

UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICHE E TECNOLOGIE INFORMATICHE



TESI DI LAUREA

**ANALISI DELLA MULTIRESISTENZA
AGLI ANTIBIOTICI IN *SALMONELLA* SPP.**

**ANALYSIS OF ANTIBIOTIC MULTIRESISTANCE IN
SALMONELLA SPP.**

Relatore: Ch.mo Prof. FORTUNATO PESARIN

Laureando: LUCA PASINATO

MATRICOLA N° 465628/STI

ANNO ACCADEMICO 2007-2008

Introduzione al problema

Scopi e descrizione

Lo scopo di questo studio è quello di valutare la prevalenza di Salmonella antibiotico-resistente in alcune popolazioni di animali e di identificare i possibili fattori di rischio per l'antibiotico-resistenza e per la multiresistenza.

I dati che verranno analizzati si riferiscono a campioni di salmonella isolati in Veneto, nel Triveneto e in altre province italiane nel biennio 2005-2006. Ogni unità viene testata nei confronti di 16 antibiotici e viene classificata come suscettibile, intermedia o resistente al dato antibiotico. La multiresistenza è definita come la resistenza ad almeno 4 antibiotici.

La prima fase dello studio si concentrerà sul matching e la pulizia dei dati (data la particolare natura del campionamento che verrà descritto nella sezione successiva). La seconda fase è centrata sull'analisi statistica dei dati, al fine di studiare i fattori di rischio per la multiresistenza, verranno utilizzate alcune tecniche di data mining per sondare la struttura dei dati e cercare di definire un modello con una buona capacità predittiva, valutando poi l'impatto delle singole covariate sul fenomeno osservato.

Il piano di monitoraggio Enter-Vet

I dati arrivano dal programma Enter-Vet.

Il sistema Enter-Vet, attivo dal 2002, ha la finalità di raccogliere i dati a livello nazionale relativi agli isolamenti di Salmonella spp. da campioni di origine veterinaria. I nodi della rete Enter-Vet sono gli istituti Zooprofilattici Sperimentali, con il coordinamento del Centro di Referenza Nazionale per le Salmonellosi. Gli istituti inviano al Centro di Referenza i dati relativi alla tipizzazione dei ceppi di Salmonella attraverso un sistema informatizzato, oltre che ad alcuni stipiti (in particolare i ceppi appartenenti ai sierotipi Enteritidis e Typhimurium) da sottoporre a tipizzazione fagica.

Tutti i dati vengono inviati dal Centro di Referenza all'Istituto Superiore di Sanità, che coordina a livello nazionale la rete europea Enter-net, che riceve anche le notifiche relative agli isolamenti da campioni di origine umana ed alimentare.

La nostra popolazione di riferimento sono dunque le salmonelle giunte in Istituto. Data la particolare natura delle salmonelle vengono

prelevati più campioni durante l'attività di raccolta dati, per avere diverse matrici da analizzare ed essere quindi sicuri di riscontrare il ceppo salmonellare che potrebbe infestare il luogo controllato. Può capitare quindi che uno stesso tipo di prelievo multiplo, proveniente da uno stesso luogo e in una stessa data dia risultati tutti uguali per quanto riguarda la sierotipizzazione e l'antibiotico-resistenza; queste osservazioni devono venire considerate sia dal punto di vista epidemiologico, sia all'interno della ricerca come una singola infezione di salmonella.

E' stato dimostrato infatti che dal punto di vista microbiologico, vista la natura altamente infettiva delle salmonelle, i ceppi che vengono isolati durante questi controlli, presentano profili genetici omogenei. Questi risultati vanno quindi 'puliti' da eventuali ridondanze che porterebbero ulteriori 'distorsioni' che si aggiungerebbero a quelle che derivano dal fatto che il programma Enter-Vet è un piano di monitoraggio e quindi non sottoposto ad una logica di campionamento statistico.

Salmonelle e determinazione del problema

Descrizione generale

Il genere *Salmonella* (dal nome del veterinario D.E. Salmon) comprende i batteri di dimensioni 0.7-1.5 nm x 2.0-2.5 nm talora mobili per la presenza di flagelli.

Habitat e diffusione

Il serbatoio delle salmonelle è rappresentato dall'intestino di tutti gli animali a sangue caldo e a sangue freddo; tali batteri possono sopravvivere per oltre 9 mesi nell'ambiente, soprattutto nei terreni umidi, nell'acqua, nei materiali fecali e negli alimenti per animali quali soprattutto le farine di ossa e di pesce.

La maggior parte dei ceppi è cosmopolita, diffusa cioè in tutte le parti del mondo, mentre altri sono localizzati in una particolare regione del globo.

Struttura antigenica

La struttura antigenica delle salmonelle è piuttosto varia e complessa. Ogni ceppo isolato può essere descritto in base ai principali antigeni che lo caratterizzano (Kauffmann-White).

La formula antigenica dei numerosissimi sierotipi di salmonelle esistenti è formata da 3 parti, ognuna separata dalle altre mediante il segno ':', che rappresentano, nell'ordine: l'antigene¹ O, l'antigene H della fase 1 e l'antigene H della fase 2.

Sensibilità agli antibiotici

Le salmonelle sono sensibili a vari antibiotici, tra i quali i più utilizzati sono le betalattamine (ampicillina, amoxicillina e cefalotina), gli aminoglicosidi (kanamicina, streptomina, gentamicina, apramicina e amikacina), le tetracicline, i chinoloni e l'associazione sulfametazolo-trimetoprim. Gli antibiotici utilizzati nel nostro studio sono: l'Acido nalidixico, l'Ampicillina, il Cefotaxime, il Ciproflaxin, Cloramfenicolo, la

¹ Antigene: Qualunque sostanza che, venendo a contatto con un organismo, è in grado di stimolare in questo la produzione di anticorpi specifici e di scatenare una risposta del sistema immunitario. (Microsoft® Encarta® Enciclopedia Online, (2008))

Gentamicina, la Colistina, la Kanamicina, la Streptomina, il Trisulfamidico, la Tetraciclina, il Sulfametolo/Trimetoprim, l'Amoxicillina, l'Enrofloxacin, le Cefalotine e il Ceftazidime.

Determinazione del problema

Va ricordato che la maggior parte degli oltre 2400 sierotipi di salmonella è in grado di procurare malattie all'uomo e circa il 2% dei casi che presentano complicazioni arriva a morte. I decessi nell'UE sono circa 200 ogni anno.

Da questo si capisce come l'antibiotico-resistenza sia uno dei problemi di maggiore attualità per questo genere di studi, visto che il problema risulta comunque molto diffuso.

Gruppi di salmonelle

Indicativamente possiamo distinguere tra 2 gruppi di salmonella, anche se questa distinzione non risulta così netta.

- Salmonelle adatte: sono salmonelle che si adattano ad uno o più specifici tipi di ospite.
- Salmonelle non adatte: sono in grado di infettare diversi ospiti e sono caratterizzate da una particolare capacità di sopravvivere nell'ambiente, tutti questi tipi di salmonelle sono potenzialmente fattori patogeni.

Fenomeno dell'antibiotico resistenza

E' un importante fenomeno di sopravvivenza selettiva nei batteri. Il fattore responsabile di tale resistenza è detto fattore R (che è un elemento citoplasmatico), scoperto da alcuni studiosi giapponesi dopo aver notato ceppi batterici da casi clinici che presentavano caratteristiche di multi-resistenza.

Il primo isolamento avvenne nel 1952 a seguito di uno studio specifico durante un drammatico incremento di infezioni batteriologiche incurabili con gli antibiotici utilizzati usualmente.

Questi studi si rivelarono molto importanti perché permisero di "calibrare" i vari antibiotici, ma si scoprì anche come l'antibiotico resistenza portasse ad una pressione selettiva² sui microrganismi patogeni. In al-

² pressione selettiva: Selective pressure is any phenomena which alters the behavior and fitness of living organisms within a given environment. "E' chiamata pressione selettiva qualsiasi fenomeno che altera il comportamento e l'idoneità di un organismo vivente all'interno di un dato ambiente".(Thomson-Gale,(2005-2006))

tri termini, in presenza dell'antibiotico, alcuni particolari ceppi patogeni possono prevalere sui ceppi commensali.

La pericolosità dell'antibiotico-resistenza singola o multipla è indotta dal cosiddetto fattore R e potenziata dalla proprietà di "trasferimento" non solo nell'ambito della stessa specie o dello stesso genere, ma anche tra diversi generi. Questo fa ben comprendere quale sia l'entità del pericolo di tale caratteristica.

E' da tener presente comunque che i ceppi con antibiotico-resistenza di tipo infettivo sono quasi sempre meno patogeni dei ceppi normali (non multiresistenti) corrispondenti.

L'Antibiogramma

Nella pratica medico-veterinaria, l'esame batteriologico dei campioni patologici viene richiesto con la finalità di isolare i microrganismi patogeni e confermare quindi la diagnosi di malattia infettiva e avere indicazioni precise sul tipo di terapia da instaurare.

Pertanto la determinazione della sensibilità del batterio patogeno isolato ai diversi farmaci antimicrobici rappresenta una delle basi su cui si potrà instaurare una terapia mirata.

Tra le prove in vitro della sensibilità la più usata è l'antibiogramma o tecnica della diffusione in agar da dischetti.

La tecnica utilizzata è quella di Kirby-Bauer.

Si basa sulla deposizione di un certo numero di dischetti di cellulosa, impregnati di quantità note di farmaci antibatterici, in una piastra di petri contenente un adatto terreno colturale solido, opportunamente insemato con il germe patogeno in esame.

Durante il periodo di incubazione delle piastre, i chemioantibiotici si diffonderanno dai dischetti nel terreno circostante e, se efficaci, inibiranno la replicazione batterica in un'area tanto più grande quanto maggiore sarà la loro attività.

Si osserverà così la comparsa di aloni di inibizione di crescita attorno al dischetto antibiotato, il cui diametro sarà proporzionale all'attività antibatterica dell'antibiotico contenuto.

L'assenza di tale alone indicherà l'inefficacia del farmaco.

In base ai risultati dell'antibiogramma il batterio verrà definito

- Sensibile
- Intermedio
- Resistente

Antibiotici usati nell'analisi

Nella tabella sono riportati gli antibiotici e la loro classificazione, rispetto alla grandezza dell'alone riscontrato nell'analisi dell'antibiogramma.

ANTIBIOTICO	Resistente	Intermedio	Sensibile
AMC -Amoxicillina	≤13	14-16	≥17
AMP -Ampicillina	≤13	14-17	≥18
CTX -Cefotaxime	≤14	15-22	≥23
CF -Cefalotina	≤14	15-17	≥18
C -Cloramfenicolo	≤12	13-17	≥18
CL -Colistina	≤8	9-10	≥11
GM -Gentamicina	≤12	13-14	≥15
K -Kanamicina	≤14	14-17	≥15
NA -Acido nalidixico	≤13	14-18	≥19
CAZ -Ceftazidime	≤12	13-14	≥15
S -Streptomina	≤11	12-14	≥15
TE -Tetraciclina	≤14	15-18	≥19
SXT -Sulfametolo/Trimetoprim	≤10	11-15	≥16
S3 -Trisulfamidico	≤12	13-16	≥17
CIP -Ciproflaxin	≤15	16-20	≥21
ENR -Enrofloxacin	≤17	18-21	≥22

Analisi tramite tecniche di data mining

Definizione di data mining

Il data mining rappresenta l'attività di elaborazione in forma grafica o numerica di grandi raccolte di dati o di flussi continui di dati con lo scopo di estrarre informazione utile a chi detiene i dati stessi.

Il data mining è una disciplina relativamente 'giovane' che si colloca nel punto di intersezione fra la statistica, l'intelligenza artificiale e la gestione dei data base.

La connessione con il mondo dei data base risulta implicita nel fatto che la gestione di grosse moli di dati provenienti da basi di dati che possono essere anche di tipo 'distribuito' impone competenze per la gestione, l'estrapolazione e l'eventuale pulizia dei dati.

L'intelligenza artificiale risulta cruciale nel momento successivo all'acquisizione dei dati, molte tecniche derivanti dal machine learning, infatti, possono risultare utili nella ricerca delle 'leggi' che regolano il fenomeno osservato sulla base dei dati.

Data Mining, gli oggetti di interesse e i pericoli

Nel campo del data mining l'elemento di interesse dell'indagine è spesso molto 'labile'.

Tipicamente infatti ci si trova ad affrontare 2 situazioni:

- la ricerca e la determinazione, tramite i dati disponibili, di un modello globale per poter spiegare il fenomeno osservato;
- la determinazione di "configurazioni speciali" nell'andamento dei dati.

I dati però, non derivando necessariamente da un piano di campionamento o sperimentale, possono mancare delle condizioni canoniche per una corretta raccolta e interpretazione. Questa implicazione iniziale comporta sicuramente, in questo ambito, maggiori difficoltà interpretative e impone maggiore attenzione durante tutto il lavoro di analisi e soprattutto nel momento della chiusura e quindi delle conclusioni a cui si perviene tramite lo studio.

I dati in studio tipicamente, infatti, non derivano da un campione casuale e quindi le conclusioni devono rifarsi al solo insieme sotto studio (dando luogo a conclusioni essenzialmente ristrette ai casi osservati) e pertanto non possono ritenersi estendibili a una generica popolazione di riferimento.

A differenza delle ricerche tipiche in ambito clinico, qui la specificazione dell'oggetto di interesse non avviene a priori, infatti questo è uno dei punti di estrema differenza fra l'ambito del data mining e altri ambiti di indagine.

Un particolare caso è quello delle variabili di tipo leaker, che sono surrogati della variabile di interesse, per esempio se la variabile di interesse è il valore della bolletta telefonica, il numero delle telefonate sarà "quasi" sicuramente una variabile di tipo leaker.

Da valutare quindi sono la comprensione del fenomeno in questione, la comprensione degli strumenti utilizzati (tipicamente informatici e matematici) e la comprensione sintetica dei risultati che questi modelli generano.

Scelta degli strumenti per l'analisi del problema

Ci troviamo di fronte alla possibilità di costruire modelli che ci aiutino a comprendere meglio la situazione in oggetto guardandola fondamentalmente da varie angolazioni.

Tra gli strumenti ritenuti adatti alla descrizione del problema sono stati individuati:

- la regressione lineare applicata a modelli di classificazione;
- i tree-models applicati alla classificazione, ovvero alberi di classificazione e foreste casuali;
- l'analisi discriminante lineare;
- la regressione logistica.

