si è verificata una valanga e di che tipo essa sia (0 nessuna valanga, 1 piccola, 2 media, 3 grande).

Osservando i dati è stato notato che le variabili FI e N avevano parecchi dati mancanti e che in alcuni giorni (ogni giorno corrispondeva ad una riga della tabella) c'erano più dati mancanti; per ovviare a questi problemi è stato deciso di non inserire nel modello le variabili FI e N e di cancellare le righe con i dati mancanti.
E' stato notato anche che la variabile risposta L1 aveva un numero di risposte positive, valore > 0, molto inferiore e variegato rispetto al numero di risposte negative, valore uguale a 0 e a '/'.



Essendo questo un primo studio di questi dati è stato deciso di concentrare la previsione in un ambito più generale e di valutare quindi come risposta se l'evento si è verificato o no, lasciando ad un prossimo studio le valutazioni sull'entità delle valanghe .
Per realizzare questa differenziazione si è scelto di creare una variabile binaria di nome *risp* che valesse 0 quando L1 valeva 0 o '/' e 1 quando L1 valeva 1, 2 o 3; con questo metodo si perdeva un po' di accuratezza nelle risposte del modello, non si riesce a distinguere se la valanga è di bassa, media o grande entità, ma lasciava una base più solida su cui costruirlo.

## II. Costruzione dei modelli

Con i dati raccolti sono stati creati vari modelli; per risolvere il problema del sovra-adattamento dei dati e per garantire l'indipendenza tra dati di stima e dati di verifica, è stato deciso di dividere il campione in due sotto-campioni uno, quello che sarebbe servito per stimare il modello, con il 75% e l'altro, quello per la verifica, con il restante 25%.
Per garantire la completa casualità i dati dei due campioni sono stati estratti dal campione generale.

### II.a GLM

Vista la natura fattoriale della variabile risposta, è stato scartato a priori il modello lineare, il quale non dà buoni risultati con variabili di questo tipo; è stato invece utilizzato come primo approccio un modello lineare generalizzato (*generalized linear model,* GLM ) che a differenza del lineare semplice è ottimo per lo studio della relazione tra variabili quantitative e la variabile risposta dicotomica .

Il primo modello è stato stimato in base alla formula che includeva tutte le variabili, tranne naturalmente N e FI che per l'eccessivo numero di dati mancanti erano state escluse a priori e usando la funzione di legame tipica per la famiglia binomiale, dato che la variabile risposta era binaria.
Il risultato (vedi appendice A1) è stato un modello in cui la devianza residua era di 139.83 su 261 gradi di libertà un valore quindi molto buono, osservando però la significatività delle variabili si vedeva che solo 6 su 16 avevano un valore p significativo (p-value < 0.05).
Il passo successivo è stato stimare un nuovo modello che nella formula includesse soltanto le 6 variabili significative quindi *risp ~ VQ1 + TMAX + HS + PR + STAZ + B*.
Il modello ridotto (vedi appendice A2) aveva una devianza residua logicamente superiore al modello precedente ma comunque molto buona (164.03 su 283 d.f.), infatti se si confrontavano le devianze dei due modelli mediante il test della devianza utilizzando le differenze tra valori delle devianze e tra i gradi di libertà si otteneva un risultato che confermava il miglioramento del modello.
*164.03 - 139.83 = 24.20*
*283 – 261 = 22*
*24.20 su 22 d.f. ha un valore p pari a circa 0.33*

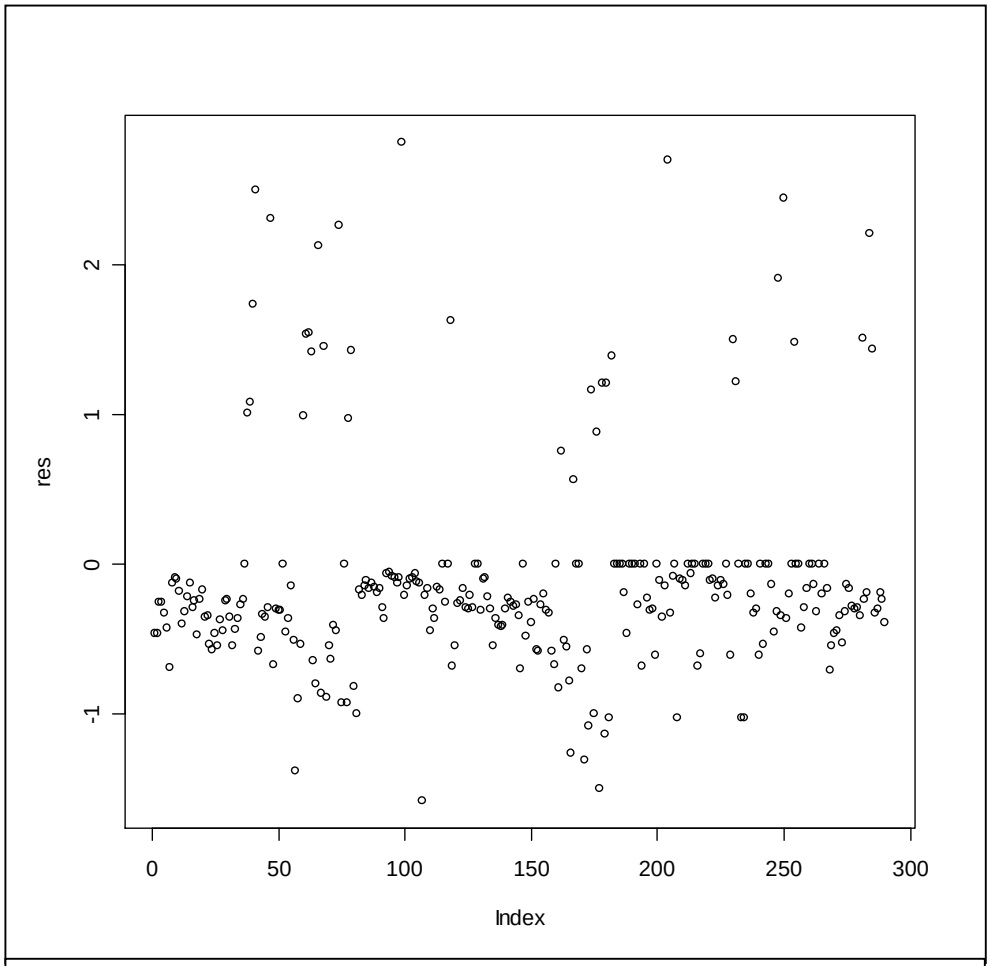
Analizzando però ogni singola variabile si poteva notare che 2 sulle 6 incluse non avevano un valore p significativo.
Si è quindi proceduto ad un'ulteriore semplificazione della formula su cui si stimava il modello togliendo le 2 variabili non significative: *risp~ TMAX + HS + STAZ + B.*
Il nuovo modello (vedi appendice A3) continuava a soddisfare il test sulla devianza residua (165.87 su 285 d.f.)
Test sulla devianza:
*165.87 – 164.03 = 1.84*
*285 – 283 = 2*

