

## CONTENUTO PROTEICO DEL GRANO E VARIABILI AGRONOMICHE: UN'ANALISI STATISTICA

D. G. Calò, F. Scotto, S. Ravaglia

### 1. INTRODUZIONE

L'interesse per il tenore proteico del grano deriva dall'importanza fondamentale svolta dal glutine (che costituisce circa l'80% del contenuto proteico) nel determinare le caratteristiche tecnologiche degli impasti, e quindi dei derivati, delle farine di frumento.

Prove varietali sono svolte regolarmente da ormai più di vent'anni su tutto il territorio nazionale, con l'obiettivo di individuare e selezionare varietà sempre più produttive e rispondenti alle diversificate destinazioni di impiego dei grani coltivati.

A fronte di una attività di confronto varietale così vasta e ben programmata dal punto di vista sperimentale, riguardo agli effetti delle componenti agronomiche sul livello proteico delle cariossidi di frumento non si riscontrano studi altrettanto pianificati, basati sull'osservazione simultanea dei diversi fattori agronomici e su campioni sufficientemente numerosi.

Questo lavoro ha preso le mosse dall'osservazione congiunta di molteplici componenti agronomiche sulle coltivazioni di frumento delle province di Bologna e Modena nell'annata agraria 1999-2000. Dopo aver controllato la qualità e l'omogeneità dei dati disponibili, gli effetti dei fattori agronomici, e delle loro interazioni, sul tenore proteico del grano sono stati studiati impiegando la metodologia di regressione non parametrica nota con il nome di alberi di regressione. Per motivi di reperibilità dei dati si è ristretta l'analisi al solo frumento tenero, che rappresenta il cereale a paglia più coltivato nell'areale considerato.

### 2. I DATI: FONTI E QUALITÀ

Le unità statistiche prese in esame corrispondono a 113 partite di grano, conferite nelle cinque agenzie del Consorzio Agrario di Bologna e Modena (CABM), e a 7 prove parcellari, ripetute per due varietà, coordinate dal Centro Ricerche Produzioni Vegetali (CRPV) della Regione Emilia Romagna. I dati relativi alle prove parcellari, in parte pubblicati sull'Informatore Agrario (Poli *et al.*, 2000), sono stati forniti

dal CRPV. I dati del Consorzio sono stati tratti dalle schede agronomiche<sup>1</sup> raccolte nell'ambito del progetto Proteina+ (avviato dal Consorzio stesso per l'annata agraria 1999-2000), riferite alle varie aziende agricole coinvolte nel progetto. In tale ambito i cerealicoltori erano stati guidati, riguardo alle tecniche colturali, dai tecnici del CABM sulla base delle conoscenze disponibili e delle caratteristiche specifiche di ogni appezzamento (analisi del terreno, precessione colturale).

Tra le variabili disponibili, in questo studio sono stati considerati i fattori agronomici che descrivono aspetti potenzialmente interessanti, ai fini della ricerca, delle principali componenti di un agroecosistema.

Ad esempio, i fattori climatici sono fonte di variabilità, nello spazio e nel tempo, ed è quindi importante tenerne conto. Tra questi è particolarmente interessante la piovosità nel periodo che va dalla semina alla raccolta, perché costituisce l'unica forma di irrigazione delle colture di frumento e perché può avere effetti di stress sulle coltivazioni.

Altrettanto nota è la variabilità spaziale delle caratteristiche chimiche e fisiche del suolo. Le prime sono state descritte dalla percentuale di sostanza organica in dotazione del terreno, fortemente connessa con la percentuale di azoto, di cui è nota l'influenza sul contenuto proteico del grano<sup>2</sup>. La struttura fisica del terreno dipende, invece, dalla composizione del suolo in argilla, limo e sabbia, detta tessitura. Queste due caratteristiche sono state rilevate, attraverso analisi chimiche e granulometriche, su prelievi di terreno eseguiti nel settembre 1999, ovvero prima che gli appezzamenti considerati venissero sottoposti alla semina.

Le caratteristiche pedologiche possono però essere modificate dagli interventi colturali. Ad esempio, la lavorazione del terreno può aumentare la porosità dello stesso. Si è voluto quindi distinguere tra la semina su un terreno lavorato (in qualunque modo) e quella su sodo, ovvero in condizioni di nessuna o minima lavorazione del terreno.