La regressione lineare applicata a modelli di classificazione

Partendo dal caso di 2 categorie. Possiamo istituire uno schema di regressione lineare in cui la variabile risposta y è in questo caso dicotomica e assume i valori 0 oppure 1.

Possiamo usare quindi il valore

$$\hat{y} = \frac{1}{2}$$

come soglia di discriminazione per la previsione delle due categorie, nel senso che i soggetti verranno allocati ai gruppi, in base al superamento o meno di tale soglia.

La regressione lineare in una delle sue forme più semplici, facendo riferimento al solo piano bidimensionale, può essere espressa come:

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \epsilon,$$

la natura dell'errore è peraltro singolare visto che si riduce ad un errore di tipo dicotomico e implica la necessità di inserire l'intercetta nel modello per arrivare a conclusioni adeguate, infatti l'unico assunto realmente importante affinché il criterio dei minimi quadrati fornisca risposte sensate è che

$$E[\epsilon] = 0,$$

ma questo risulta praticamente soddisfatto nel caso l'intercetta faccia parte del modello, il valore β_0 infatti ingloba l'eventuale valore non nullo.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 z_1 + \hat{\beta}_2 z_2 = \frac{1}{2}$$

Questa retta, che in conclusione eguaglia il valore di soglia, risulta essere la stima dei minimi quadrati per cui (in questo caso il piano bidimensionale) viene diviso in 2 porzioni e le eventuali previsioni assegnate in base al loro 'peso' rispetto alla soglia a uno dei due gruppi. In questo caso specifico si allocherebbe il soggetto al gruppo 1 se il corrispondente \hat{y} superasse $\frac{1}{2}$, al gruppo 0 altrimenti.

Naturalmente la generalizzazione del metodo si estende anche a più di $K = 2$ categorie, con $K_n = N$ categorie e inserendo nel predittore funzioni non lineari delle z_i .

Analisi discriminante

L'impostazione propria del problema di classificazione è quella che tipicamente riguarda i modelli di analisi discriminante.

Qui infatti ci si riferisce ad una variabile casuale p-dimensionale X ed una variabile casuale categoriale y che rappresenta la classe a cui appartiene un soggetto.

La popolazione complessiva è costituita di K sub-popolazioni (le modalità della variabile categoriale), aventi la rispettiva funzione di densità di probabilità uguale a

$$p_1(x), p_2(x), \dots, p_K(x)$$

per la distribuzione di X , e con peso

$$\pi_1, \pi_2, \dots, \pi_K,$$

rispetto al totale della popolazione, tenendo conto che

$$\sum_{k=1}^K \pi_k = 1.$$

Da cui la densità complessiva per la popolazione è

$$p(x) = \sum_{k=1}^K \pi_k p_k(x).$$

A priori quindi la probabilità che un soggetto appartenga alla k -esima popolazione è data da π_k .

Se per quel soggetto è noto il valore assunto da X allora per il teorema di Bayes la probabilità a posteriori che quel soggetto appartenga al gruppo k è data da

$$\{Py = k | X = x_0\} = \frac{\pi_k p_k(x_0)}{p(x_0)}.$$

Equivalentemente il confronto tra 2 classi avviene sulla base dell'operatore (sottoposto a funzione logaritmo)

$$\lg \frac{\{Py=k|X=x_0\}}{\{Py=m|X=x_0\}} = \lg \frac{\pi_k}{\pi_m} + \lg \frac{p_k(x_0)}{p_m(x_0)}$$

Quindi ci si può riferire ad ogni classe, per il confronto, tramite la funzione discriminante

$$d_k(x_0) = \lg \pi_k + \lg p_k(x_0).$$

Da ciò si intuisce che il valore di k che massimizza la funzione discriminante individua il gruppo a cui attribuiamo il nuovo soggetto.

Per le stime che rendono effettivamente operativo l'impianto teorico abbiamo una strada naturale per i π_k che possono essere visti come la frazione di popolazione rispetto al totale, mentre per i $p_k(x)$ dobbiamo decidere se utilizzare approcci di tipo parametrico o meno.

Nell'analisi discriminante lineare si utilizza l'ipotesi più semplice, quella secondo la quale ogni densità $p_k(x)$ è normale multipla con parametri dipendenti dalla popolazione di riferimento k , diciamo

$$N_p(\mu_k, \Sigma_k),$$

tale per cui risulta

$$p_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}} \det(\Sigma_k)^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right\}$$

Nell'analisi discriminante lineare, semplificando, assumendo che tutte le matrici di varianza siano uguali ad una stessa matrice Σ , la funzione discriminante prende la forma,

$$d_k(x) = \lg \pi_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + x^T \Sigma^{-1} \mu_k.$$

Espressione che, essendo una funzione lineare, dà il nome all'analisi.

La stima dei parametri è

$$\hat{\mu} = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\Sigma} = \frac{1}{n-K} \sum_{i:y_i=k} \sum_{k=1} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

Peraltro si può pervenire alla funzione discriminante lineare indicata sopra senza ricorrere all'ipotesi di normalità multipla, appoggiandosi solo ad ipotesi del secondo ordine. Questo giustifica l'utilizzo della tecnica anche quando X non è una variabile normale multipla e anzi può avere componenti non continue.

Alberi di classificazione

Il modo più intuitivo e semplice di approssimare una funzione è quello che implica l'utilizzo di una funzione a "gradini", una funzione cioè costante a tratti su intervalli.

D'altronde si rende necessario valutare

- quali intervalli considerare
- quante suddivisioni utilizzare e quali punti scegliere
- quale valore di ordinata assegnare ad ogni intervallo

In ambito di regressione con variabili continue e non categoriali risulta immediato porre

$$\frac{\int_{R_j} f(x) dx}{|R_j|}$$

Dove il termine R_j rappresenta lo j -esimo intervallo e lo stesso all'interno del modulo rappresenta la lunghezza dell'intervallo. Per la scelta dei punti di suddivisione conviene scegliere punti più vicini laddove la funzione è più ripida, mentre il numero di punti dipende dalla bontà dell'approssimazione che viene richiesta alla funzione.

Per approssimare una certa funzione di regressione $f(x)$ non nota, utilizzeremo quindi le n osservazioni 'campionarie' generate da un generico modello che si assume essere generatore dei dati.

Nel caso di variabili risposta di tipo categoriale o dicotomico dobbiamo utilizzare delle strutture in grado di poter adattare ai nostri dati una funzione di classificazione.

In linea teorica, questa funzione dovrebbe stimare attraverso una procedura sub-ottimale, dato l'alto costo computazionale, elementi utili alla rappresentazione di una funzione a gradini, quali le grandezze degli intervalli, il numero di gradini e il valore della funzione in un dato intervallo.

Nel caso categoriale vi è quindi necessità di trovare una funzione che possa essere utilizzata a tale scopo.

Se indichiamo con 0 e 1 le due classi e con

$$p(x) = P \{1|x\}$$

la probabilità che un individuo con x caratteristiche (dove x si riferisce al vettore delle specifiche dell'individuo, quindi il vettore che contiene le covariate) appartenga alla classe 1, si può approssimare $p(x)$ tramite una funzione a gradini del tipo:

$$\hat{p}(x) = \sum_{j=1}^J P_j I(x \in R_j)$$

Con P_j che rappresenta la probabilità che una osservazione appartenga all'intervallo moltiplicato per la funzione indicatrice che determina o meno l'appartenenza di x a tale intervallo.

Essendo qui le variabili di tipo dicotomico risulta immediato utilizzare la media aritmetica per stimare i P_j che quindi diverranno le frequenze relative di 1 riferite alla regione R_j .

Ora, operativamente si lavora attraverso una ottimizzazione passo-passo che genera una sequenza di stime sempre più raffinate, e ad

ogni passaggio si minimizza la devianza relativamente al passaggio dall'approssimazione corrente a quella successiva.

Questo procedimento viene applicato iterativamente fino ad arrivare, almeno in linea teorica, a costruire un albero con n foglie.

Questo tipo di strategia però non risulta molto utile, poiché non da informazione sintetica rispetto alla struttura effettiva dei dati.

Nasce quindi il bisogno di utilizzare una potatura per rendere di qualche utilità il modello generato.

Per fare questo, però, dobbiamo istituire una forma di devianza per rendere possibile l'utilizzo di una qualche funzione di penalizzazione.

Per la devianza si può utilizzare la devianza connessa alla distribuzione binomiale.

$$D = -2 \sum_{i=1}^n \{y_i \lg \hat{p}_i + (1 - y_i) \lg(1 - \hat{p}_i)\},$$

Che accorpendo gli elementi appartenenti alla stessa regione R_j , dove la probabilità vale costantemente P_j , risulta

$$D = \sum_{j=1}^J -2n_j [\hat{P}_j \lg \hat{P}_j + (1 - \hat{P}_j) \lg(1 - \hat{P}_j)] = \sum_j D_j$$

Quindi si può riscrivere la devianza come

$$D = 2n \sum_j \frac{n_j}{n} Q(\hat{P}_j),$$

che, a meno della costante $2n$, è una media delle entropie pesate con la numerosità delle foglie, $Q()$ infatti può essere visto come

$$- \sum_{k=0,1} P_{jk} \lg P_{jk},$$

quindi come un indice di impurità e può essere sostituito con altri indici, come quello di Gini.

Per la potatura quindi si introduce tipicamente una funzione obiettivo che incorpora una penalizzazione per il costo-complessità dell'albero, cioè per la dimensione J dell'albero.

$$C_\alpha(J) = \sum_{j=1}^J D_j + \alpha J,$$

con α parametro di penalizzazione. Si seleziona quindi l'albero che minimizza la funzione C_α .

Quando la classificazione avviene attraverso metodi più instabili, come ad esempio alberi o reti neurali, è fortemente influenzata dalla scelta specifica dell'insieme di dati usato per la stima. Se tale insieme viene modificato di poco, si può ottenere un modello completamente diverso dall'originale, con circa lo stesso errore di previsione.

Per ottenere una migliore capacità previsiva dal modello, una possibilità è quella di combinare le previsioni ottenute da metodi diversi. Una possibilità per ottenere combinazioni di modelli consiste nel considerare per la previsione ad ogni interazione diversi sottoinsiemi delle variabili esplicative, ottenendo così delle stime da combinare.

Una strategia di questo tipo è stata proposta utilizzando come classificatori originali gli alberi e scegliendo le variabili da inserire in ciascun modello attraverso selezione casuale; tale procedura ha perciò preso il nome di Foresta Casuale.

La procedura consiste nel selezionare in modo casuale, ad ogni nodo di un albero, un piccolo gruppo di variabili esplicative che verranno ispezionate per trovare il punto di suddivisione ottimale, secondo il criterio di crescita utilizzato. Seguendo questo metodo, quindi, per far crescere l'albero, anziché esplorare tutti i possibili punti di suddivisione in ciascun nodo, vengono esplorate solo F variabili scelte a caso.

Quando l'albero raggiunge la sua massima dimensione però non viene potato, sarà infatti l'operazione di combinazione dei diversi alberi che permetterà di evitare i problemi di sovra-adattamento.

Il numero di variabili da selezionare in ciascun nodo è un parametro di regolazione da determinare e generalmente viene mantenuto costante su tutti i nodi. Spesso viene scelto considerando foreste costruite con valori diversi di F e determinando quel valore che minimizza l'errore su un insieme di verifica.

L'altro parametro di regolazione da determinare, chiamato B , rappresenta il numero di alberi che costituiscono la foresta. Si dimostra che l'errore globale converge a una soglia inferiore al crescere di B e che non si presentano problemi di sovra-adattamento quando vengono aggiunti ulteriori alberi. Se si utilizza, assieme alla selezione randomizzata delle variabili anche l'uso del bagging è possibile stimare delle misure di importanza per le variabili esplicative. Si può infatti utilizzare l'errore di previsione ottenuto dai dati out-of-bag (non utilizzati per la stima) per scegliere il parametro di regolazione. Una misura dell'importanza di ciascuna variabile esplicativa nella previsione della risposta si può procedere nel modo seguente. Dopo aver costruito ogni albero, si effettua la previsione sull'insieme di

dati out-of-bag e sullo stesso insieme con i valori della j-esima variabile permutati casualmente.

Si misura quindi la differenza tra l'errore di previsione nei due casi e, al termine della procedura, si considera la media delle differenze tra i vari alberi divisa per l'errore standard. Tale indicatore fornisce una misura di quanto la variabile influisce sulle previsioni.

La regressione logistica

Nei modelli lineari generalizzati i modelli binomiali vengono utilizzati quando si hanno dati che esprimono il numero di "successi" rispetto al totale di prove effettuate.

In questo caso il modello generale rientra nella classe dei glm con

$$Z_i \sim \text{Bin}(x_i, \pi_i),$$

dove il generico Z_i è l'osservazione i-esima.

Obiettivo del modello è studiare la dipendenza delle variabili risposta rispetto al valore delle variabili di regressione. Questo avviene cercando di trovare il valore della media condizionale degli Z_i data la matrice delle variabili indipendenti X .

Questo valore si esprime tramite la formula $E(Z|x)$.

In questa particolare forma legata ai glm appare più appropriato utilizzare come variabile risposta la proporzione di successi $Y_i = Z_i/m_i$ e quindi utilizzare $E(Y_i) = \mu_i = \pi_i$ come probabilità all'interno del modello binomiale.

Quindi μ_i è un parametro compreso tra $[0,1]$. Una scelta naturale per la funzione legame è una funzione che rispetti questo vincolo, una reale possibilità è assegnarle una forma lineare nei parametri

$$\eta_i = \beta_1 x_i + \beta_2 x_i + \dots + \beta_n x_i$$

collegata a μ_i tramite una funzione di ripartizione Ψ tale che

$$\mu_i = \Psi(\eta_i).$$

Da cui la funzione link del modello diventa

$$g(\mu) = \Psi^{-1}(\mu).$$

Tra le varie scelte possibili per la funzione di ripartizione possiamo ricordare:

- probit dove la funzione di ripartizione è quella di una normale standard e $g(\mu) = \Phi^{-1}(\mu)$, funzione link, non ha una forma analitica;
- c-log-log dove $\Psi(\eta_i) = 1 - e^{-e^\eta}$ è la funzione di ripartizione e la funzione link risulta essere

$$g(\mu) = \Psi^{-1}(\mu) = \log(-\log(1 - \mu)).$$

La funzione link che useremo noi invece sarà la funzione logit, che viene così definita:

$$\Psi(\eta_i) = \frac{e^\eta}{1+e^\eta}.$$

Il modello binomiale con link logit prende il nome di "*proportional odds*", in questo caso infatti i parametri lineari η sono collegati a μ tramite la funzione link ovvero la sua inversa

$$\eta = \log \frac{\mu}{1-\mu} = \Psi^{-1}(\mu).$$

Accenno agli errori nella regressione logistica

Nella regressione lineare si assume che ogni osservazione possa essere espressa come

$$y = E(Y|x) + \epsilon,$$

dove ϵ rappresenta il termine d'errore, tipicamente viene definito come la deviazione di un'osservazione dalla media condizionale.