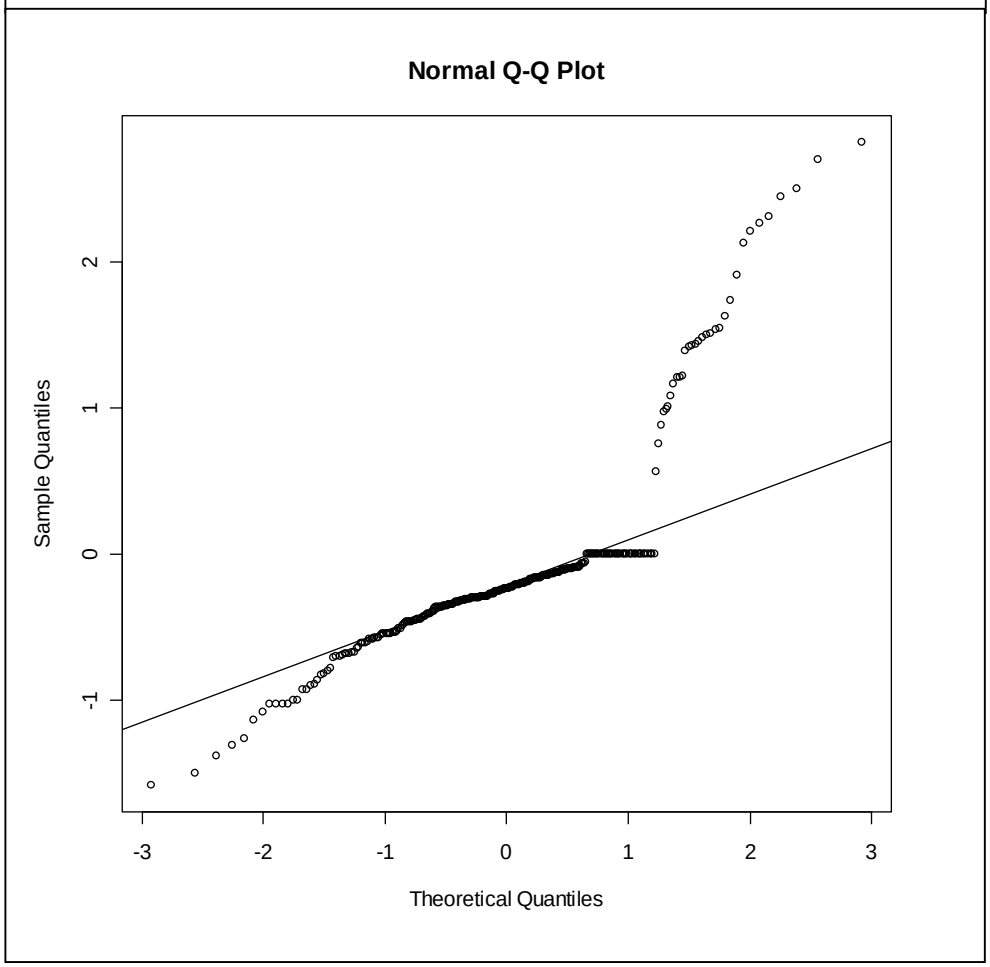
*1.84 su 2 d.f. ha un valore p pari a circa 0.40*

Il modello in questione non aveva variabili non significative all'interno della sua formula.
Esaminando meglio la formula del modello si nota che tre sono variabili che indicano uno stato della neve, mentre la quarta è la variabile che indica le stazioni; si potrebbe pensare che le tre variabili quantitative abbiano un peso diverso per ogni stazione perché forse in un certo luogo è maggiore l'influenza della temperatura massima o dello strato di neve accumulato e che quindi la relazione che le lega alla variabile STAZ non si additiva ma moltiplicativa.
Viene ipotizzata quindi la seguente formula:
*risp ~ (TMAX + HS + B) * STAZ*
e di conseguenza viene stimato il modello (vedi appendice A4). Nell'analisi del modello si vede che le tre variabili TMAX:STAZ, HS:STAZ, B:STAZ hanno un livello di significatività basso, non si può quindi confermare l'ipotesi che i tre fattori abbiano diverso peso per le tre diverse stazioni; questo fatto se confermato può essere importante in quanto mette tutte le cause di valanga allo stesso piano.
Si è quindi ritornati al modello precedente senza interazioni tra variabili, e lo si è scelto come il migliore della categoria dei GLM.

Come ultima verifica della bontà del modello è stata fatta l'analisi dei residui.
L'analisi dei residui serve per vedere se il modello nel predirre i valori è abbastanza equilibrato da produrre residui che siano equamente distribuiti e che non risultino anomali.
Il fatto che i residui non siano concentrati in aree particolari del grafico significa che il modello non ha una predisposizione predeterminata ad un certo tipo di risposta e che i dati sono equilibrati.Nella realtà dei fatti è molto difficile che si abbiano

residui perfettamente uniformi all'interno dell'intervallo e che non ci siano residui "a macchie".



I dati utilizzati per la verifica del funzionamento del modello non sono quelli utilizzati per la stima dello stesso.
La previsione viene effettuata con l'utilizzo del 25% dei dati totali che è stato trattenuto per la verifica appunto.

Nel primo grafico che visualizza i residui si può notare questi ultimi non sono distribuiti in modo uniforme all'interno dello spazio, ciò indica una non perfetta casualità.
Il risultato ottenuto nel primo grafico si rispecchia nel secondo, il quale mette a confronto i quantili dei residui con i quantili della normale; avendo dei residui eccezionalmente buoni si potrebbe vedere la linea presente nel grafico completamente coperta.
Nel nostro caso si può notare come nel centro del grafico i residui convergano verso la linea mentre nelle code si allontanino; questo fatto è accentuato nella coda di destra, dove più i residui si discostano dal centro del grafico più si distanziano dalla linea.
I residui in questo modello non si comportano molto bene; le cause di queste anomalie si posso ricercare nel basso numero di casi favorevoli ( casi in cui si è verificata la valanga) contro un alto numero di casi sfavorevoli( caso in cui non si è verificata la valanga ).