Tra le pratiche colturali altrettanto importante è l'apporto di fertilizzanti, del quale vanno considerati aspetti diversi. Motivazioni agronomiche suggeriscono infatti di tenere conto oltre che dell'aspetto quantitativo della concimazione, anche dei modi e dei tempi di somministrazione. Accanto alla quantità totale di azoto distribuito, si è considerato, perciò, il numero di frazionamenti in cui tale quantità è stata somministrata, e se è stata effettuata una concimazione tardiva, cioè successiva al 15 aprile. La scelta, in parte arbitraria, di questa data limite è dovuta al fatto che indicativamente dopo la metà di aprile, essendo ormai fissato il numero di foglie e di spighe, la concimazione dovrebbe influire proprio sul rapporto tra proteine e amido nella cariosside.

Tra le pratiche colturali è stato considerato anche il periodo di semina, che può avere effetti sul risultato produttivo e qualitativo.

---

<sup>1</sup> Per i dati relativi al periodo di semina e alla lavorazione pre-semina, non richiesti nelle schede agronomiche del Consorzio, sono stati interpellati gli agricoltori stessi. In tal modo è stato possibile risalire al 37,5% delle informazioni sul periodo di semina e al 16,6% di quelle sulla lavorazione del terreno. I dati relativi alle precipitazioni nei comuni delle cinque agenzie del Consorzio sono stati in parte tratti dal sito dell'Azienda Regionale Prevenzione e Ambiente (ARPA Emilia Romagna) e in parte richiesti ad Enti operanti nelle diverse zone.

<sup>2</sup> Tra tutti gli elementi presenti nel terreno, l'azoto è quello che più direttamente influisce sul livello proteico del grano. I cereali non riescono a sintetizzare l'azoto dall'aria e quindi possono assorbirlo solo dalla terra.

Sulla coltura possono inoltre influire altri organismi biologici, come la coltura precedente, che lascia il terreno più o meno deteriorato, o come insetti, microrganismi patogeni e parassiti: l'attuazione di un trattamento fungicida o pesticida potrebbe essere indicativa della loro presenza.

Poiché le interazioni tra i fattori agronomici possono avere effetti diversi su varietà diverse, si è voluto infine distinguere tra le due principali varietà di frumento tenero coltivate nell'areale considerato: Mieti e Sagittario. Di fatto, queste due varietà sono di consolidata tradizione, per resistenza alle avversità ambientali e colturali, per produzione ettariale e contenuto proteico (Barbanti *et al.*, 1995).

In sintesi, per ogni unità statistica si è osservato il contenuto proteico (espresso in quantità percentuale sul totale delle componenti della cariosside), rilevato su prelievi di qualche etto di granella con apparecchiature NIR o NIT (*Near Infrared Reflectance* o *Transmittance*), e le seguenti variabili esplicative:

- precipitazioni (in mm di pioggia da ottobre a maggio)
- sostanza organica nel terreno (in percentuale)
- tessitura del terreno (12 modalità, secondo la classificazione della G.U. n. 125, 1992)
- lavorazione pre-semine (tradizionale, sodo)
- azoto distribuito (in Kg/ha)
  - numero di frazionamenti della concimazione (da 1 a 3)
  - frazionamento tardivo (sì, no)
- periodo di semina (mesi da settembre a gennaio)
- precessione colturale (12 modalità, tra cui barbabietola da zucchero, sorgo, mais e soia)
- difesa fitosanitaria (sì, no)
- varietà coltivata (Mieti, Sagittario).

Poiché i dati proteici non erano stati rilevati con il medesimo apparecchio, si è svolto preliminarmente un controllo sperimentale sulla qualità degli strumenti di rilevazione impiegati presso le diverse agenzie. Presi dieci prelievi di granella (con valori proteici compresi tra 10% e 17%) si è misurato con ognuna delle macchine il tenore proteico di ciascun prelievo, confrontando il valore ottenuto con quello dedotto dall'analisi chimica (realizzata con il *microkjeldhal*), che costituisce un valore di riferimento. Sono emerse distorsioni positive delle varie macchine rispetto ai risultati dell'analisi chimica, giudicate statisticamente significative dal test *t* di Student per dati appaiati, come mostrato nella tavola 1.