L'assunzione più comune è quella che ϵ si distribuisca come una normale di media 0 e varianza σ^2 considerata costante rispetto alla variabile indipendente (o di regressione).

Da questo segue che

$$Y \sim N_n(E(Y|x), \sigma^2).$$

Questo non si può adattare al caso della regressione logistica, infatti nella regressione di tipo logistico la nostra variabile risposta y data x può essere espressa come $y = \pi(x) + \epsilon$, qui la quantità ϵ può assumere due soli valori

se $y = 1$ allora ϵ assumerà valore $\epsilon = 1 - \pi(x)$ con probabilità $\pi(x)$.

se $y = 0$ allora ϵ assumerà valore $\epsilon = -\pi(x)$ con probabilità $1 - \pi(x)$,

quindi, poichè ϵ ha una distribuzione con media 0 e varianza uguale a $\pi(x)(1 - \pi(x))$ la variabile risposta seguirà una binomiale con la probabilità data dalla media condizionale $\pi(x)$.

Adeguare il modello

Iniziamo introducendo la formula per la log-verosimiglianza di un glm, tale formula è data da:

$$\ell(\beta) = \sum_{i=1}^n \ell_i(\beta),$$

con

$$\ell_i(\beta) = \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi).$$

Una volta introdotta la stima di massima verosimiglianza possiamo introdurre il concetto di modello "massimale", tale modello avrà le seguenti caratteristiche:

- ha funzione di distribuzione uguale a quella del modello corrente;
- ha un numero parametri pari a n ;
- ha la medesima funzione link del modello corrente;

Il confronto dovrà quindi avvenire fra il modello saturo e il modello ridotto

$$D = -2 \ln \left[\frac{\text{verosimiglianza del modello ridotto}}{\text{verosimiglianza del modello saturo}} \right],$$

da cui con semplici passaggi, una volta definiti per convenzione $\hat{\theta}$ come la massima verosimiglianza per il modello ridotto e $\tilde{\theta}$ come il modello massimale nella sua stima di massima verosimiglianza, si arriva alla seguente definizione:

$$D(y; \hat{\theta}) = 2\phi[\ell(\tilde{\theta}) - \ell(\hat{\theta})] = \phi \sum_{i=1}^n D_i,$$

con

$$D_i = 2[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]$$

e

$$\frac{D(y; \hat{\theta})}{\phi} = \sum_{i=1}^n D_i$$

detta devianza scalata e che risulta sempre positiva.

Nei glm di tipo binomiale con funzione di link logit la devianza può essere espressa come:

$$D(y; \hat{\theta}) = 2 \sum_{i=1}^n o_i \log \frac{o_i}{e_i},$$

dove gli o_i sono le frequenze osservate e gli e_i sono le frequenze attese.

La differenza tra modello saturo e quello in esame data dalla devianza rappresenta una misura della diminuzione della bontà di adattamento dovuta al passaggio dal modello saturo a quello corrente con $p < n$ variabili esplicative. Quindi il modello saturo non risulta essere un modello d'interesse pratico, non sintetizzando i dati, ma costituisce un valore di riferimento per il confronto con il modello corrente.

Per controllare la significatività nel modello di una variabile indipendente compariamo i valori della devianza tra il modello contenente la variabile e il modello ridotto, che non contiene la variabile in questione.

Chiamiamo tale confronto

$$G = -2 \ln \left[\frac{\text{(modellononcontenentelavariabile)}}{\text{(modellocontenentelavariabile)}} \right]$$

che equivale a:

$$2[\ell(\hat{\beta}) - \ell(\hat{\beta}_{mr})].$$

Quindi

$$\frac{D(Y; \hat{\theta}_{mr}) - D(Y; \hat{\theta})}{\phi},$$

che per $n \rightarrow \infty$ si distribuisce come un $\chi_{p-p_0}^2$ sotto H_0 .

Dove

H_0 = l'informazione apportata dalla variabile aggiuntiva è statisticamente poco rilevante

H_1 = l'informazione apportata dalla variabile aggiuntiva è statisticamente rilevante

Da ciò si passa al confronto con la distribuzione nulla e si rifiuta l'ipotesi nulla sulla base di p-value troppo elevati.

Nel caso in cui il parametro di dispersione non fosse noto, si sostituisce a questo la sua stima consistente $\hat{\phi} = \frac{D(Y; \hat{\theta})}{n-p}$ e si procede allo stesso modo.

Analisi descrittiva per i dati in esame

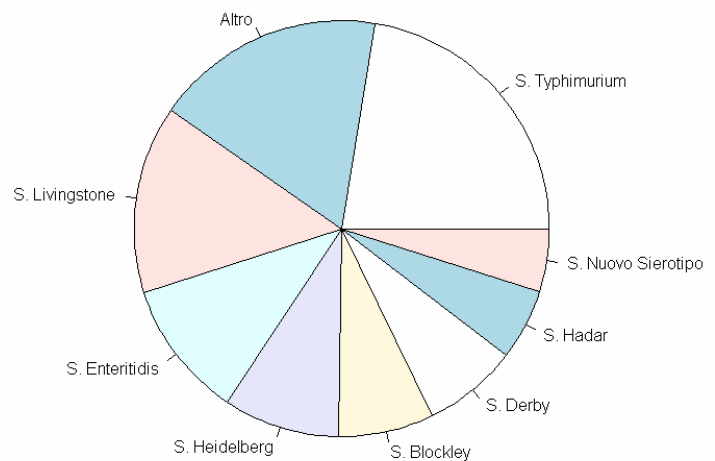
Verranno inseriti in questa sezione alcuni grafici e alcune tabelle ottenute da una prima indagine sui dati in esame.

Verranno effettuate alcune indagini sulle resistenze e sui soggetti campionati: animali, sierotipi e luoghi di prelievo.

Sierotipi

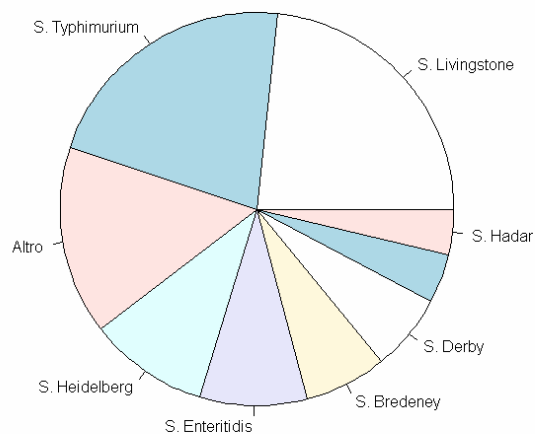
Anno 2005

S. Typhimurium	0,178
Altro	0,1424
S. Livingstone	0,1163
S. Enteritidis	0,0851
S. Heidelberg	0,072
S. Blockley	0,059
S. Derby	0,059
S. Hadar	0,0434
S. Nuovo Sierotipo	0,0391
	0,0356
S. Tennessee	0,0217
S. Agona	0,0208
S. London	0,0208
S. Bredeney	0,0182
S. Infantis	0,0156
S. Thompson	0,0156
S. Anatum	0,0139
S. Saintpaul	0,013
S. Virchow	0,0122
S. Mbandaka	0,0104
S. Rissen	0,0078



Anno 2006

S. Livingstone	0.1849
S. Typhimurium	0.1712
Altro	0.1225
S. Heidelberg	0.0784
S. Enteritidis	0.07
S. Bredeney	0.0533
S. Derby	0.0502
	0.032
S. Hadar	0.0297
S. Blockley	0.0244
Gruppo B	0.0236
S. Saintpaul	0.0236
S. Anatum	0.0198
S. Virchow	0.0198
S. Mbandaka	0.019
S. Thompson	0.019
S. Agona	0.0167
S. Rissen	0.0167
S. London	0.0137
S. Infantis	0.0099
S. Tennessee	0.0015



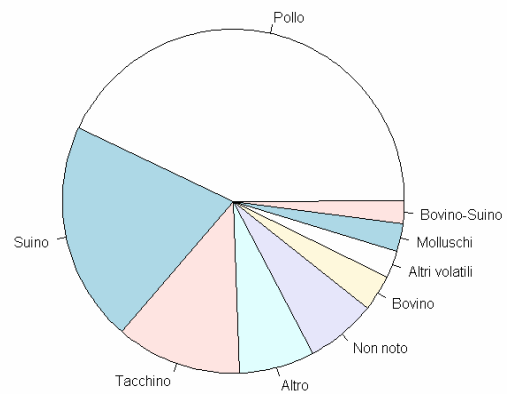
Dove le immagini si riferiscono alla distribuzione "pura", senza altre osservazioni dei primi 9 sierotipi in classifica per entrambi gli anni.

I sierotipi S. Typhimurium e S. Livingstone sono i maggiormente presenti nello studio.

Specie animali

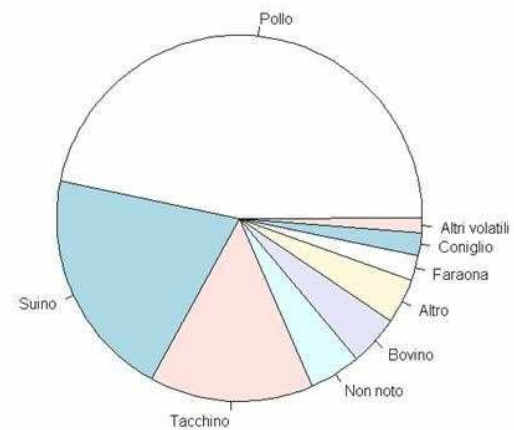
Anno 2005

Pollo	0,4201
Suino	0,2023
Tacchino	0,1155
Altro	0,0694
Non noto	0,0651
Bovino	0,033
Altri volatili	0,0252
Molluschi	0,0252
Bovino-Suino	0,0208
Coniglio	0,0139
Faraona	0,0095



Anno 2006

Pollo	0,4612
Suino	0,2002
Tacchino	0,1423
Non noto	0,0457
Bovino	0,0434
Altro	0,0396
Faraona	0,0221
Coniglio	0,019
Altri volatili	0,0129
Bovino-Suino	0,0091
Molluschi	0,0046



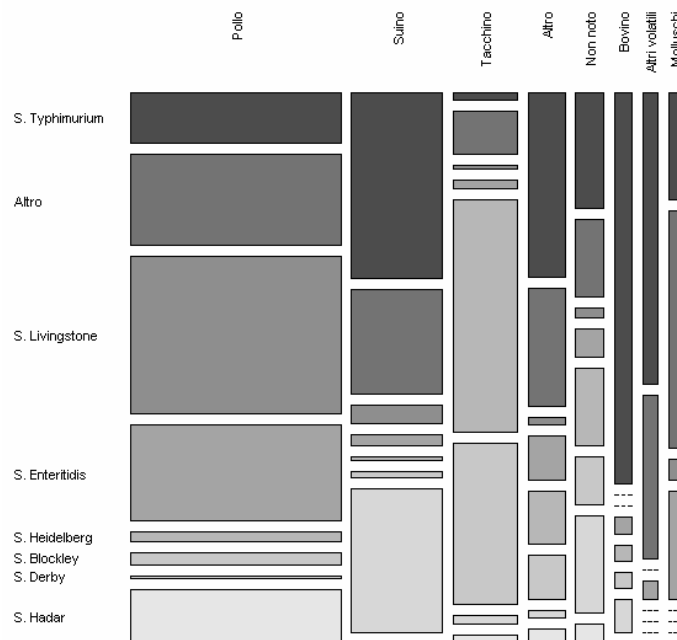
Da questo notiamo come le specie avicole siano le più numerose, assieme a suini e bovini.

Specie Animali e Sierotipi

Anno 2005

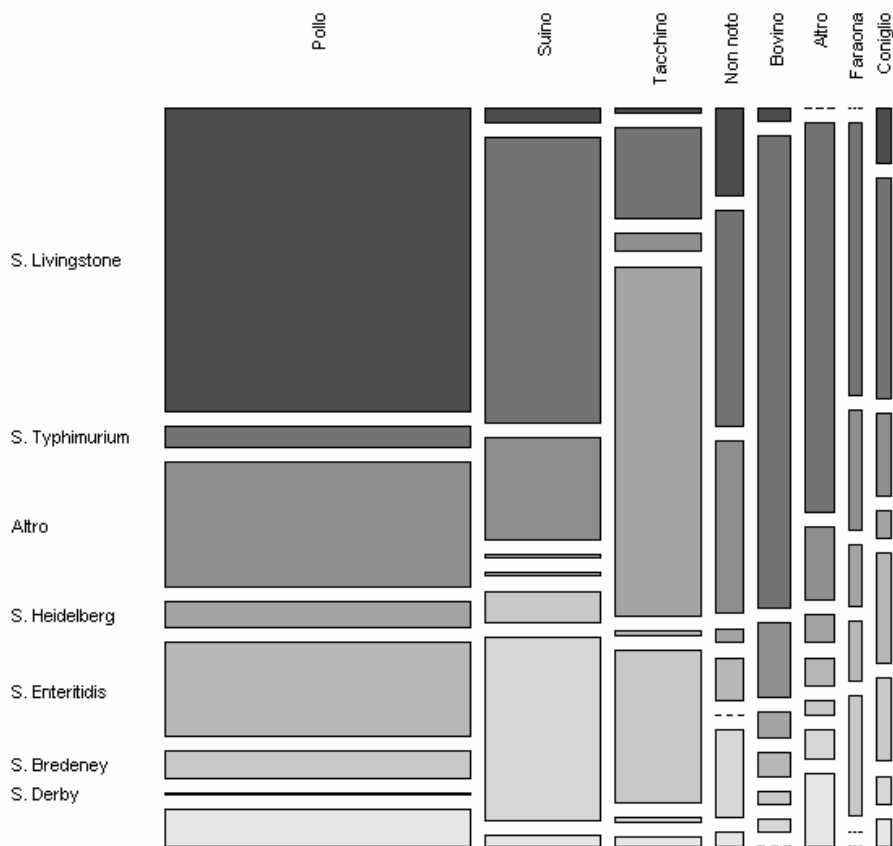
	S. Typhimurium	Altro	S. Livingstone	S. Enteritidis	S. Heidelberg	S. Blockley	S. Derby	S. Hadar	S. Nuovo Sierotipo
Pollo	0,035	0,061	0,107	0,064	0,007	0,008	0,002	0,036	0,003
Suino	0,054	0,030	0,005	0,003	0,001	0,002	0,042	0,000	0,018
Tacchino	0,002	0,009	0,001	0,002	0,048	0,033	0,002	0,002	0,003
Altro	0,022	0,014	0,001	0,005	0,006	0,005	0,001	0,002	0,000
Non noto	0,010	0,007	0,001	0,003	0,007	0,004	0,009	0,002	0,007
Bovino	0,021	0,000	0,000	0,001	0,001	0,001	0,002	0,000	0,004
Altri volatili	0,014	0,008	0,000	0,001	0,000	0,000	0,000	0,000	0,001
Molluschi	0,004	0,010	0,001	0,004	0,000	0,000	0,000	0,000	0,000
Bovino-Suino	0,008	0,002	0,001	0,001	0,002	0,003	0,002	0,000	0,001
Coniglio	0,009	0,000	0,000	0,001	0,000	0,003	0,001	0,000	0,000
Faraona	0,000	0,003	0,000	0,000	0,001	0,001	0,000	0,002	0,001

Non sono state inserite tutte le osservazioni, ci siamo soffermati alle prime 9, per cui è stato anche prodotto un grafico, le osservazioni sono state classificate in ordine decrescente rispetto alle distribuzioni marginali. Quindi il sierotipo maggiormente riscontrato nel 2005 è risultato il sierotipo S. Typhimurium e la specie che ha maggiore frequenza è il Pollo.