## II.b ALBERO DI CLASSIFICAZIONE

Dopo aver realizzato questo primo modello per avere una più ampia panoramica delle soluzioni al problema della previsione delle valanghe si è deciso di stimare con i dati un albero di classificazione.
La prima sostanziale differenza tra questo metodo e i GLM e che gli alberi di classificazione non sono parametrici, cioè non

viene costruito il modello in base a dei parametri, siano essi costanti o da trovare.

Tra le varie qualità di questo metodo c'è:

- la semplicità di comunicazione dei risultati a persone anche fuori dall'ambito statistico, l'albero viene utilizzato in moltissimi ambiti.
- La rapidità di calcolo, la procedura infatti non è molto onerosa dal punto di vista computazionale.
- L'uso di variabili discrete e categoriali, molto importante nel nostro caso in cui oltre alla variabile risposta *risp* altre variabili sono qualitative.
- La selezione naturale delle variabili più significative all'interno del modello, se nella realizzazione dell'albero una certa variabile non viene utilizzata è logico pensare che essa non si così importante.

Questo metodo porta anche degli svantaggi, i più grossi sono:

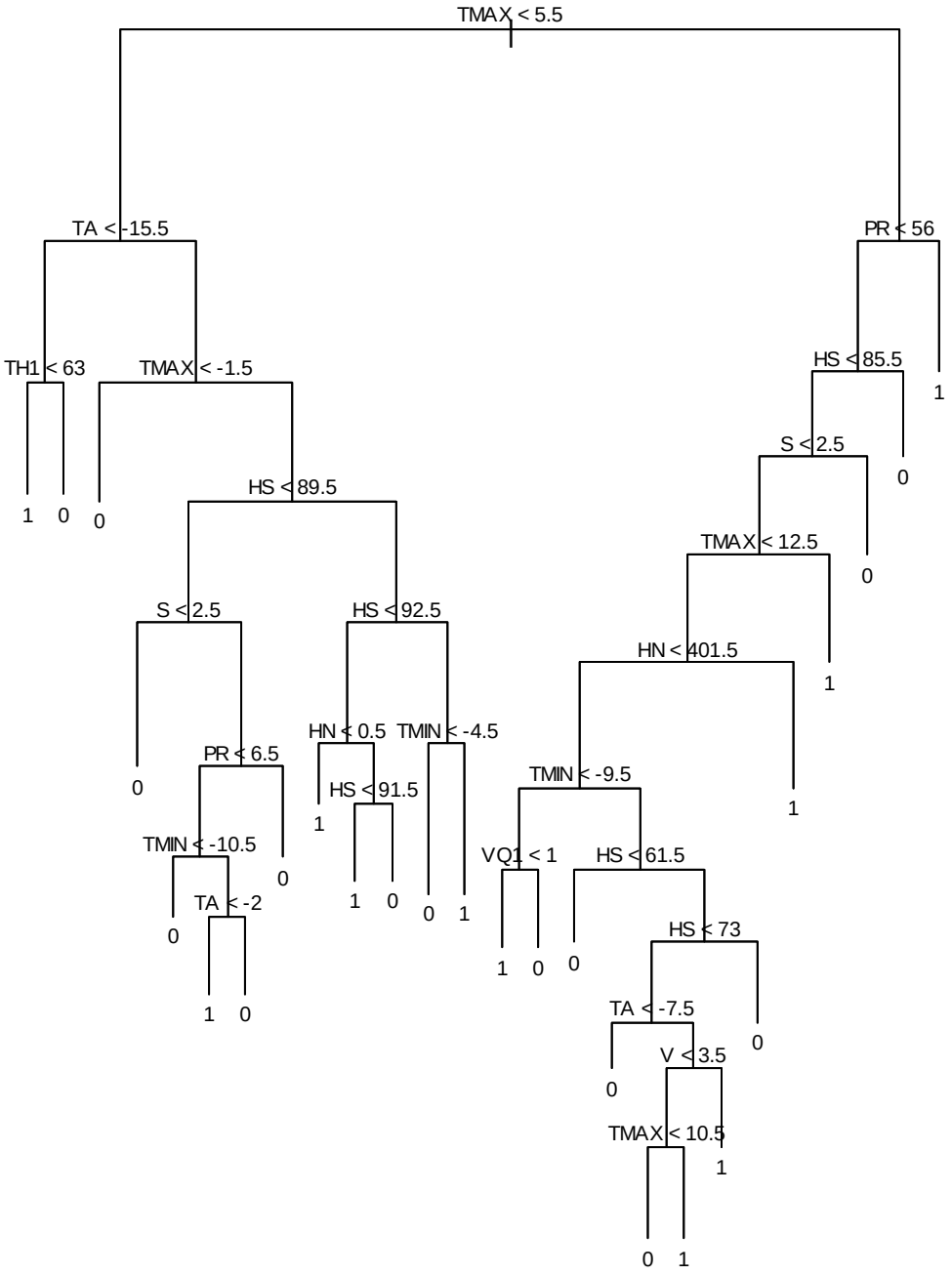
- Il fatto che sia non parametrico non permette di utilizzare procedimenti formali di inferenza statistica, come la verifica di ipotesi, la stima intervallare, etc.
- Non è semplice valutare realmente l'ordine di importanza delle variabili.

L'albero che viene a crearsi serve per decidere dove un'osservazione verrà posizionata, in base alle variabili che compongono il modello.

In ogni nodo finale dell'albero di ha un valore della variabile risposta mentre il percorso che bisogna fare per arrivare a quel nodo è deciso dalle altre variabili.

Un'osservazione partendo dalla radice dell'albero scende giù per ogni foglia, e ad ogni foglia prende una certa via in base al valore della sua variabile rispetto a delle costanti.

Alla fine l'osservazione raggiunge un nodo finale, il valore del quale indica il valore predetto dall'albero per quella osservazione.




Nell'albero che è stato creato ( vedi appendice B) si può notare che le variabili più presenti nelle foglie sono le stesse che compongono la formula del modello GLM, questa può essere indirettamente una conferma della validità dello strumento.

I valori predetti dall'albero di classificazione, sempre basandosi sul campione di verifica, a differenza del GLM che predice delle probabilità, sono dei valori esatti della variabile risposta, nel nostro caso 0 o 1.
Grazie a questa caratteristica possiamo utilizzare uno strumento diverso per verificare la bontà del modello, la tabella di errata classificazione.