TAVOLA 1

*Risultati del test di normalità e del test t di Student sulle differenze tra i valori del tenore proteico rilevati da ciascuna macchina e quelli dedotti dall'analisi chimica su 10 prelievi di granella.*

variabili differenza		test di normalità di Shapiro-Wilk		media aritmetica	deviazione standard	test <i>t</i> di Student $H_0 : \delta_i = 0$	
$D_i = X_i - X_0$	$n_i$	$w_i$	<i>p-value</i>	$\bar{d}_i$	$s_{d_i}$	$t_i$	<i>p-value</i> (2 code)
$X_1 - X_0$	10	0,917	0,380	0,379	0,331	3,625	0,006
$X_2 - X_0$	10	0,975	0,924	1,679	0,427	12,446	<0,001
$X_3 - X_0$	10	0,879	0,159	0,544	0,401	4,291	0,002
$X_4 - X_0$	10	0,904	0,304	0,706	0,278	8,029	<0,001

Nella tavola 1 le variabili  $X_1, \dots, X_4$  descrivono i dati proteici rilevati, nell'esperimento, da ciascuno dei quattro apparecchi impiegati presso le diverse agenzie.

Esse si riferiscono, nell'ordine, alla macchina dell'agenzia CABM di Ozzano, a quella dell'Agenzia di Crevalcore, a quella di San Felice e, infine, all'apparecchio del CRPV. La variabile  $X_0$  rappresenta invece i risultati dell'analisi chimica.

Prima di procedere, quindi, era necessario eliminare le differenze indotte dall'uso di macchine diverse. Così, i dati proteici rilevati da ogni macchina sono stati corretti sottraendo loro la stima della distorsione di quella macchina rispetto all'analisi chimica.

Inoltre non tutte le informazioni erano state reperite. In particolare, per il periodo di semina solo il 37,5% dei dati era disponibile e riguardo al tipo di lavorazione pre-semina mancavano 16 dati (il 12,5% del totale); inoltre, per 6 appezzamenti era stata indicata più di una precessione colturale, il che equivaleva a non disporre di tale dato. Non si è voluto attribuire un valore fittizio ai dati mancanti: per le prime due variabili è stata attribuita la modalità 'non disponibile' a tutti i valori non pervenuti; per i casi di precessione colturale non definita univocamente si è invece utilizzata la modalità 'mista'.

### 3. UN'ANALISI ESPLORATIVA PRELIMINARE

È stata condotta un'analisi preliminare di tipo esplorativo, che ha confortato sulla qualità dei dati, nei quali si sono ritrovate alcune forme di relazione tra fattori agronomici ben note in agronomia. Ma soprattutto ha suggerito spunti di riflessione interessanti, in particolare sulla concimazione. A tale riguardo, come si è già accennato, i cerealicoltori avevano ricevuto indicazioni da parte di tecnici competenti, basate sulla percentuale di sostanza organica in dotazione del terreno e sulla precessione colturale.

Esisteva dunque una forte e intrinseca relazione tra queste tre variabili di importanza cruciale: la sostanza organica, la coltura precedente e l'entità della concimazione effettuata. Se, nel piano di concimazione, le dosi consigliate di azoto fossero state determinate in modo da compensare perfettamente la diversa quantità di sostanza organica presente nel terreno e gli effetti delle varie precessioni colturali – ammesso che ciò sia possibile – e se gli agricoltori avessero distribuito esattamente la quantità di azoto loro consigliata, non sarebbe stato rilevato alcun effetto nella variabile risposta. In realtà le aziende agricole non hanno rispettato sempre le indicazioni ricevute, come mostra la figura 1, che illustra la relazione tra la quantità di azoto consigliata e quella effettivamente distribuita.

La concimazione effettuata segue abbastanza scrupolosamente le indicazioni fornite, fino a quando queste risultano, per così dire, contenute, mentre oltre una certa dose di azoto consigliato la fiducia nelle indicazioni ha lasciato ampio spazio ad una prudente parsimonia.

Si è voluto quindi controllare che il comportamento degli agricoltori non abbia alterato sensibilmente le corrispondenze tra precessione colturale e dosaggi medi

di azoto istituite nei piani di concimazione. Le differenze nei valori medi della concimazione consigliata, illustrati nella tavola 2, sono imputabili alla coltura precedente. Non si può pensare che siano in realtà un effetto mascherato della diversa dotazione di sostanza organica, poiché i valori medi della percentuale di sostanza organica sono di fatto risultati simili per tutte le precessioni considerate, ad esclusione del valore elevato relativo all'erba medica<sup>3</sup>, una coltura tradizionalmente considerata "miglioratrice".