Anno 2006

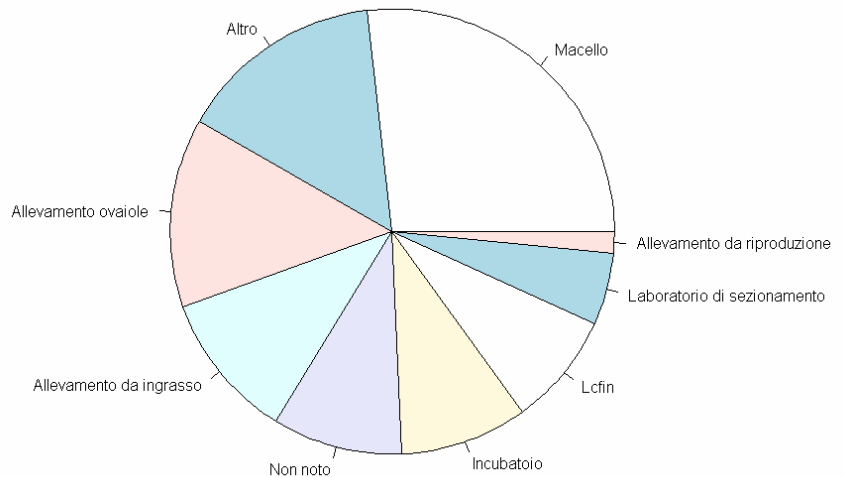
	S. Livingstone	S. Typhimurium	Altro	S. Heidelberg	S. Enteritidis	S. Bredeney	S. Derby	S. Hadar
Pollo	0,174	0,012	0,072	0,015	0,054	0,015	0,001	0,021
Suino	0,003	0,062	0,022	0,001	0,001	0,007	0,040	0,002
Tacchino	0,001	0,014	0,003	0,056	0,001	0,024	0,001	0,002
Non noto	0,005	0,011	0,009	0,001	0,002	0,000	0,005	0,001
Bovino	0,001	0,029	0,005	0,002	0,002	0,001	0,001	0,000
Altro	0,000	0,021	0,004	0,002	0,002	0,001	0,002	0,004
Faraona	0,000	0,007	0,003	0,002	0,002	0,003	0,000	0,000
Coniglio	0,002	0,006	0,002	0,001	0,003	0,002	0,001	0,001
Altri volatili	0,000	0,005	0,003	0,000	0,002	0,000	0,001	0,001
Bovino- Suino	0,000	0,004	0,000	0,000	0,001	0,000	0,000	0,000
Molluschi	0,000	0,001	0,000	0,000	0,002	0,000	0,001	0,001



Luogo Prelievo

Anno 2005

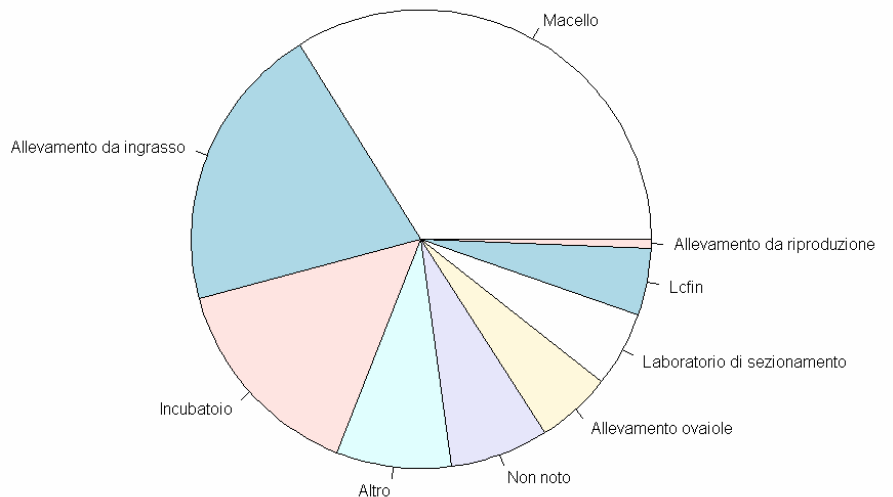
Macello	0,268
Altro	0,149
Allevamento ovaiole	0,137
Allevamento da ingrasso	0,108
Non noto	0,095
Incubatoio	0,093
Lcfin	0,082
Laboratorio di sezionamento	0,052
Allevamento da riproduzione	0,016



Dove Lcfin è l'abbreviazione di luogo_consumatore_finale che accorpa le modalità ristorazione collettiva e punto vendita al dettaglio. Questo accorpamento è stato effettuato per poter valutare l'esistenza nel punto finale della filiera alimentare di salmonelle.

Anno 2006

Macello	0,3387
Allevamento da ingrasso	0,2032
Incubatoio	0,1484
Altro	0,0807
Non noto	0,0693
Allevamento ovaiole	0,0533
Laboratorio di sezionamento	0,0533
Lcfin	0,0472
Allevamento da riproduzione	0,0061



Resistenze e sierotipi

Anno 2005

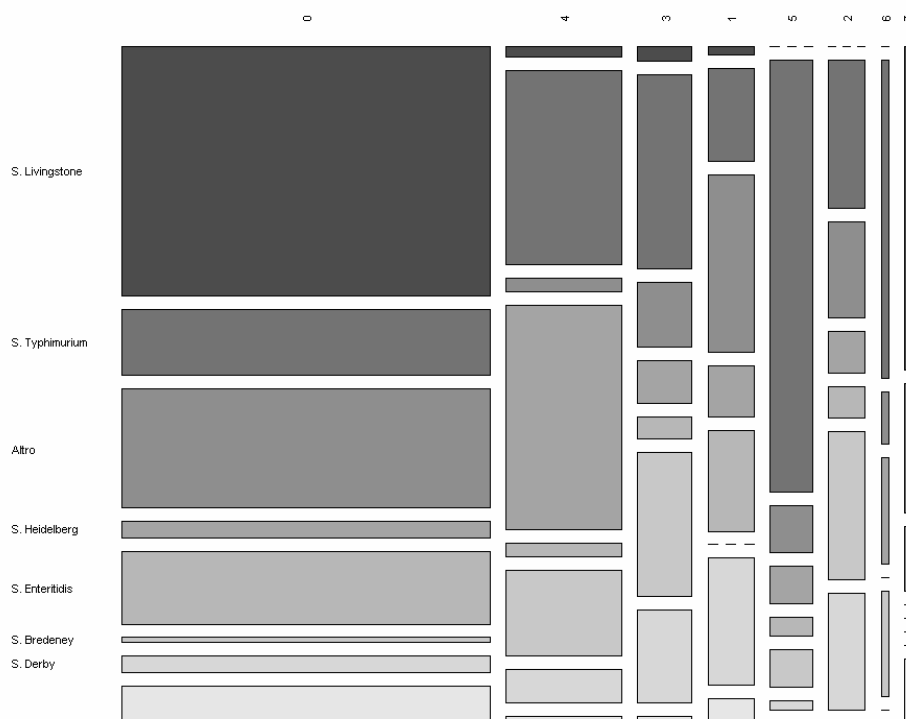
	S. Typhimurium	Altro	S. Livingstone	S. Enteritidis	S. Heidelberg	S. Blockley	S. Derby	S. Hadar	S. Nuovo Sierotipo
0	0,05122	0,08594	0,10851	0,06076	0	0	0,01562	0	0
4	0,03212	0,00434	0,00087	0,00087	0,05729	0,03646	0,00608	0,01562	0,01302
3	0,02257	0,00955	0,00087	0,00087	0,00347	0,01823	0,01823	0,00694	0,00347
1	0,00781	0,01736	0,00174	0,02083	0,00087	0	0,01389	0	0,00608
5	0,02083	0,00694	0,00087	0	0,00521	0,0026	0,00174	0,01215	0,00868
2	0,00694	0,01389	0,00087	0,00174	0,00174	0,00087	0,00174	0,00781	0,00087
6	0,02517	0,00434	0	0	0,0026	0,00087	0,00087	0,00087	0,00174
7	0,01042	0	0,00174	0	0,00087	0	0	0	0,00521
8	0	0	0,00087	0	0	0	0	0	0
9	0,00087	0	0	0	0	0	0,00087	0	0



Anno 2006

	S. Livingstone	S. Typhimurium	Altro	S. Heidelberg	S. Enteritidis	S. Bredeney	S. Derby	S. Hadar	
0	0,17656	0,04718	0,08447	0,01142	0,05175	0,00304	0,01218	0,02664	0
4	0,00228	0,04338	0,00304	0,05023	0,00304	0,01903	0,00761	0,00152	0,00761
3	0,00152	0,02055	0,00685	0,00457	0,00228	0,01522	0,00989	0,00076	0,00381
1	0,00076	0,00837	0,01598	0,00457	0,00913	0	0,01142	0,00228	0
5	0	0,03501	0,00381	0,00304	0,00152	0,00304	0,00076	0	0,01218
2	0	0,01065	0,00685	0,00304	0,00228	0,01065	0,00837	0	0,00152
6	0	0,00457	0,00076	0,00152	0	0,00152	0	0	0,00381
7	0,00381	0,00152	0,00076	0	0	0	0	0,00076	0,00076
8	0	0	0	0	0	0,00076	0	0	0
9	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0

Una cosa interessante da notare è che le salmonelle, prevalentemente o non presentano resistenze o presentano resistenza a 4 o 3 antibiotici.



Dai dati del 2006, notiamo quindi come la maggior parte delle salmonelle, campionate durante l'anno, siano riconducibili al sierotipo S. Livingstone e non multiresistenti, anzi completamente sensibili agli antibiotici.

Analisi dei dati tramite tecniche di data mining

Veniamo ora alla determinazione dei modelli che utilizzeremo per lo studio, per iniziare verrà stilata una breve descrizione delle variabili che entreranno a fare parte dei modelli e verrà costruita una funzione con le sole covariate che troveremo interessanti e utili ai fini di una buona sintesi dei dati. Verranno eliminate anche variabili considerate di tipo leaker. In caso venissero riscontrate variabili che presentano micronumerosità o assumono sempre lo stesso valore probabilmente dovranno essere scartate o accorpate durante il lavoro.

Finita questa prima fase costruiremo una "funzione" di regressione con le variabili di interesse che verrà utilizzata dai vari modelli e eventualmente ridotta nel numero di variabili o arricchita di interazioni successivamente, nella fase di adattamento del modello stesso.

Variabili

ID1	Identificativo univoco
ID	Identificativo da laboratorio
LABORATORIO_ORIGINE	laboratorio che ha condotto l'indagine sul campione
PROVINCIA_LABORATORIO	provincia del laboratorio
TIPO_CAMPIONE	tipologia del campione: es. Animale, Ambientale...
CSPEC	Sottotipologia del campione, vuoto per tutte le osservazioni
COMUNE_PRELIEVO	Comune dove è stato fatto il campionamento
PROVINCIA_PRELIEVO	Provincia dove viene fatto il prelievo
ASL	ASL di competenza
COD_ORIGINE	Codice di origine
DATA_PRELIEVO	Data del prelievo
DATA_ACCETTAZIONE	data dell'accettazione del campione
NOME	Nome dell'azienda da cui sono stati tratti i dati
LUOGO_PRELIEVO	Luogo Prelievo: per esempio macello, allevamento ovaiole
SPECIFICA- RE_LUOGO_PRELIEVO	Luogo Specifico: vuoto per tutte le osservazioni
SPECIE	Specie animale
SPECIE_SPECIF	Specie animale specifica : vuoto per tutte le osservazioni
TIPO_PRELIEVO	Tipo del prelievo: organi_tessuti, feci...
TIPO_PRELIEVO_SPEC	Tipo prelievo Specifico
LAB_RIF	Laboratorio di riferimento
COD_CAMP_LAB	Codice di origine cui il laboratorio assegna un sottocodice
SOTTOSPECIE	Sottospecie della salmonella
SIEROTIPO	Risultato della sierotipizzazione che è la catalogazione base agli antigeni (sostanze che l'organismo giudica estranee o pericolose) che contiene o agli anticorpi che possono contrastarla.
FAGOTIPO	Risultato della fagotipizzazione
DATA_SIERO	Data della sierotipizzazione

NA	Variabile dicotomica che indica o meno resistenza
AMP	Variabile dicotomica che indica o meno resistenza
CTX	Variabile dicotomica che indica o meno resistenza
CIP	Variabile dicotomica che indica o meno resistenza
C	Variabile dicotomica che indica o meno resistenza
GM	Variabile dicotomica che indica o meno resistenza
CAZ	Variabile dicotomica che indica o meno resistenza
COL	Variabile dicotomica che indica o meno resistenza
K	Variabile dicotomica che indica o meno resistenza
S	Variabile dicotomica che indica o meno resistenza
S3	Variabile dicotomica che indica o meno resistenza
TE	Variabile dicotomica che indica o meno resistenza
SXT	Variabile dicotomica che indica o meno resistenza
AMC	Variabile dicotomica che indica o meno resistenza
ENR	Variabile dicotomica che indica o meno resistenza
CF	Variabile dicotomica che indica o meno resistenza
REFERENTE	Referente per le salmonelle, per tutti i campi è Padova
MULTIRESISTENZA	Somma delle resistenze
MULTIRES.DICO	Indica se l'osservazione è multiresistente o no