*Analisi dei valori predetti dall'albero nella tabella di errata classificazione:*

```
     osserv
pred  0   1
   0  77   4
   1  14   2
errore totale:  0.185567
falsi positivi/negativi:  0.87500000 0.04938272 7
```

In questa tabella stilizzata sono stati messi a confronto i valori presenti ( *osserv* ) nei dati con i valori predetti dal modello ( *pred* ).
- 77 è il numero di valori predetti 0 che veramente valgono 0.($n_{11}$)
- 4 è il numero di valori predetti 0 che in realtà valgono 1, detti anche *falsi negativi.* ($n_{12}$)
- 14 è il numero di valori predetti a 1 che in realtà valgono 0, detti anche *falsi positivi.* ($n_{21}$)
- 2 è il numero di valori predetti a 1 che veramente valgono 1. ($n_{22}$)

i due test sotto la tabella indicano uno l'errore totale, cioè la somma dei falsi positivi e negativi diviso il numero complessivo delle osservazioni e l'altro molto simile ma diviso

per positivi e negativi è il quoziente dei falsi fratto la somma delle osservazioni per valori predetti.
I risultati ottenuti dalla tabella e dal test confermano le difficoltà incontrate nell'analisi dei residui, cioè il faticoso lavoro del modello per la bassa percentuale di casi favorevoli.
Per effettuare una ulteriore prova si è utilizzato l'insieme di stima per predirre i valori e poi è stata visualizza la tabella di errata classificazione.

```
  osserv
pred   0   1
   0 250   8
   1   8  24
errore totale:  0.05517241
falsi positivi/negativi:  0.25000000 0.03100775
```

In questo caso i dati certificano la bontà del modello nello stimare giustamente i dati che sono stati usati per crearlo, ma evidenziano una certa difficoltà nell'utilizzo di nuovi dati.

## II.c ANALISI DEL DISCRIMINANTE LINEARE

Il nuovo metodo proposto è l'analisi del discriminante lineare.
Questo metodo a differenza dei precedenti non è un adattamento di metodi già esistenti ma è stato creato apposta per i problemi di classificazione.
Altre qualità sono:

- La semplicità di calcolo nella stima dei parametri, nella funzione discriminante ed in generale del metodo.
- La qualità e la stabilità dei risultati, il metodo è affidabile e produce risultati validi in moltissimi casi.
- E' robusto rispetto alle ipotesi, anche se alcune ipotesi vengono violate produce comunque risultati validi.

Questo metodo ha in sé anche delle difficoltà quali:

- Ipotesi restrittive, il modello è costruito sotto ipotesi molto dettagliate.
- Graduatoria delle variabili, non ci sono tecniche semplici per valutare l'importanza di una variabile all'interno del modello.
- Le stime dei parametri non sono robuste a valori anomali.

Il metodo si basa su una funzione che viene chiamata discriminante.
La funzione è così formata:

$$d(x_o) = log\ \pi_k + log\ p_k\ (x_o)$$

in cui $\pi$ è il peso dato ad ogni classe mentre $p$ è la funzione di densità di probabilità di ogni classe, $k$ è l'indicatore della classe.
Il valore di $k$ che massimizza la funzione $d$ per un determinato $x_o$ è la classe a cui appartiene l'osservazione.

Il modello creato (vedi appendice C) ha dato questi risultati:
*campione di verifica*

```
     osserv
pred  0  1
   0 87  6
   1  4  0
errore totale:  0.1030928
falsi positivi/negativi:  1.00000000 0.06451613
```

*campione di stima*

```
      osserv
pred   0   1
   0 252  26
   1   6   6
errore totale:  0.1103448
falsi positivi/negativi:  0.50000000 0.09352518
```

Il metodo dell'analisi del discriminante non si comporta bene come vorremmo, con dei dati nuovi ( campione di verifica ) non riesce a assegnare bene i valori di risposta, infatti $n_{22}$, dati predetti a 1 che realmente valgono 1, vale 0; mentre nelle previsioni con i dati con cui è stato stimato raggiunge risultati migliori, $n_{22}$ vale 6, ma non ancora soddisfacenti.

## II.d RETI NEURALI

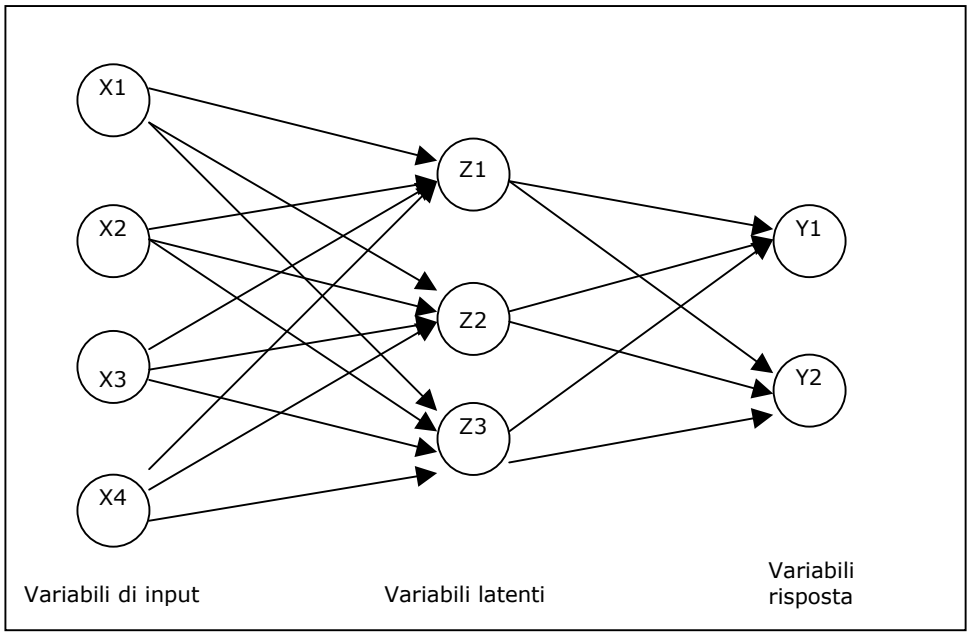
L'ultimo metodo che è stato utilizzato è quello delle reti neurali.
Le reti sono un'altro metodo di classificazione non parametrico, i loro principali vantaggi sono:

- La flessibilità, il metodo nella maggior parte dei casi riesce ad adattarsi bene ai dati.
- L'aggiornabilità sequenziale, il modello può essere semplicemente aggiornato ad ogni arrivo di nuovi dati.

Gli svantaggi invece sono:

- Arbitrarietà, nel modello alcuni parametri devono essere scelti arbitrariamente.
- Inferenza, come tutti i metodi non parametrici le reti neurali non hanno strumenti di inferenza.