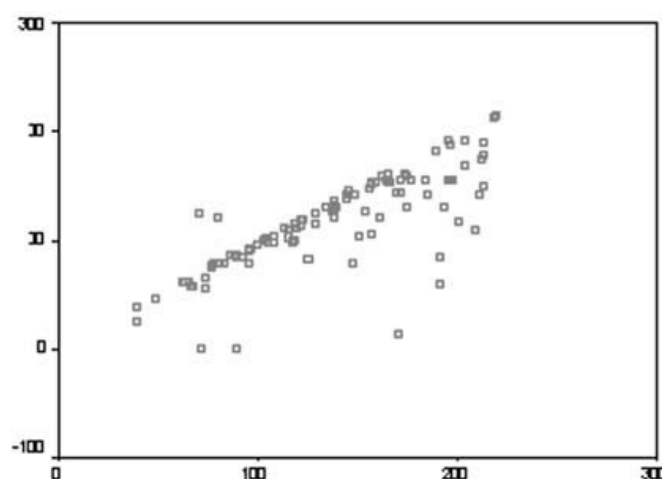


Figura 1 – Unità (Kg/ha) di azoto distribuite (in ordinata) dai cerealicoltori e unità di azoto consigliate (in ascissa).

TAVOLA 2

*Valori medi della quantità di azoto consigliata (in Kg/ha) e di quella distribuita sugli appezzamenti con la medesima precessione colturale.*

Coltura precedente	Frequenza assoluta	Unità di azoto consigliate (media aritmetica)	Unità di azoto distribuite (media aritmetica)
Barbabietola	40	95	92
Mais	35	162	139
erba medica	2	39	32
Soia	13	109	87
Sorgo	20	192	148
altre colture	17	119	117

È interessante osservare che dopo l'erba medica, sulla quale è comprensibile dunque che si concimi poco, la precessione ritenuta migliore è la barbabietola da zucchero, per la quale si consigliano in media 95 Kg/ha; seguono la soia con 109 e il mais con 162 unità consigliate in media; infine il sorgo con 192 unità. La media per tutte le altre precessioni risulta uguale a 119 Kg/ha. Il sorgo è una precessione colturale non auspicabile per il grano, non solo perché è anch'esso una graminacea, ma anche per il fatto che deteriora la struttura del suolo a causa anche delle fitte radici che compattano il terreno; un'abbondante concimazione è dunque comprensibile. Maggiori perplessità desta invece la differenza tra la concimazione

<sup>3</sup> Per questo motivo si è voluto dare rilievo alla precessione 'erba medica' decidendo di non includerla nella modalità 'altre colture' sebbene fosse presente in appena 2 unità statistiche.

consigliata dopo il mais e quella consigliata dopo la barbabietola da zucchero: l'azoto consigliato dopo la barbabietola è poco più della metà di quello consigliato dopo il mais. Per citare solo una fonte, in Crescentini e Governatori (1998) la barbabietola da zucchero e il mais vengono entrambe definite "colture da rinnovo" e vengono consigliate 130-150 unità d'azoto dopo la barbabietola e 150-160 unità dopo il mais.

Considerando finalmente le quantità di azoto realmente apportate si osserva che le differenze si riducono leggermente (coerentemente con quanto osservato a proposito del fatto che la quantità di azoto distribuita si discosta maggiormente da quella consigliata quanto più quest'ultima è alta). La differenza tra la quantità media di azoto distribuita dopo il mais e quella apportata dopo la barbabietola da zucchero sembra comunque piuttosto elevata. Dopo il sorgo sono state distribuite in media 148 unità d'azoto, che sono quasi una cinquantina in meno rispetto quelle consigliate. Probabilmente gli agricoltori avevano una loro opinione riguardo alla quantità da distribuire dopo le diverse colture, oppure non hanno distribuito volentieri dosi di azoto superiori alle 140-150 unità.

Le indicazioni tratte da questi semplici confronti hanno messo in luce aspetti importanti della relazione che lega il fattore concimazione alla precessione colturale. È stato utile tenere conto di tali considerazioni nell'interpretazione dei risultati delle analisi statistiche multivariate, i cui aspetti metodologici sono illustrati nel prossimo paragrafo.