Variabili eliminate

ID1	Identificativo, codici univoci
ID	Identificativo da laboratorio, codici univoci
LABORATORIO_ORIGINE	Purtroppo sono dati molto frazionati, non danno grossa informatività e non vengono inseriti anche per una questione di privacy
CSPEC	Sottotipologia del campione, vuoto per tutte le osservazioni
COMUNE_PRELIEVO	Comune dove è stato fatto il campionamento, dati molto frazionati, data la natura del monitoraggio si sceglie di fare entrare nel modello la provincia del prelievo
CSPEC	Sottotipologia del campione, vuoto per tutte le osservazioni
ASL	ASL di competenza, dati che possono essere riassunti anche dalla provincia prelievo, di cui è una classificazione in grana più fine
COD_ORIGINE	Codice di origine, univoco per campionamento, quindi è molto frazionato, arbitrario e non rende informazioni
DATA_PRELIEVO	Data del prelievo, le date presentano micronumerosità, sono quasi codici univoci, e vengono eliminate
DATA_ACCETTAZIONE	Stessa situazione delle date precedenti
DATA_SIERO	Data della sierotipizzazione
NOME	Nome dell'azienda da cui sono stati tratti i dati, non è possibile inserirla per questioni di privacy
SPECIFICA-RE_LUOGO_PRELIEVO	Luogo Specifico: vuoto per tutte le osservazioni
SPECIE_SPECIF	Specie animale specifica : vuoto per tutte le osservazioni
TIPO_PRELIEVO_SPEC	Tipo prelievo Specifico: vuoto per tutte le osservazioni
LAB_RIF	Laboratorio di riferimento: uguale per tutte le osservazioni
COD_CAMP_LAB	Codice di origine cui il laboratorio assegna un sottocodice, univoco per ogni osservazione
NA	Variabile dicotomica che indica o meno resistenza

AMP	Variabile dicotomica che indica o meno resistenza
CTX	Variabile dicotomica che indica o meno resistenza
CIP	Variabile dicotomica che indica o meno resistenza
C	Variabile dicotomica che indica o meno resistenza
GM	Variabile dicotomica che indica o meno resistenza
CAZ	Variabile dicotomica che indica o meno resistenza
COL	Variabile dicotomica che indica o meno resistenza
K	Variabile dicotomica che indica o meno resistenza
S	Variabile dicotomica che indica o meno resistenza
S3	Variabile dicotomica che indica o meno resistenza
TE	Variabile dicotomica che indica o meno resistenza
SXT	Variabile dicotomica che indica o meno resistenza
AMC	Variabile dicotomica che indica o meno resistenza
ENR	Variabile dicotomica che indica o meno resistenza
CF	Variabile dicotomica che indica o meno resistenza
REFERENTE	Referente per le salmonelle, per tutti i campi è Padova
MULTIRESISTENZA	Somma delle resistenze

Le variabili che rappresentano la resistenza al singolo antibiotico non vengono inserite poiché si ricerca la multiresistenza del campione, sono da considerare quindi variabili di tipo leaker.

Le variabili provincia prelievo e provincia laboratorio vengono mantenute e studiate singolarmente modello per modello.

La formula da cui partiremo per la determinazione del problema, quindi uguale per tutti i modelli e poi successivamente adattata ad ognuno di essi sarà (scritta in codice R, dove la tilde sta a significare che la multiresistenza dicotomica, viene predetta tramite le covariate inserite di seguito):

MULTIRES.DICO~PROVINCIA LABORATORIO+TIPO CAMPIONE+
PROVINCIA PRELIEVO+LUOGO PRELIEVO+SPECIE+TIPO PRELIEVO
+SOTTOSPECIE + SIEROTIPO+FAGOTIPO+Anno

Si procede senza inserire interazioni, questo perché il numero di covariate è alto, e quindi per ora anche la ricerca di eventuali interazioni non è semplice e verrà dettata dal modello che si utilizza.

Vi è poi da considerare che si è interessati anche all'aspetto puro dei singoli fattori di rischio. Una volta individuato il modello che consideriamo "buono" per descrivere il nostro problema, eventuali interazioni potrebbero essere difficili, o scontate, da interpretare. Quindi per il momento ci addentreremo nella ricerca del modello senza particolari assunzioni sulle interazioni fra fattori.

Modelli generati

Si è scelto di operare costruendo 2 set di dati, il primo con dati non accorpati e il secondo con dati accorpati, questo perché essendo tutte le variabili di tipo categoriale-nominale, l'alto numero di micronumerosità porta spesso a distorsioni nella stima dei parametri (per esempio rendendo non invertibile la matrice $X^T X$ dei modelli lineari) data la presenza di collinearità o multicollinearità.

Il primo set quindi sarà il set dati_a2, mentre il secondo sarà chiamato dati_n2.

Ogni set di dati è stato diviso in un insieme di stima e in un insieme di prova o confronto, il primo con 1850 osservazioni, mentre il secondo con 616, che poi potrebbero diminuire, visto che probabilmente alcune osservazioni "singolari" dovranno essere rimosse perché non essendo state incluse nell'insieme precedente non avranno modo di essere stimate.

Il set di dati dati_n2 verrà utilizzato, soprattutto, come confronto per i parametri, poiché l'alto frazionamento delle osservazioni lo rende difficile da utilizzare quando si utilizzano diagnostiche per la ricerca del modello migliore, quali curve lift o roc.

I modelli utilizzati saranno: analisi discriminante lineare, modelli lineari, modelli lineari generalizzati, foreste casuali.

Modelli lineari

Dal punto di vista puramente operativo si è studiata la funzione scelta, applicata ai modelli lineari, questa poi si è rivelata avere delle collinearità, si è quindi scelto di accorpare una categoria (provincia prelievo TN) e di eliminare la variabile FAGOTIPO; si è proceduto, quindi, ad una selezione automatica delle variabili ritenute significative per il modello. Naturalmente questo è stato fatto tenendo conto anche dell'importanza biologica della variabile e non miopicamente.

Fatto questo, si è giunti alla determinazione (qui riporteremo le indicazioni solo per i dati accorpati, set dati_a2)

Di 2 modelli lineari.

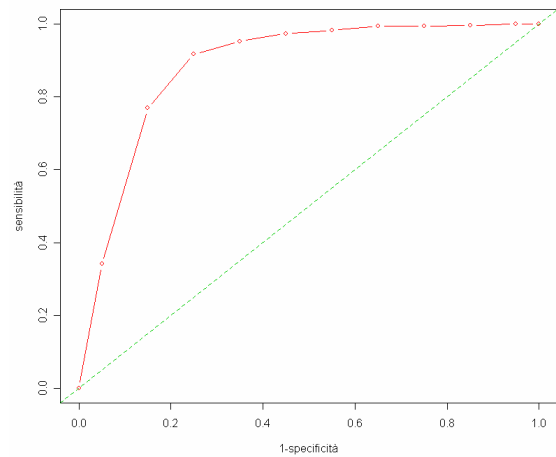
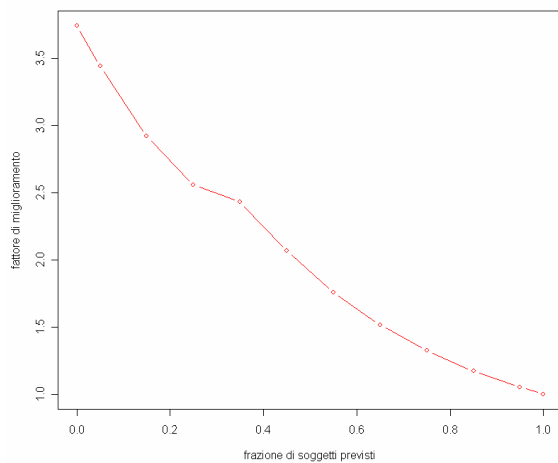
Modello m1_a2_m-dati accorpati e fagotipo non incluso nel modello

Formula:

MULTIRES.DICO ~ PROVINCIA_LABORATORIO + TIPO_CAMPIONE + PROVINCIA_PRELIEVO + LUOGO_PRELIEVO + SPECIE + TIPO_PRELIEVO + SOTTOSPECIE + SIEROTIPO + Anno

Previsione:

Osservati	Previsti				
	0	1			
FALSE	392	49	441	Falsi Negativi	0,111111
TRUE	58	115	173	Falsi Positivi	0,33526
	450	165	614	errore Totale	0,174267



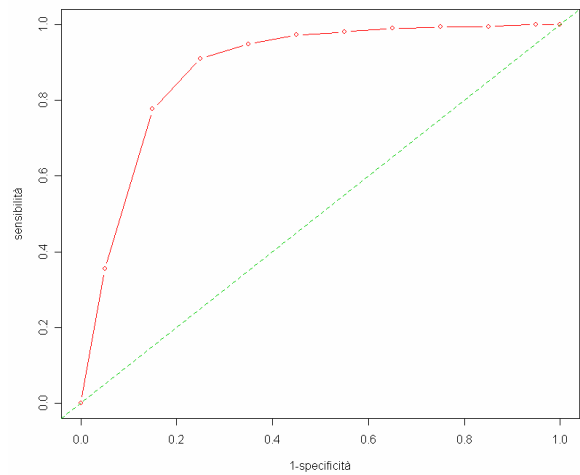
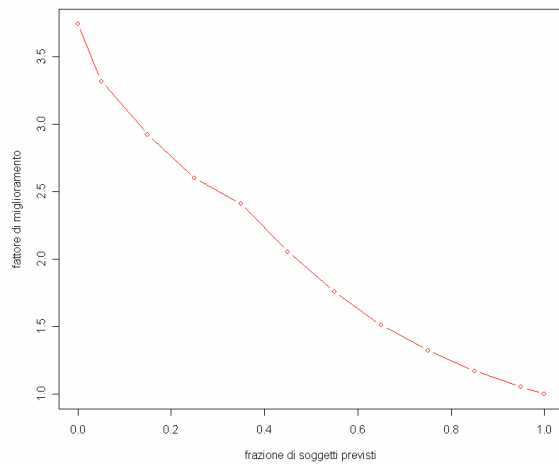
Modello m1_a2_m_s-dati accorpati e fagotipo non incluso nel modello

Formula:

MULTIRES.DICO ~ PROVINCIA_LABORATORIO + LUOGO_PRELIEVO + SPECIE + SIEROTIPO + Anno

Previsione:

Osservati	Previsti				
	0	1			
FALSE	397	49	446	Falsi Negativi	0,109865
TRUE	53	115	168	Falsi Positivi	0,315476
	450	165	614	errore Totale	0,166124



Analisi discriminante lineare

Data la particolare natura dell'analisi discriminante lineare, si cercherà, di creare un modello senza accorpare la modalità "tn" della provincia_prelievo, quindi verranno eliminate le variabili fagotipo e provincia_laboratorio che risultava significativa nel modello lineare, in modo da creare un livello di confronto rispetto al modello precedente e contemporaneamente iniziare a valutare le variabili provincia_prelievo e provincia_laboratorio.

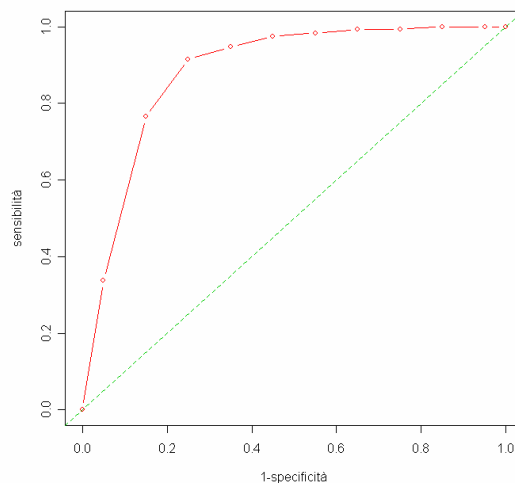
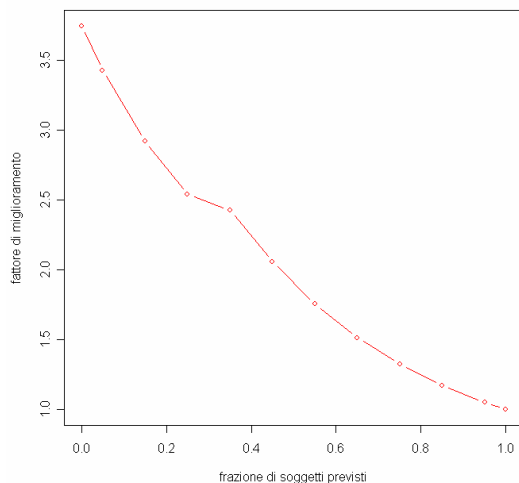
Modello mlda_a22

Formula :

MULTIRES.DICO ~ TIPO_CAMPIONE + PROVINCIA_PRELIEVO + LUOGO_PRELIEVO +SPECIE + TIPO_PRELIEVO + SOTTOSPECIE + SIEROTIPO + Anno

Previsione:

Osservati	Previsti				
	0	1			
FALSE	389	50	439	Falsi Negativi	0,113895
TRUE	61	114	175	Falsi Positivi	0,348571
	450	165	614	errore Totale	0,180782



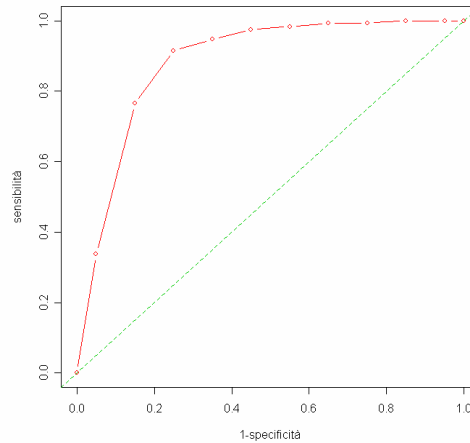
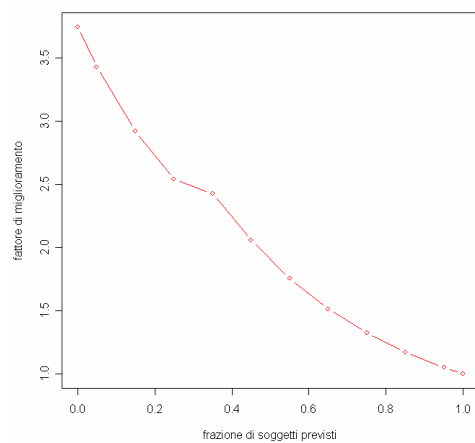
Modello mlda_a2234

Formula :

MULTIRES.DICO ~ TIPO_CAMPIONE + PROVINCIA_LABORATORIO + LUOGO_PRELIEVO + SPECIE + TIPO_PRELIEVO + SOTTOSPECIE + SIEROTIPO + Anno

Previsione:

Osservati	Previsti			Falsi Negativi	0,105505
	0	1			
FALSE	390	46	436	Falsi Positivi	0,337079
TRUE	60	118	178	errore Totale	0,172638
	450	165	614		



Le due curve lift e roc sono pressoché uguali, quindi dovrebbe risultare preferibile, utilizzare la provincia_laboratorio come variabile di regressione, infatti assicura un errore di previsione più piccolo. Questa è l'informazione resa anche dal modello lineare, durante la procedura di selezione delle variabili.