L'idea da cui parte il metodo è che il collegamento tra variabili di input e la variabile risposta non si diretto, cioè che tra le due tipologie ve ne sia una terza che stia in mezzo.
Le variabili di questa terza tipologia si chiamano latenti e fanno parte di uno strato invisibile che però è fondamentale per arrivare dalle variabili di input alla variabile risposta.
Aver spezzato il collegamento in due parti ci permette di utilizzare due diverse funzioni di collegamento, una per ogni segmento.




Le due funzioni che collegano le variabili sono:

$$z_j = f_o(\Sigma_{h->j} a_{hj} x_h), \quad y_k = f_1(\Sigma_{j->k} \beta_{jk} z_j)$$

dove $a_{hj}$ e $\beta_{jk}$ sono parametri da stimare.
Il numero di variabili latenti è a discrezione di chi applica il metodo, in genere se ne provano vari e poi si decide di usare quello che da i risultati migliori

Il modello creato (vedi appendice D) ha dato questi risultati:

*campione di verifica*

osserv

pred 0 1

0 85 5

1 6 1

errore totale: 0.1134021

falsi positivi/negativi: 0.85714286 0.05555556

*campione di stima*

osserv

pred 0 1

0 253 17

1 5 15

errore totale: 0.07586207

falsi positivi/negativi: 0.25000000 0.06296296

I risultati ottenuti rispecchiano un po' quello che è accaduto con gli altri modelli, le reti neurali trovano difficoltà a lavorare su dati nuovi (campione di verifica), si comportano bene invece con i dati di stima.

## III. CONFRONTO FRA I MODELLI

Da una prima analisi dei quattro modelli che sono stati stimati, si può notare che il GLM è quello che meglio si adatta ai dati.

Per avere una maggiore certezza che il modello lineare generalizzato si il migliore tra i quattro è stato adottato un ulteriore strumento di confronto, le curve lift e roc.

Queste due curve in un grafico, rivelano l'efficacia di un modello mediante l'analisi della tabella di errata classificazione.

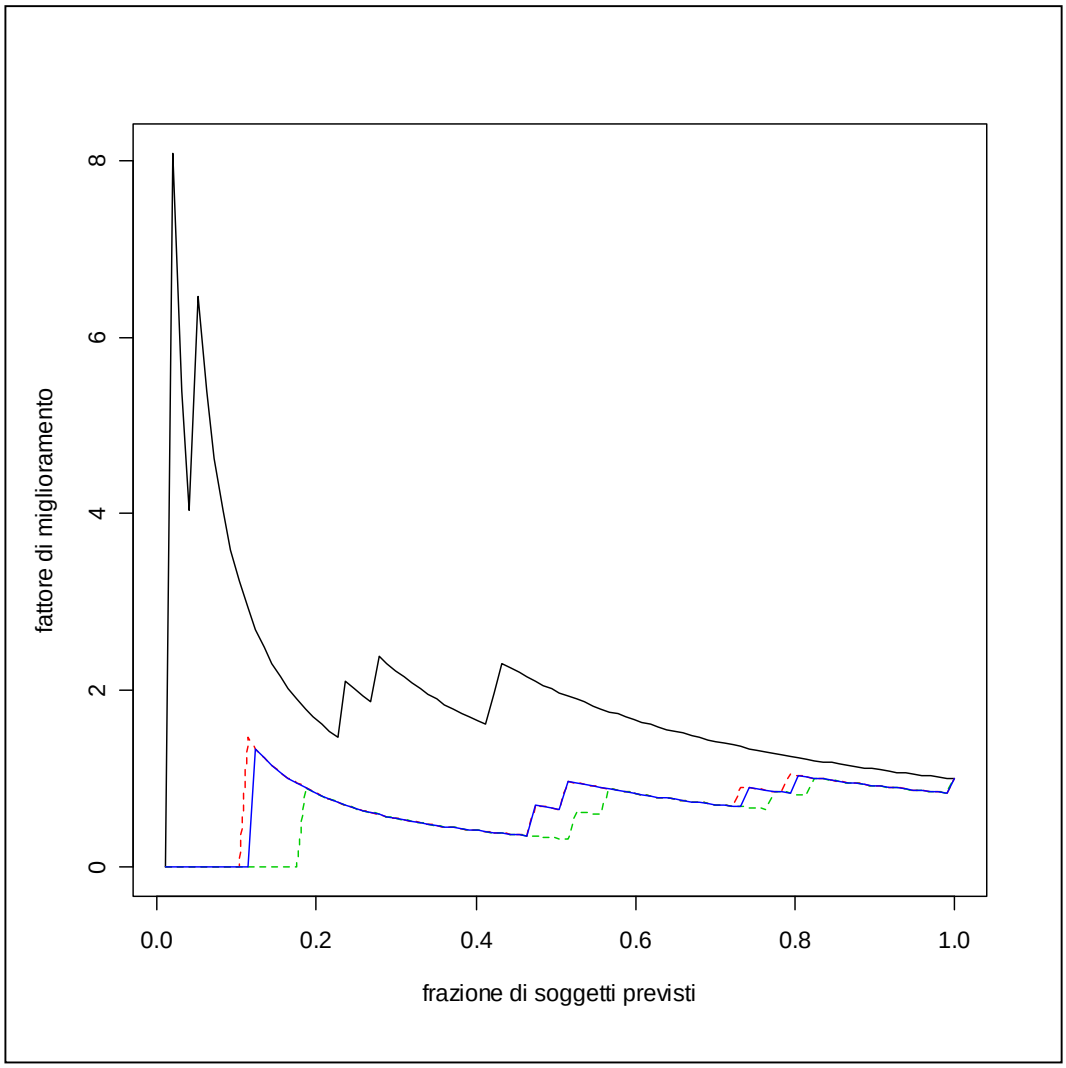
La curva lift va a vedere se i valori con maggiore probabilità predetti dal modello sono quelli effettivamente che hanno maggior frequenza nel campione originale.

La curva roc mette a confronto la scelta casuale, rappresentata dalla bisettrice del grafico, con il modello in esame; se la curva del modello è sempre maggiore della bisettrice allora esso porta un miglioramento; più la curva si distanzia dalla retta migliore è il modello.

Nella ordinata la curva roc ha la sensibilità, mentre nell'ascissa ha la 1-specificità.