#### 4. GLI ALBERI DI REGRESSIONE

Analizzare l'effetto delle singole variabili osservate sulla variabile risposta attraverso metodi di analisi bivariata comporta il rischio che variabili irrilevanti appaiano influenti poiché correlate con altre variabili, che sono invece dotate di un reale potere esplicativo della variabilità della risposta. La comprensione delle relazioni di dipendenza di questa dalle variabili esplicative richiede che tutte le variabili osservate siano prese in esame contemporaneamente.

Tra i metodi che permettono di studiare la relazione di dipendenza di una variabile risposta da un insieme di variabili, la metodologia più nota e impiegata è la regressione lineare multipla. Tuttavia l'ipotesi di linearità della relazione di dipendenza e le assunzioni del modello sui termini di errore ne condizionano l'impiego. Nel problema in esame la qualità dell'adattamento ai dati del modello di regressione lineare multipla è risultata scarsa. Il metodo di regressione per partizioni ricorsive (Breiman *et al.*, 1984) rappresenta un'alternativa particolarmente interessante per risolvere problemi di regressione in spazi a dimensione elevata ed eventualmente con regressori misti, capace di fornire indicazioni sulla struttura dei dati che un'analisi di regressione lineare potrebbe non rivelare. Esso si è dimostrato un utile strumento di analisi per il problema in esame, soprattutto in virtù della semplicità interpretativa dei risultati che produce e della sua flessibilità, comune a tutti i metodi non parametrici.

Indicata con  $f$  la funzione incognita che descrive la relazione di dipendenza congiunta della variabile  $Y$  dalle variabili  $X_1, X_2, \dots, X_p$  secondo un generico modello di regressione

$$Y = f(x_1, x_2, \dots, x_p) + \varepsilon \quad E(\varepsilon) = 0, \quad (1)$$

i metodi di regressione non parametrici si propongono di determinare, a partire dai dati, un'approssimazione di  $f$  senza assumere per questa una forma funzionale che possa essere specificata da un numero finito di parametri. In altre parole si ammette che  $f$  appartenga ad uno spazio di funzioni  $F$  di dimensione non finita:

$$f \in F(\theta) \quad \dim(F) = \infty, \quad (2)$$

laddove nel modello di regressione lineare multipla si suppone per  $f$  la seguente forma funzionale

$$f(x_1, x_2, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (3)$$

indicizzata da  $p+1$  parametri incogniti.

Nell'ambito dei metodi di regressione non parametrici, il modello di regressione per partizioni ricorsive più semplice, e anche più utilizzato, approssima la superficie di regressione  $f$  con una funzione  $\hat{f}$  costante a tratti, che assume valore  $a_k$  sulla regione rettangolare  $R_k$  dello spazio  $p$ -dimensionale  $D$  dei regressori:

$$\hat{f}(\mathbf{x}) = a_k \quad \text{per } \mathbf{x} \in R_k \subset D \quad k = 1, \dots, K, \quad (4)$$

dove l'insieme  $\{R_k\}$  costituisce una partizione di  $D$  in un numero  $K$ , non noto a priori, di sottoinsiemi (per una trattazione di alcune estensioni del metodo, si veda Friedman, 1991).

Il problema statistico è quello di stimare congiuntamente gli elementi-regioni  $R_k$  e i corrispondenti valori  $a_k$ . La soluzione proposta in Breiman *et al.* (1984) è di tipo ricorsivo. A partire dall'intero spazio  $D$ , e quindi da un predittore iniziale costante su  $D$ , si ricercano:

- la suddivisione dicotomica (*split*) di tale regione in due sottoregioni rettangolari (ciò impone che la suddivisione sia definita rispetto a una modalità-soglia di uno dei regressori),
- i valori delle costanti  $a_k$  corrispondenti alle suddette sottoregioni rettangolari.

Le soluzioni a questi due problemi devono essere tali che la funzione costante a tratti così definita si approssimi il più possibile alla funzione  $f$ . Tale accostamento (ovviamente stimato sull'insieme delle osservazioni) rappresenta quindi la funzione obiettivo da ottimizzare per scegliere lo *split* e i valori delle costanti  $a_k$ .

Ad ogni passo della procedura di Breiman *et al.* il problema della bipartizione è ripetuto per ciascuna delle regioni identificate al passo precedente, e il processo prosegue finché non si verificano alcune condizioni terminali (ad esempio, numero minimo di unità statistiche nella regione).