Foreste Casuali

Per quanto riguarda le foreste casuali ci troviamo ad utilizzare un modello di regressione non parametrica, che assicura una certa robustezza nella stima e ci consente, dal punto di vista operativo, anche di stilare una classifica delle variabili che risultano "importanti" ai fini della determinazione o meno della multiresistenza.

A partire dalla funzione che abbiamo determinato all'inizio cerchiamo di vedere quali sono le capacità predittive delle foreste casuali e quale sia l'ordine di importanza assegnato dal modello alle variabili.

Modello rf_a2

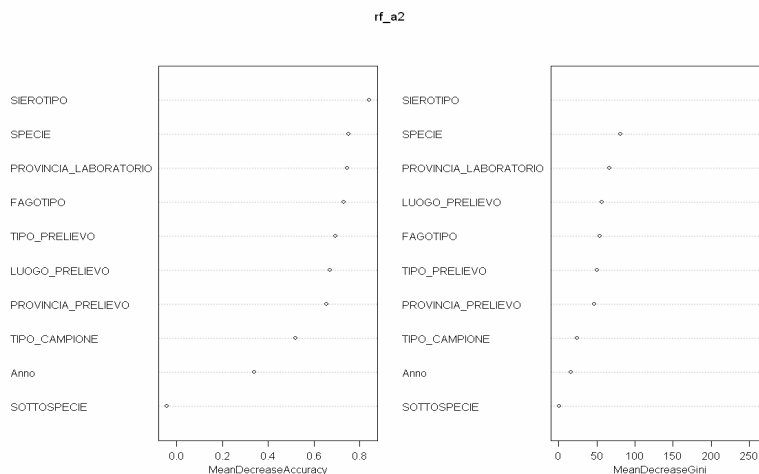
Formula:

MULTIRES.DICO ~ PROVINCIA_LABORATORIO + TIPO_CAMPIONE + PROVINCIA_PRELIEVO + LUOGO_PRELIEVO + SPECIE + TIPO_PRELIEVO + SOTTOSPECIE + SIEROTIPO + Anno

Previsione:

Osservati	Previsti					
	0	1				
FALSE	398	42	440	Falsi Negativi	0,095455	
TRUE	52	122	174	Falsi Positivi	0,298851	
	450	165	614	errore Totale	0,153094	

Tabella di "classificazione" delle covariate :



Si vede immediatamente come, anche qui, il sierotipo e la specie animale siano covariate fondamentali, si potrebbe provare anche ad eliminare la sottospecie dal modello, questa ultima infatti sembra non apportare nessun miglioramento nella fase di previsione della multiresistenza.

La nuova formula sarà quindi:

MULTIRES.DICO ~ PROVINCIA_LABORATORIO + TIPO_CAMPIONE + PROVINCIA_PRELIEVO + LUOGO_PRELIEVO + SPECIE + TIPO_PRELIEVO + SIEROTIPO + FAGOTIPO + Anno

Che non include la sottospecie all'interno dei regressori .

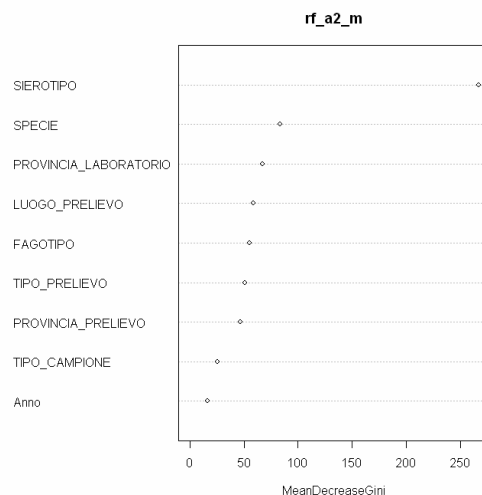
Modello rf_a2_m

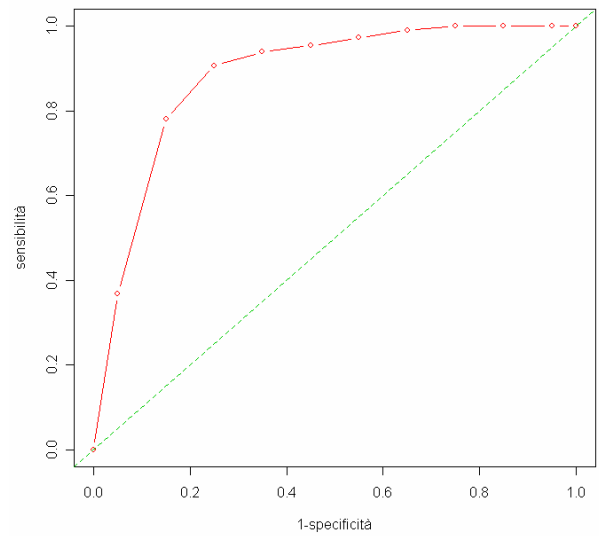
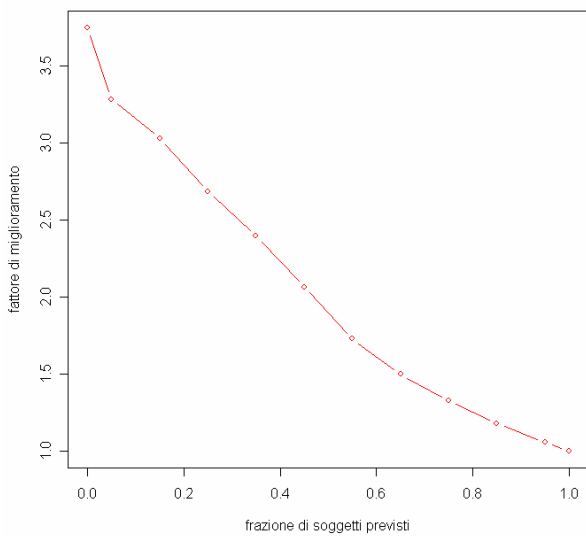
Formula:

MULTIRES.DICO ~ PROVINCIA_LABORATORIO + TIPO_CAMPIONE + PROVINCIA_PRELIEVO + LUOGO_PRELIEVO + SPECIE + TIPO_PRELIEVO + SIEROTIPO + FAGOTIPO + Anno

Previsione:

Osservati	Previsti				
	0	1			
FALSE	400	41	441	Falsi Negativi	0,092971
TRUE	50	123	173	Falsi Positivi	0,289017
	450	165	614	errore Totale	0,148208





Come si può notare il comportamento migliora sensibilmente e anche la quantità di falsi positivi diminuisce.

Manteniamo quindi l'ultimo modello. Il fatto che la sottospecie sia risultata quasi "deleteria" è da attribuire probabilmente al fatto che presenta alcune micronumerosità che si sono ulteriormente "spezzate" nel momento della creazione dei due gruppi di stima e prova, questo può portare a distorsioni all'interno della stima e nella fase di classificazione.

Ricordiamo che la sottospecie è stata inserita all'inizio quale variabile di regressione motivata da un fattore "biologico", vale a dire strettamente legato alla struttura del problema, scegliendo poi di toglierla nel caso il modello ci suggerisse di farlo, come in questa occasione, o nei modelli lineari.

Regressione logistica

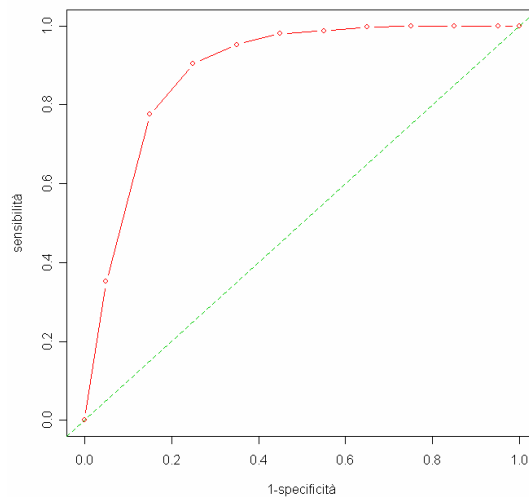
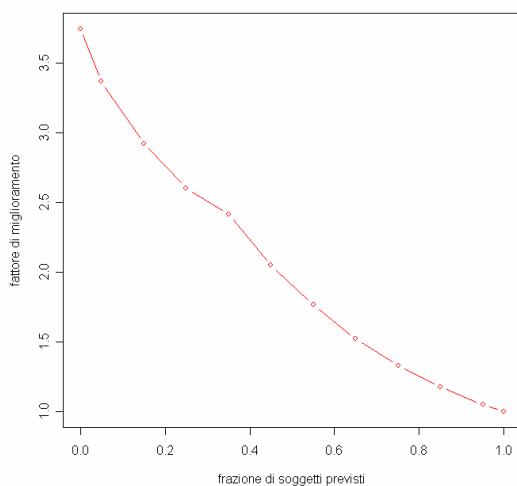
Modello mglm_a2_s

Formula:

MULTIRES.DICO ~ PROVINCIA_LABORATORIO + LUOGO_PRELIEVO + SPECIE + SIEROTIPO + Anno

Previsione:

Osservati	Previsti			
	0	1		
FALSE	400	50	450	Falsi Negativi 0,111111
TRUE	50	114	164	Falsi Positivi 0,304878
	450	165	614	errore Totale 0,162866



E' stato inserito direttamente il modello generato dalla selezione automatica delle variabili. Si è scelto di mantenere l'anno, anche se risultava non significativo, lo si è fatto perché, avendo solo 2 modalità con una buona numerosità, comunque non appesantiva troppo il modello finale, la devianza risultava comunque buona e poteva essere interessante mantenere un parametro di confronto utile per quel che riguarda le informazioni che questi modelli renderanno nel momento in cui ci si occuperà dello studio dei parametri. Non sono state inserite interazioni.

Scelta del modello

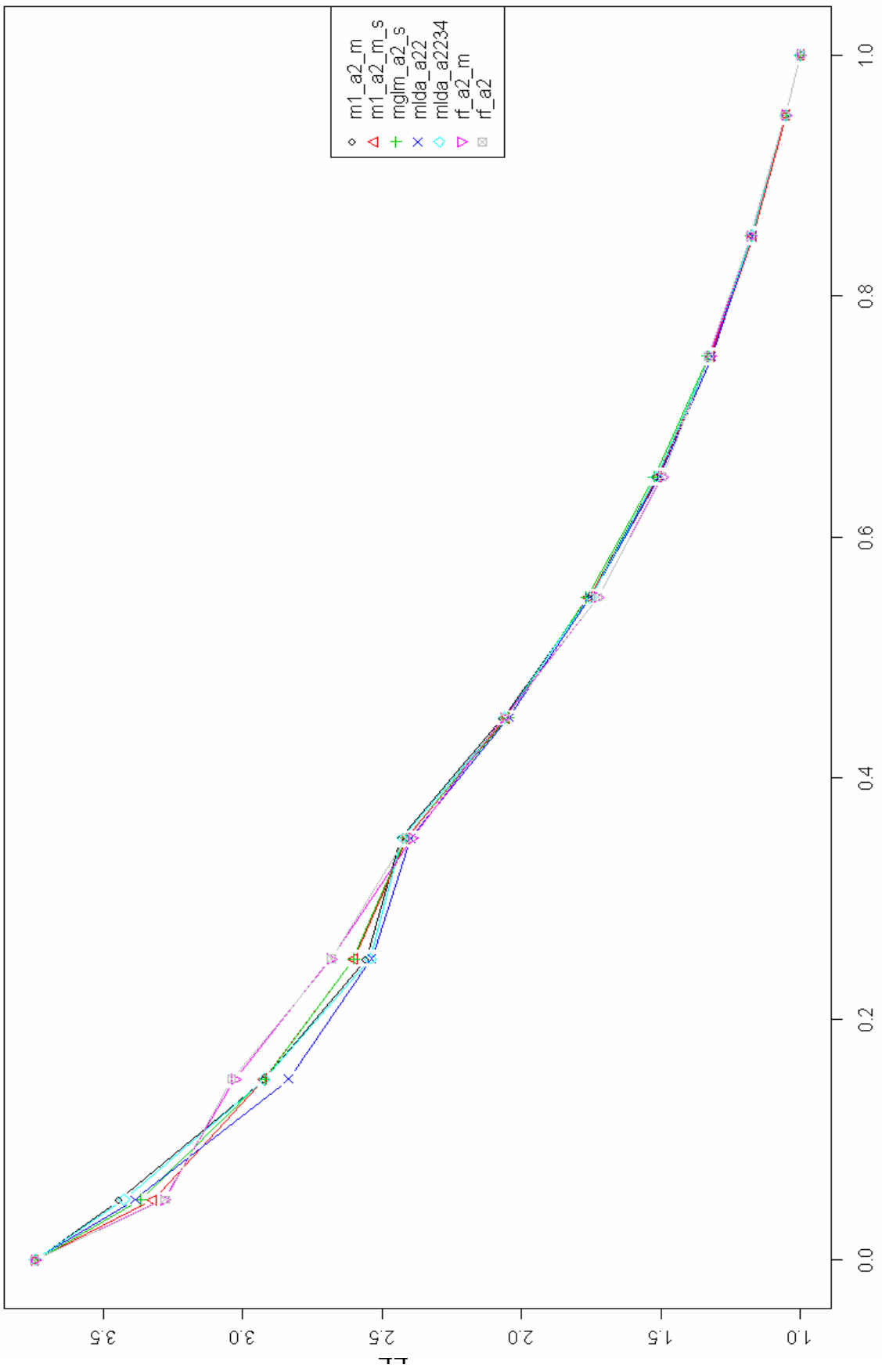
Sceghieremo il modello in base alla sua capacità di classificazione, alla sua capacità di "miglioramento" ovvero tramite il confronto delle curve lift e in base alla semplicità di interpretazione.