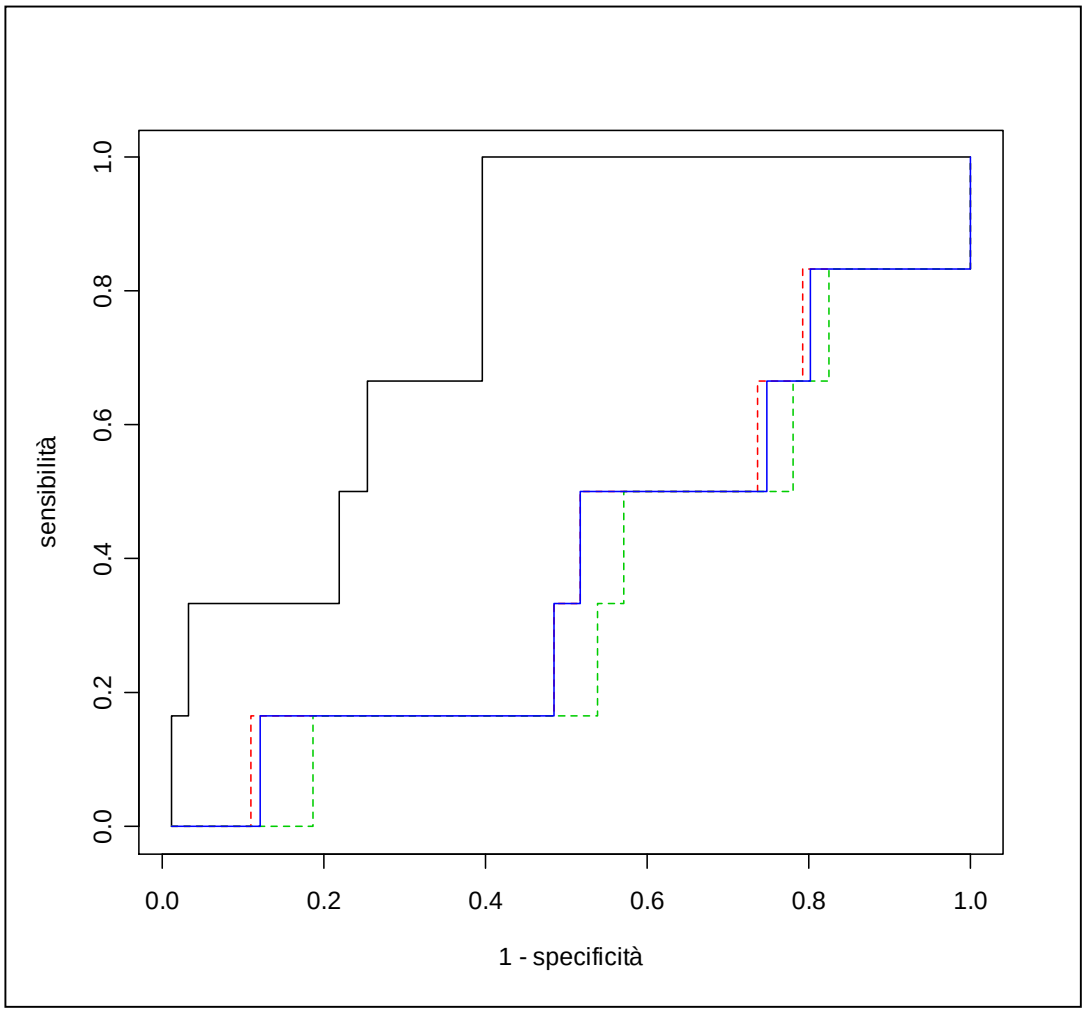
$$Sensibilità = n_{22}/n_{12}+n_{22} \quad Specificità = n_{11}/n_{11}+n_{21}$$

CURVA LIFT




Nella curva lift si può notare come il modello del GLM ( nero ) sia di molto migliore degli altri modelli, infatti il fattore di miglioramento è particolarmente elevato nella frazione di soggetti che realmente hanno una frequenza maggiore.

CURVA ROC




Nella curva roc si nota subito che la spezzata del modello GLM ( nero ) è l'unica che è sopra la bisettrice e quindi l'unica che apporta un miglioramento rispetto alla scelta casuale.

Il modello GLM si è dimostrato sotto tutti gli aspetti il migliore nell'adattarsi ai dati, ed è quindi quello che verrà scelto per un'eventuale previsione.