Questa ultima assunzione non è così banale se si pensa che le reti neurali hanno la possibilità, aumentando i nodi nello strato latente, di approssimare pressoché tutte le funzioni, ma questo implica il dover dipanare un modello forse maggiormente complesso di quello iniziale dal punto di vista interpretativo.

	Falsi Negativi	Falsi Positivi	errore Totale
rf_a2_m	0,09297	0,28902	0,14821
rf_a2	0,09546	0,29885	0,15309
mglm_a2_s	0,11111	0,30488	0,16287
m1_a2_m_s	0,10987	0,31548	0,16612
mlda_a2234	0,10551	0,33708	0,17264
m1_a2_m	0,11111	0,33526	0,17427
mlda_a22	0,11390	0,34857	0,18078

Il modello che sembra comportarsi meglio dal punto di vista della classificazione è la foresta casuale privata della sottospecie, con un errore totale del 14,8%. Tutti i modelli comunque sembrano comportarsi in maniera quantomeno "dignitosa".

Per quanto riguarda le curve lift otteniamo questo grafico che comunque risulta "abbastanza" interpretabile.



Si riscontra infatti che il modello lineare con tutti le covariate, quindi non sottoposto a selezione automatica delle variabili, risulta il miglior classificatore per percentuali "alte" mentre le foreste casuali, vista anche la minor tendenza a classificare falsi positivi tendono a rimanere più "stabili" nella classificazione per il resto delle percentuali, questo andamento si ripete per tutti i classificatori con un alto numero di covariate e con un alto numero di falsi positivi.

Consideriamo quindi i modelli creati con foreste casuali più stabili, ma teniamo a mente che probabilmente per percentuali assegnate "alte" risulta leggermente sottostimato il rischio di multiresistenza. Per questo si potrebbe anche pensare di lavorare sulla soglia dei modelli lineari, si potrebbe scegliere una soglia più alta per diminuire il numero di falsi positivi.

Vediamo ora l'interpretabilità.

Le foreste casuali risulterebbero totalmente non interpretabili dal punto di vista dei parametri, se non ci fosse data comunque la possibilità di stimare una graduatoria dell'impatto sulla diminuzione media dell'entropia all'inserimento di una delle covariate. I modelli lineari e gli altri utilizzati rendono invece delle stime che possono essere utilizzate, parametro per parametro per stimare l'importanza della singola modalità assunta dalla variabile.

La scelta ricade quindi sulle foreste casuali, che rispetto agli altri modelli hanno un errore globale basso e un comportamento rispetto alla curva lift e all'interpretabilità apprezzabili per chiarezza e robustezza.

Questo però implica che potremo utilizzare questo tipo di modellazione come "solco" per lo studio dei parametri, risulta infatti evidente che vi sono delle covariate che, venendo scelte sempre dai modelli quali variabili statisticamente significative e avendo spesso anche un forte significato biologico, hanno un forte impatto sulla determinazione delle multiresistenze.

Di queste covariate può risultare utile confrontare le varie modalità fra i modelli, per determinare quali siano i fattori di rischio effettivi, questo tipo di studio è indicativo per quanto riguarda i modelli lineari, mentre può risultare abbastanza preciso nell'ambito di modelli lineari generalizzati. Naturalmente tutto questo viene eseguito tenendo sempre a mente che si confrontano 2 o più modelli che per struttura possono dare risposte anche diverse (spesso per piccole numerosità) ma indicativamente si può pensare di strutturare un confronto fra parametri.

Confronto fra i parametri

Per prima cosa iniziamo con lo studiare quali siano le cause che portano ogni modello ad inserire la variabile `provincia_laboratorio` all'interno dei modelli.

Queste due variabili sono state inserite nello studio entrambe, anche se si poteva supporre una certa correlazione, perché sarebbe stato utile cercare di vedere se vi fosse una maggioranza di multiresistenze all'interno di una certa provincia o vi fosse una tendenza di qualche laboratorio a sovrastimare la resistenza di un campione.

Questo è il motivo iniziale dell'inserimento, ora però vi è la necessità di controllare quale sia il motivo della maggior "importanza" di `provincia_laboratorio` rispetto a `provincia_prelievo`.

Questo tipo di riscontro è dovuto al cosiddetto effetto cluster, che a causa del tipo di campionamento ci troviamo spesso a riscontrare, ma che si sperava di eliminare attraverso la pulitura dei dati o piuttosto ad un qualche tipo di diversità da provincia a provincia per quel che riguarda le multiresistenze.

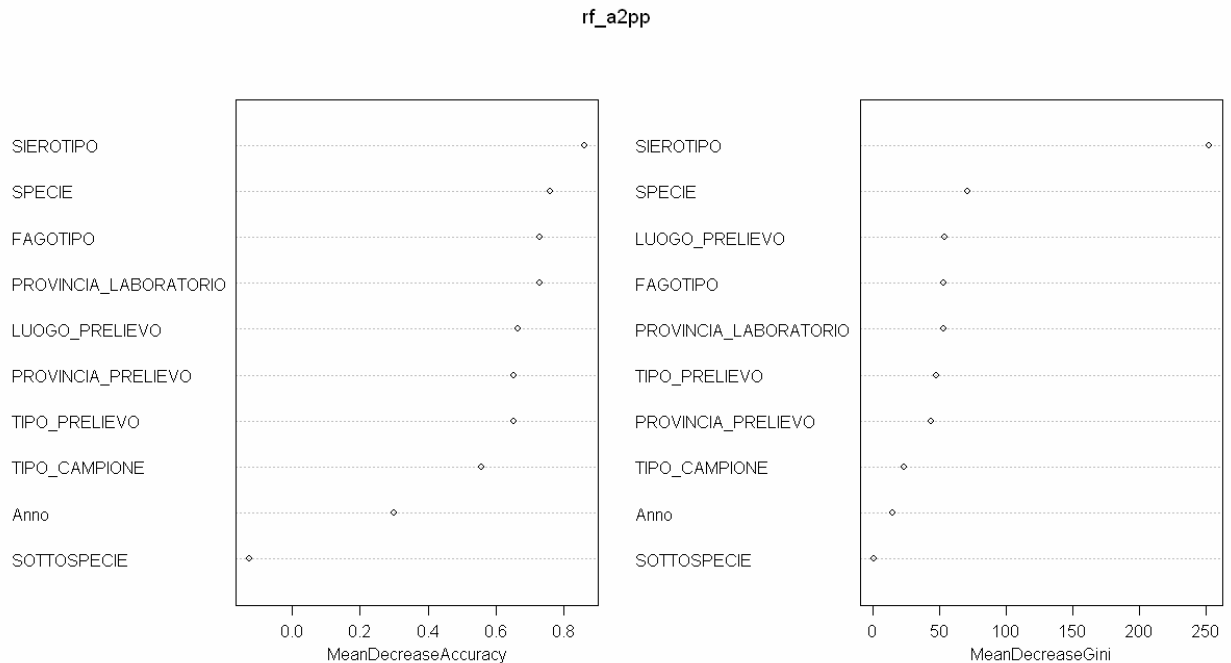
Guardando più a fondo, con un'analisi capillare, notiamo come vi sia un cluster di multiresistenze per quel che riguarda la provincia `laboratorio` di BO, i campioni però provenivano dalla provincia di TV, che risultava però bilanciata nella quantità di multiresistenze rispetto al valore di confronto assegnato.

Tutti Questi campioni, che creano un cluster molto grande di multiresistenze, però provengono da una stessa azienda anche se da sedi distaccate, quindi, dal punto di vista biologico i dati sono attendibili, mentre dal punto di vista della localizzazione possono sorgere alcuni problemi.

Si potrebbe provare a ricostruire i dati eliminando queste osservazioni, vedendole come un gruppo a parte, considerandole una entità diversa per poi vedere come si comportano i dati confrontandoli con i risultati precedenti, per vedere quanto cala la pericolosità collegata alle province e da questo notare se esista una omogeneità o meno nella distribuzione delle salmonella rispetto al luogo da cui viene prelevata.

Utilizzando solo lo strumento dei modelli lineari come confronto, per quanto riguarda i parametri ritroviamo una situazione pressoché identica alla precedente, senza però la provincia di Bologna, quindi pro-

viamo a vedere cosa succede utilizzando uno strumento leggermente più "sensibile" alla conformazione dei dati come le foreste casuali, per la classificazione dei regressori.



La situazione risulta pressoché uguale a prima, anche se vediamo che "abbastanza logicamente" il regressore provincia_laboratorio cala in importanza andandosi a inserire dopo il fagotipo, anche se effettivamente abbiamo la stessa capacità di diminuzione dell'entropia.

La cosa interessante è che Luogo_prelievo diventa il terzo regressore in ordine di importanza, questo risulta interessante, perché a differenza del sierotipo, questa variabile è qualcosa su cui si può "lavorare" anche in ambito pratico.

Possiamo considerare la provincia laboratorio associata alla provincia prelievo e comunque un indicatore importante per questo set di dati, che sembra avere indicato zone di maggiore pericolosità per quanto riguarda le multiresistenze nelle province di Udine e Venezia, anche se, data la non enorme quantità di osservazioni a disposizione e la provenienza da piani di monitoraggio e non propriamente di campionamento non possiamo che trarre un elemento di informazione indicativa non potendo esprimere pareri definitivi, ma solo utili ad uno studio più approfondito.

Sierotipo

In tutti i modelli il sierotipo risulta essere la covariata più importante.

Confrontiamo quindi i parametri del modello lineare con quelli del glm e della lda per vedere quali siano i parametri in accordo.

C'è da premettere che sono stati creati due gruppi, uno con dati accorpati e uno con dati non accorpati, questi due gruppi sono stati controllati in modo incrociato, per vedere se ci fossero discordanze, che non si sono riscontrate, quindi, per maggior chiarezza espositiva vengono inseriti i sierotipi presenti nel modello accorpato che erano comunque i più numerosi e quelli di maggior interesse.

Modello Lineare

PARAMETRO	STIMA	STD.ERROR	Sig
	-0,333839	0,058487	***
Altro	-0,350071	0,033417	***
Gruppo B	0,182746	0,086406	*
S. Agona	-0,207870	0,065191	**
S. Anatum	-0,212637	0,073294	**
S. Blockley	0,175105	0,050694	***
S. Derby	-0,336625	0,044742	***
S. Enteritidis	-0,374884	0,040818	***
S. Hadar	0,290487	0,052181	***
S. Heidelberg	0,206729	0,048902	***
S. Infantis	-0,413866	0,080998	***
S. Livingstone	-0,372292	0,038768	***
S. London	-0,264353	0,073702	***
S. Mbandaka	-0,349434	0,075859	***
S. Nuovo Sierotipo	0,227366	0,066634	***
S. Rissen	-0,425917	0,080894	***
S. Tennessee	0,387399	0,095009	***
S. Thompson	-0,429013	0,074635	***
S. Virchow	-0,336191	0,073895	***

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Poiché gli errori standard e quindi l'inferenza sulla significatività non possano essere paragonate a quelli dei modelli lineari (ci sono peraltro risultati incoraggianti in alcuni studi che segnalano come possano essere considerati attendibili) scegliamo di prendere quei sierotipi che presentano alta significatività. Sono stati evidenziati quei sierotipi che

risultano maggiormente multiresistenti. Risultano notevoli le pericolosità dei sierotipi

Gruppo B
S. Blockley
S. Hadar
S. Heidelberg
S. Nuovo Sierotipo
S. Tennessee

Regressione logistica

PARAMETRO	STIMA	STD.ERROR	T.VALUE	Sig	ODDS
	-1,97E+00	5,12E-01	-3,837	***	0,140016
Altro	-2,09E+00	2,66E-01	-7,866	***	0,123811
Gruppo B	7,81E-01	5,08E-01	1,537		2,183218
S. Agona	-8,56E-01	4,30E-01	-1,989	*	0,424773
S. Anatum	-9,94E-01	4,59E-01	-2,163	*	0,370278
S. Blockley	7,81E-01	3,13E-01	2,496	*	2,183
S. Bredeney	1,17E-01	3,02E-01	0,388		1,124344
S. Derby	-1,75E+00	3,13E-01	-5,611	***	0,173253
S. Enteritidis	-2,74E+00	4,58E-01	-5,98	***	0,0647
S. Hadar	1,49E+00	3,40E-01	4,38	***	4,42823
S. Heidelberg	9,49E-01	3,05E-01	3,115	**	2,583125
S. Infantis	-3,15E+00	1,06E+00	-2,978	**	0,042938
S. Livingstone	-2,70E+00	3,71E-01	-7,291	***	0,067071
S. London	-1,28E+00	5,01E-01	-2,562	*	0,276927
S. Mbandaka	-2,07E+00	7,69E-01	-2,688	**	0,126692
S. Nuovo Sierotipo	9,16E-01	4,30E-01	2,133	*	2,499773
S. Rissen	-3,06E+00	1,04E+00	-2,93	**	0,047076
S. Saintpaul	-2,21E-01	3,93E-01	-0,561		0,802118
S. Tennessee	2,48E+00	8,49E-01	2,924	**	11,96517
S. Thompson	-1,63E+01	4,31E+02	-0,038		8,76E-08
S. Virchow	-2,09E+00	6,88E-01	-3,04	**	0,123317

Signif. codes:
 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

In questo caso, visto che utilizziamo uno strumento adeguato al nostro tipo di classificazione, manteniamo tutti i parametri, in modo da averne un quadro più dettagliato, come indicatore utilizzeremo gli odds. Ogni parametro è considerato al netto di tutti gli altri parametri.

In dettaglio cala solamente la significatività del sierotipo Gruppo B, mentre gli altri

S. Blockley
S. Hadar
S. Heidelberg

S. Nuovo Sierotipo
S. Tennessee

Vengono confermati nella loro "inclinazione" alla multiresistenza. Il sierotipo S. Tennessee, non presenta grosse numerosità, ma è pressoché sempre multiresistente pur provenendo da zone differenti e da tipi di campionamento differenti, quindi sembra plausibile che venga considerato un fattore di rischio importante.

Lda

PARAMETRO	0	1
	0,0346	0,0087
Altro	0,1745	0,0398
Gruppo B	0,0055	0,0225
S. Agona	0,0236	0,0156
S. Anatum	0,0165	0,0138
S. Blockley	0,0197	0,0917
S. Bredeney	0,0259	0,0623
S. Derby	0,0645	0,0294
S. Enteritidis	0,1077	0,0104
S. Hadar	0,0149	0,09
S. Heidelberg	0,0228	0,2007
S. Infantis	0,0173	0,0017
S. Livingstone	0,2154	0,0225
S. London	0,0181	0,0104
S. Mbandaka	0,0204	0,0035
S. Nuovo Sierotipo	0,0071	0,0467
S. Rissen	0,0173	0,0017
S. Saintpaul	0,0181	0,0277
S. Tennessee	0,0016	0,0277
S. Thompson	0,0228	0
S. Virchow	0,0204	0,0052

In questo caso non abbiamo indicazioni inferenziali, ma in definitiva risultano le stesse indicazioni che vengono fornite dagli altri modelli, possiamo quindi concludere che nel nostro studio i sierotipi maggiormente pericolosi sono :

S. Blockley
S. Hadar
S. Heidelberg
S. Nuovo Sierotipo
S. Tennessee

In particolare il sierotipo Hadar a fronte di una buona numerosità presenta una tendenza alla multiresistenza che può essere considerata

circa 4 volte superiore a quella del sierotipo di riferimento S. Typhimurium.

Specie

La specie risulta essere una covariata "privilegiata", in quanto fa parte di quelle categorie di variabili di determinazione biologica, quindi ritenute importanti per localizzare la presenza della multiresistenza.

Modello Lineare

PARAMETRO	STIMA	STD.ERROR	Sig
Bovino	0.121754	0.051074	*
Bovino-Suino	0.171635	0.082398	*
Coniglio	0.159207	0.071654	*
Non noto	0.086231	0.044773	.
Suino	0.091733	0.033651	**
Tacchino	0.092930	0.038693	*

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Tutte le specie risultano maggiormente multiresistenti rispetto al pollo, soprattutto i bovini e i conigli, ma anche i suini e i tacchini che comunque hanno numerosità elevate.

Regressione logistica

PARAMETRO	STIMA	STD.ERROR	T.VALUE	Sig.	ODDS
Bovino	7,54E-01	3,38E-01	2,23	*	2,124422
Bovino-Suino	1,06E+00	5,47E-01	1,938	,	2,886371
Coniglio	8,99E-01	4,90E-01	1,833	,	2,45739
Faraona	5,55E-01	4,72E-01	1,176		1,742115
Molluschi	-6,28E-01	7,01E-01	-0,895		0,533925
Non noto	5,87E-01	3,22E-01	1,824	,	1,798764
Suino	6,54E-01	2,47E-01	2,651	**	1,923026
Tacchino	5,16E-01	2,57E-01	2,006	*	1,674476

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Lda

PARAMETRO	0	1
Bovino	0,0299	0,0554
Bovino-Suino	0,0126	0,0208
Coniglio	0,0126	0,026
Faraona	0,0102	0,0277
Molluschi	0,0157	0,0069
Non noto	0,0503	0,0657
Suino	0,1903	0,2197
Tacchino	0,0668	0,2751

Nella variabile specie bisogna ben definire alcuni valori, probabilmente infatti dal punto di vista puramente numerico il valore che maggiormente si avvicina ai dati è quello del modello Lda. Ora risulta chiaro come molte specie siano maggiormente multiresistenti del Pollo, che è specie di riferimento, ma se facessimo un confronto noteremmo che, comunque, il tacchino resta la specie maggiormente multiresistente (questo potrebbe essere dovuto alla minore numerosità delle altre specie che portano gli algoritmi a sovrastimare il rischio). Una precisazione è necessaria per la specie Bovino-Suino, questa non è propriamente una specie, in realtà di solito si utilizza questa categoria quando vi è un misto di carni su cui viene fatto un prelievo; queste modalità derivano in massima parte da luoghi Lcfin, quindi luoghi in cui il consumatore finale acquista le merci. Questo potrebbe essere considerato un segnale di allarme per quanto riguarda i controlli su questo tipo di carni.

Fagotipo

Sul fagotipo non si può dire nulla, se non che è considerata dalle foreste casuali una variabile di regressione abbastanza importante, purtroppo è troppo frazionata per essere stimato dai modelli parametrici.

Luogo Prelievo

Modello Lineare

PARAMETRO	STIMA	STD.ERROR	Sig
Allevamento da ingrasso	0.118403	0.038817	**
Allevamento da riproduzione	0.162230	0.083928	.
Incubatoio	0.138312	0.045696	**
Laboratorio di sezionamento	0.094540	0.052130	.
Macello	0.113250	0.037069	**

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Come si evince da questi parametri i luoghi in cui si riscontrano maggiormente le multiresistenze sono luoghi in cui gli animali sono ammassati (incubatoi, allevamenti da ingrasso, allevamenti da riproduzione) oppure le carni vengono lavorate, come i macelli e i laboratori di sezionamento. Questi dati devono essere confrontati con il valore di riferimento che è la modalità Allevamento ovaiole nella quale l'incidenza della multiresistenza risulta bassa; Questo è anche dovuto al fatto che i polli risultano essere tra le specie più "protettive" rispetto alla multiresistenza.

Incuriosisce però vedere quale sia l'incidenza delle multiresistenze anche nei polli, rispetto al luogo in cui, data la specie pollo, viene fatto il prelievo.

Solo dal punto di vista descrittivo controlliamo quale sia l'andamento dell'incidenza delle multiresistenze nei polli.

	Numerosità	Multiresistenze	%
Incubatoio	189	8	4,232804
Allevamento ovaiole	225	31	13,77778
Allevamento da ingrasso	229	39	17,03057
Allevamento da riproduzione	4	2	50
Macello	305	64	20,98361
Laboratorio di sezionamento	28	2	7,142857
Lcfin	14	4	28,57143
Altro	42	10	23,80952
Non noto	54	4	7,407407

Solo da questo si riesce a intuire come il pollo sia poco propenso a produrre multiresistenza, vista l'incidenza molto bassa che ha negli incubatoi, ma in ambienti quali macelli (i punti di vendita al dettaglio "lcfin" sono troppo poco numerosi) e allevamenti da ingrasso tenda a sviluppare percentuali maggiori di multiresistenza, questo diversamente dai tacchini che fin dall'incubatoio presentano percentuali molto alte (88 su 103, 85%).

Queste osservazioni sembrano far propendere per il fatto che alcuni luoghi favoriscano il manifestarsi di multiresistenza.

Regressione logistica

PARAMETRO	STIMA	STD.ERROR	T.VALUE	Sig.	ODDS
Allevamento da ingrasso	1,16E+00	3,65E-01	3,183	**	3,19632
Allevamento da riproduzione	1,30E+00	6,26E-01	2,071	*	3,654649
Altro	7,51E-01	3,96E-01	1,896	,	2,11933
Incubatoio	1,39E+00	4,64E-01	3,006	**	4,030942
Laboratorio di sezionamento	9,27E-01	4,33E-01	2,141	*	2,527675
Lcfin	4,66E-01	4,29E-01	1,086		1,593766
Macello	1,08E+00	3,46E-01	3,115	**	2,941736
Non noto	3,94E-01	4,07E-01	0,969		1,483345

Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Dalla regressione logistica otteniamo pressoché le stesse informazioni, risulta comunque preoccupante che gli incubatoi risultino altamente propensi a multiresistenza che si concentra nella specie Tacchino.

Lda

PARAMETRO	0	1
Allevamento da ingrasso	0,1509	0,1696
Allevamento da riproduzione	0,0086	0,0225
Altro	0,1187	0,0986
Incubatoio	0,114	0,1384
Laboratorio di sezionamento	0,0495	0,0588
Lcfin	0,0597	0,0761
Macello	0,3019	0,3183
Non noto	0,077	0,0779

CONCLUSIONI

Lo studio ha portato alla luce dei fattori di rischio strettamente correlati alla natura biologica dei campioni osservati.

Sierotipo e specie risultano fondamentali nel classificare la presenza o meno di multiresistenza, questo significa che, partendo da queste due covariate, si può pensare di classificare come più o meno a rischio alcune specie e alcuni sierotipi, in particolare i sierotipi:

S. Blockley
S. Hadar
S. Heidelberg
S. Nuovo Sierotipo
S. Tennessee

Risultano fattori importanti per la determinazione di eventuali multiresistenze. Anche se non particolarmente "importante" dal punto di vista statistico l'anno 2006 risulta essere maggiormente protettivo rispetto all'anno 2005 nei confronti della multiresistenza, questo dato è sufficientemente confortante ma va comunque mantenuto nell'ambito delle ipotesi vista la natura dei dati.

La cosa realmente interessante risulta il legame fra luoghi molto affollati o in cui la carne viene lavorata e l'aumento di multiresistenza.

Da questo riscontro si potrebbe partire per creare un piano di campionamento ambientale al fine di studiare meglio tale evidenza, questo infatti sembra uno dei pochi punti sui quali si possa realmente lavorare per cercare di far diminuire il rischio.

L'osservazione risulta tanto più importante se pensiamo che gli isolamenti da campioni di specie bovino-suino sono localizzati prevalentemente nella parte finale della filiera produttiva.

Il programma di monitoraggio potrebbe essere ristrutturato per gestire la raccolta dei dati in modo maggiormente capillare e uniforme rispetto a tutte le province, cercando di bilanciare il numero di campioni, magari inserendo il numero di tutte le analisi effettuate rispetto al numero di campioni spediti, in modo tale da poter assegnare anche un determinato "peso" alle informazioni ottenute, alcune province infatti sembrano inviare un numero molto minore di osservazioni e visto che si tratta solo di campioni in cui è stata riscontrata salmonella non si può risalire al numero di prelievi effettuati.

Conoscendo il numero complessivo di campioni raccolti forse si riuscirebbe ad ottenere una maggiore informatività dallo studio.

Sarebbe altresì interessante, data l'importanza assunta nello studio, sottoporre a tipizzazione fagica tutti i campioni in modo da ottenere maggiori informazioni e poter utilizzare strumenti parametrici per stimare in "profondità" la pericolosità dei vari fattori di rischio collegati a questa covariata.

Dal punto di vista operativo quindi potrebbe risultare utile controllare quale sia l'incidenza di multiresistenza e per quali antibiotici viene sviluppata resistenza all'interno di allevamenti di cui si conoscono la pressione farmacologica e la pressione selettiva.

Utilizzando questo tipo di dati si potrebbe infatti esaminare quale sia il collegamento fra consanguineità, selezione di matrice umana e multiresistenza e quale sia il rapporto fra l'utilizzo di strumenti farmacologici e sviluppo delle resistenze nelle popolazioni animali.

Considerando che alla luce dei risultati esiste una disparità tra multiresistenza riscontrata nei tacchini rispetto ai polli la causa potrebbe essere una pressione selettiva esagerata?

Nota: è stato costruito un piccolo programma per la pulitura dei dati su piattaforma vba (con il fermo proposito di convertirlo in codice Java), il programma e il codice sono a disposizione del centro ospitante.

SOMMARIO:

<u>Introduzione al problema.....</u>	<u>1</u>
Scopi e descrizione	3
Il piano di monitoraggio Enter-Vet	3
<u>Salmonelle e determinazione del problema.....</u>	<u>5</u>
Descrizione generale.....	5
Habitat e diffusione.....	5
Struttura antigenica.....	5
Sensibilità agli antibiotici	5
Determinazione del problema.....	6
Gruppi di salmonelle	6
Fenomeno dell'antibiotico resistenza	6
L'Antibiogramma	7
Antibiotici usati nell'analisi	8
<u>Analisi tramite tecniche di data mining</u>	<u>9</u>
Definizione di data mining	9
Data Mining, gli oggetti di interesse e i pericoli.....	9
Scelta degli strumenti per l'analisi del problema	10
La regressione lineare applicata a modelli di classificazione	10
Analisi discriminante	11
Alberi di classificazione.....	13
La regressione logistica	17
Accenno agli errori nella regressione logistica.....	18
Adeguare il modello	19
<u>Analisi descrittiva per i dati in esame.....</u>	<u>23</u>
Sierotipi	23
Anno 2005.....	23
Anno 2006.....	24
Specie animali.....	25
Anno 2005.....	25
Anno 2006.....	25
Specie Animali e Sierotipi	26
Anno 2005.....	26
Anno 2006.....	27
Luogo Prelievo	28
Anno 2005.....	28
Anno 2006.....	28
Resistenze e sierotipi	29
Anno 2005.....	29
Anno 2006.....	30
<u>Analisi dei dati tramite tecniche di data mining</u>	<u>31</u>
Variabili.....	31
Variabili eliminate.....	32
Modelli generati.....	34
Modelli lineari.....	34

Analisi discriminante lineare.....	37
Foreste Casuali.....	39
Regressione logistica.....	42
<u>Scelta del modello.....</u>	<u>43</u>
<u>Confronto fra i parametri.....</u>	<u>47</u>
Sierotipo.....	49
Modello Lineare.....	49
Regressione logistica.....	50
Lda.....	51
Specie.....	52
Modello Lineare.....	52
Regressione logistica.....	52
Lda.....	53
Fagotipo.....	53
Luogo Prelievo.....	54
Modello Lineare.....	54
Regressione logistica.....	55
Lda.....	55
<u>CONCLUSIONI.....</u>	<u>57</u>

Riferimenti Bibliografici:

A. Azzalini, B. Scarpa (2004). *Analisi dei dati e data mining*. Springer-Verlag Italia.

David W. Hosmer, Stanley Lemeshow (2000). *Applied Logistic Regression*. Wiley Series in Probability and Statistics, II Edition.

L. Ventura. *Modelli Statistici II, Modelli lineari generalizzati*. Dipartimento di Scienze Statistiche, Università degli Studi di Padova. (Dispensa)

A. Ricci, M. Mancin, V. Cibin, L. Busani. *Entervet 2004: Rapporto Annuale*(2005). Istituto Zooprofilattico Sperimentale delle Venezie.

G. Poli, A. Cocilovo et. Al.. *Microbiologia e immunologia veterinaria* (2000). Utet Torino, II Edizione.

"Antigene," Microsoft® Encarta® Enciclopedia Online (2008)
<http://it.encarta.msn.com> © 1997-2008 Microsoft Corporation. Tutti i diritti riservati.

Thomson Gale ."*Selective Pressure*," from World of Biology(2005-2006). Thomson Gale, a part of the Thomson Corporation.