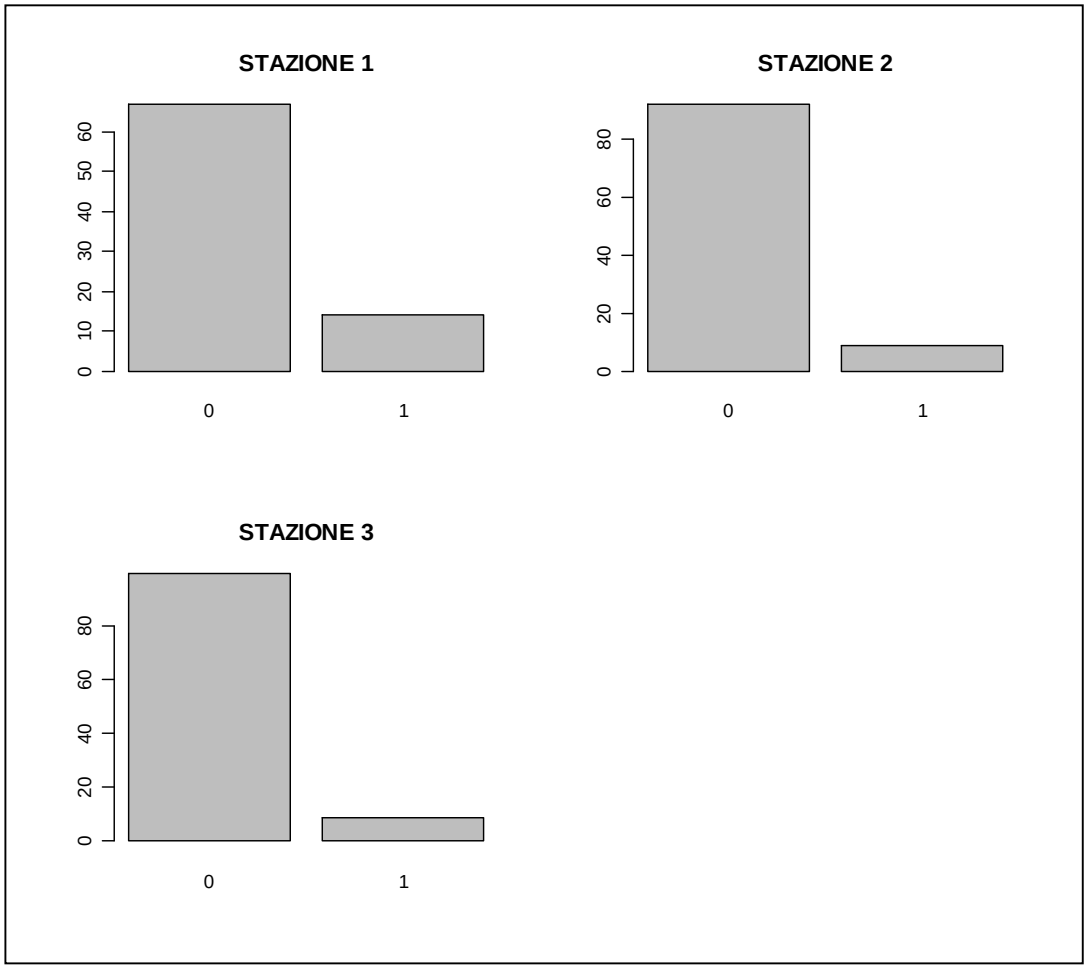
## IV. CONCLUSIONI

Il modello GLM si basa sulla formula:

*risp~ TMAX + HS + STAZ + B.*

Da questa formula si può capire da cosa dipenda maggiormente una valanga, infatti le quattro variabili in gioco sono quelle che ne influenzano di più il verificarsi.
La temperatura massima nelle 24 ore precedenti, l'altezza del manto nevoso, B (variabile non descritta) e la stazione sono le variabili più rilevanti.

L'inclusione della stazione meteorologica nella formula indica che ci sono luoghi più soggetti alle valanghe e luoghi meno soggetti; sarebbe interessante studiare quale stazione ha registrato una maggiore frequenza per poter identificare se questo fatto ha una rilevanza.
Nei grafici successivi viene realizzato un istogramma che visualizza il numero di casi favorevoli e casi sfavorevoli per ogni stazione.



Dando uno sguardo veloce agli istogrammi si può già intuire che la prima stazione ha un numero di eventi favorevoli maggiore delle altre due, ma per confrontare ancora meglio i vari casi, dato che le stazioni hanno un numero di osservazioni differenti, sono state realizzate le percentuali di casi favorevoli su tutto il campione:

| *STAZIONE 1* | *STAZIONE 2* | *STAZIONE 3* |
|---|---|---|
| *14 su 81* | *9 SU 101* | *9 SU 108* |
| *0.1728395* | *0.08910891* | *0.08333333* |
| Pejo tarlenta | Passo Valles | San Martino di Castrozza |

Come si può notare la stazione numero 1 ha una percentuale di casi favorevoli quasi del doppio superiore alle altre due stazioni che invece hanno percentuali tra di loro pressoché uguali.
Se torniamo a vedere la cartina del Trentino alto adige (Figura 1) si vede chiaramente che la prima stazione è ad ovest della regione mentre la seconda e la terza si trovano a est.
Con questi dati seppur pochi si potrebbe ipotizzare un maggiore rischio di valanghe nelle montagne a ovest del Trentino.

## V. STUDI FUTURI

Sarebbe interessante in futuro svolgere degli studi su un numero più elevato di stazioni, scegliendo appositamente i siti in modo che si possa confermare o smentire l'ipotesi suggerita prima.
Un altro studio interessante potrebbe essere quello sulla mole delle valanghe, lasciando la variabile risposta L1 e avendo quindi 4 classi in essa che determinano la dimensione del fenomeno naturale, si potrebbe vedere quali fattori sono i più determinanti per ogni tipo di valanga.

## APPENDICE A1

```
glm(formula = risp ~ WW + V + VQ1 + VQ2 + TA + TMIN +
TMAX + HS + HN + TH1 + TH3 + PR + CS + S + STAZ + B,
family = binomial, data = datas)

Deviance Residuals:
       Min              1Q         Median              3Q
Max
-1.5822806      -0.4212070      -0.2369731      -0.0002281
2.8209875

Coefficients:
             Estimate Std. Error   z value Pr(>|z|)
(Intercept) -2.076e+01  4.244e+03    -0.005 0.996098
WW          -6.583e-03  1.567e-02    -0.420 0.674434
V            1.464e-01  3.149e-01     0.465 0.641975
VQ1          6.933e-01  3.450e-01     2.010 0.044476 *
VQ2         -4.380e-01  3.263e-01    -1.342 0.179490
TA           1.271e-01  1.063e-01     1.195 0.231944
TMIN        -1.245e-01  1.139e-01    -1.094 0.274124
TMAX         1.613e-01  7.637e-02     2.112 0.034673 *
HS           3.430e-02  1.161e-02     2.954 0.003136 **
HN           1.456e-04  8.437e-04     0.173 0.862948
TH1         -9.774e-03  2.389e-02    -0.409 0.682438
TH30         1.701e+01  4.244e+03     0.004 0.996802
TH350        1.609e+01  4.244e+03     0.004 0.996975
TH351       -2.219e+00  5.745e+03 -0.000386 0.999692
TH352        1.563e+01  4.244e+03     0.004 0.997062
TH353        1.531e+01  4.244e+03     0.004 0.997121
TH354        1.458e+01  4.244e+03     0.003 0.997258
TH355        1.572e+01  4.244e+03     0.004 0.997045
TH356        1.585e+01  4.244e+03     0.004 0.997021
TH357       -1.984e+00  5.028e+03 -0.000395 0.999685
TH358       -2.846e+00  4.944e+03    -0.001 0.999541
TH359        1.633e+01  4.244e+03     0.004 0.996931
TH360       -1.776e+00  8.707e+03 -0.000204 0.999837
TH361       -1.685e+00  6.464e+03 -0.000261 0.999792
PR           3.116e-02  1.351e-02     2.307 0.021040 *
CS           4.538e-02  9.533e-02     0.476 0.634037
S           -9.910e-04  2.762e-01    -0.004 0.997137
STAZ        -7.885e-01  3.933e-01    -2.005 0.044968 *
B            2.158e+00  6.179e-01     3.493 0.000478 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 201.40  on 289  degrees of freedom
Residual deviance: 139.83  on 261  degrees of freedom
AIC: 197.83

Number of Fisher Scoring iterations: 18
```

## APPENDICE A2

```
glm(formula = risp ~ VQ1 + TMAX + HS + PR + STAZ + B,
family = binomial,
    data = datas)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2298  -0.4809  -0.3103  -0.1724   3.2592

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -4.29788    1.11034  -3.871 0.000108 ***
VQ1          0.11701    0.24425   0.479 0.631903
TMAX         0.20651    0.05322   3.880 0.000104 ***
HS           0.02991    0.00965   3.100 0.001935 **
PR           0.01424    0.01041   1.368 0.171313
STAZ        -0.62682    0.29759  -2.106 0.035178 *
B            1.61454    0.49910   3.235 0.001217 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 201.40  on 289  degrees of freedom
Residual deviance: 164.03  on 283  degrees of freedom
AIC: 178.03

Number of Fisher Scoring iterations: 6
```

## APPENDICE A3

```
glm(formula = risp ~ TMAX + HS + STAZ + B, family =
binomial,
    data = datas)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2635  -0.4640  -0.3133  -0.1822   3.2228

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.531649   0.934650  -3.779 0.000158 ***
TMAX         0.167005   0.042116   3.965 7.33e-05 ***
HS           0.028494   0.009584   2.973 0.002948 **
STAZ        -0.709353   0.290434  -2.442 0.014590 *
B            1.534871   0.490054   3.132 0.001736 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 201.40  on 289  degrees of freedom
Residual deviance: 165.87  on 285  degrees of freedom
AIC: 175.87

Number of Fisher Scoring iterations: 6
```

## APPENDICE A4

```
glm(formula = risp ~ (TMAX + HS + B) * STAZ, family =
poisson,
    data = dati)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0207  -0.4392  -0.2897  -0.1826   2.7736

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.569299   1.772370  -2.578  0.00994 **
TMAX         0.204271   0.084929   2.405  0.01616 *
HS           0.033496   0.020198   1.658  0.09724 .
B           -1.774297   2.378505  -0.746  0.45568
STAZ        -0.236402   0.864995  -0.273  0.78462
TMAX:STAZ   -0.040056   0.043316  -0.925  0.35510
HS:STAZ     -0.003091   0.010080  -0.307  0.75915
B:STAZ       1.165941   0.876852   1.330  0.18362
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 176.38  on 386  degrees of freedom
Residual deviance: 136.50  on 379  degrees of freedom
AIC: 228.5

Number of Fisher Scoring iterations: 6
```

## APPENDICE B

```
  #seleziono due parti del campione che mi serviranno
per valutare
  #in seguito l'albero.
  parte1 <- sample(1:NROW(datas), 200)
  parte2 <- setdiff(1:NROW(datas), parte1)
  #creazione dell'albero
  t1 <-
tree(fvala,data=datas[parte1,],control=tree.control
           (nobs=length(parte1), minsize=2,
mindev=0.002))
  plot(t1)
  text(t1,cex=0.6)
  t2 <- prune.tree(t1, newdata=datas[parte2,])
  #t2  serve  per  vedere  dove  a  che  ramo  l'albero
raggiunge la sua                                #massima
efficienza, perchè poi inizia a decadere
  plot(t2)
  J <- t2$size[t2$dev==min(t2$dev)]
  #poto l'albero t1 al ramo che mi ha suggerito t2
  t3<-prune.tree(t1, best=J)
  plot(t3)
  text(t3,cex=0.8)
  ptree<- predict(t3, newdata=datav, type="class")
  ptree2<- predict(t3, newdata=datas, type="class")
  tabella.sommario(ptree, datav$risp)
  tabella.sommario(ptree2, datas$risp)
```

## APPENDICE C

```
 library(MASS)
  mlda<-lda(risp ~ WW + V + VQ1 + VQ2 + TA + TMIN + TMAX
  + HS + HN +     TH1 + TH3 + PR + CS + S + STAZ + B,
  data=datas)
  plda<-predict(mlda, newdata=datav)
  plda2<-predict(mlda, newdata=datas)
  tabella.sommario(plda $class, datav$risp)
  tabella.sommario(plda2 $class, datas$risp)
```

## APPENDICE D

```
library(nnet)
  mnnet<- nnet(risp ~ WW + V + VQ1 + VQ2 + TA + TMIN +
TMAX + HS + HN +    TH1 + TH3 + PR + CS + S + STAZ + B,
data=datas, decay=0.002, size=5,  maxit=1000)
#maxit è il numero massimo di iterazioni, size è il
numero delle variabili latenti, decay è la precisione di
calcolo.
pnnet<- predict(mnnet, newdata=datav, type="class")
pnnet2<- predict(mnnet, newdata=datas, type="class")
tabella.sommario(pnnet, datav$risp)
tabella.sommario(pnnet2, datas$risp)
```

## APPENDICE E

```
a1<-lift.roc(pglm, g,type="crude")
premere <cr>
a2<-lift.roc(plda$class, g,type="crude")
premere <cr>
 a3<-lift.roc(pnnet, g,type="crude")
premere <cr>
 a4<-lift.roc(ptree, g,type="crude")
```

```
premere <cr>
#sovrapposizione dei grafici delle curve lift
plot(a1[[1]], a1[[2]], type="l", col=1, pch=1, xlab="frazione
soggetti previsti", ylab="fattore di miglioramento")
 lines(a2[[1]], a2[[2]], type="l", cex=0.75, col=2, lty=2, pch=2)
 lines(a3[[1]], a3[[2]], type="l", cex=0.75,  col=3, pch=7,lty=2)
 lines(a4[[1]], a4[[2]], type="l", cex=0.75, col=4,lty=1, pch=4)

#sovrapposizione dei grafici delle curve roc
 plot(a1[[3]], a1[[4]], type="l", col=1, pch=1)
 lines(a2[[3]], a2[[4]], type="l", cex=0.75, col=2, lty=2, pch=2)
 lines(a3[[3]], a3[[4]], type="l", cex=0.75,  col=3, pch=7,lty=2)
 lines(a4[[3]], a4[[4]], type="l", cex=0.75, col=4,lty=1, pch=4)
 lines(a5[[3]], a5[[4]], type="l", cex=0.75, col=6,lty=1, pch=4)
```

## BIBLIOGRAFIA


(1) Azzalini, A. ; Scarpa B. Analisi dei dati e data mining **2004** 92 – 95 , 100 – 105 , 125 – 131 , 136 – 141

(2) Regione Trentino Alto Adige http://www.meteotrentino.it archivio dati , metodo di raccolta

(3) McClung, D. ; Schaerer, P. Manuale delle valanghe **1996** 1 – 5

(4) Bortot, P., Ventura, L. e Salvan, A. (**2000**). Inferenza statistica: applicazioni con S-Plus e R. Cedam, Padova.

(5) R project for statistical computing http://www.r-project.org programma R, library nnet, library MASS, library tree.

(6) McCullagh, P. e Nelder, J.A. (1989). Generalized Linear Models, 2nd Edition. Chapman & Hall, London.