La sequenza di bipartizioni annidate così definita può essere rappresentata efficacemente mediante un albero binario: i nodi che si diramano dal nodo radice rappresentano la successione degli *split* e ciascun nodo terminale (detto foglia) corrisponde a un elemento  $R_k$  della partizione finale di  $D$ .

Se si valuta l'accostamento di  $\hat{f}$  a  $f$  in termini quadratici, la stima per risostituzione dell'errore quadratico medio sull'insieme delle  $n$  unità statistiche coinvolte nell'analisi è

$$R(\hat{f}) = \frac{\sum_{i=1}^n [y_i - \hat{f}(\mathbf{x}_i)]^2}{n} \quad (5)$$

Secondo questo criterio di accuratezza il predittore ottimale è  $E(Y|\mathbf{X} = \mathbf{x})$ : in tal caso, fissato lo *split*, il valore che la funzione costante a tratti assume in ciascuna delle due sottoregioni così identificate è dato dalla media aritmetica di  $Y$  calcolata sulle unità statistiche per le quali  $\mathbf{x}$  appartiene alla sottoregione corrispondente. Ciò comporta che ad ogni passo del procedimento ricorsivo sono prese in esame tutte le possibili bipartizioni generabili da ciascun regressore e si sceglie quella che dà luogo a sottoinsiemi massimamente omogenei al loro interno in termini di devianza entro di  $Y$  (e quindi, massimamente "distanti" tra loro in termini di devianza tra di  $Y$ ).

Resta così definita una gerarchia dei regressori rispetto al loro potere esplicativo e la natura locale dell'adattamento fa emergere con chiarezza eventuali effetti interattivi dei regressori stessi. Inoltre, l'effetto della variabile che genera uno *split* è valutato al netto delle variabili che hanno generato gli *split* precedenti.

Al crescere del numero degli elementi  $K$  nella partizione di  $D$ , la funzione (5) diminuisce sempre, al pari dell'indice di determinazione lineare  $R^2$  nel modello di regressione lineare multipla, che aumenta (o, almeno, non decresce) al crescere del numero di regressori nel modello. Per scegliere la struttura ottimale dell'albero, nella quale non compaiano gli *split* che non comportano un sensibile miglioramento nell'accostamento, Brieman *et al.* (1984) hanno introdotto un procedimento di potatura (*pruning*) che si applica all'albero espanso totalmente (secondo le regole d'arresto stabilite) e si basa sulla valutazione di una funzione di costo-complessità, nella quale si tiene conto della complessità della struttura dell'albero oltre che della sua accuratezza. Tale funzione permette di individuare una sequenza di alberi annidati, ciascuno dei quali è ottimale tra tutti gli alberi aventi il medesimo numero di nodi terminali. La scelta dell'albero ideale all'interno di tale sequenza si basa sulla stima dell'accuratezza dell'albero su un campione di *test* o attraverso la *cross-validation*.

I risultati dell'applicazione della metodologia fin qui descritta ai dati in esame sono illustrati nel prossimo paragrafo. L'elaborazione dei dati è stata svolta attraverso la procedura C&RT (acronimo di *Classification and Regression Trees*), implementata nel software statistico *AnswerTree* e basata sul procedimento di Brieman *et al.* (1984).



## 5. I RISULTATI

Una misura relativa dell'accuratezza di un albero di regressione è data dal rapporto della (5) alla varianza di  $Y$  calcolata sulle  $n$  unità statistiche in esame, ovvero dal rapporto della devianza di  $Y$  entro le foglie alla devianza totale di  $Y$ : il suo complemento a 1 è la stima per risostituzione della proporzione di varianza di  $Y$  spiegata dall'albero. Per l'albero costruito sui dati in esame il rapporto devianza entro/devianza totale è uguale a 0,385, corrispondente a poco meno del 62% di variabilità spiegata. La stima di *cross validation*<sup>4</sup> del medesimo rapporto è di poco superiore al 50%: un risultato che appare soddisfacente se si considera che tale stima può addirittura non essere minore di 1 (la stima per risostituzione è sempre minore di 1 per costruzione).

Nell'albero, illustrato in figura 2, la piovosità è risultata la variabile agronomica più rilevante per la resa proteica: lo *split* corrispondente ha prodotto il più alto decremento nella varianza entro. A piovosità maggiori corrispondono valori proteici mediamente più bassi. Evidentemente le precipitazioni, che nella stagione 1999-2000 si sono concentrate prevalentemente tra ottobre e gennaio, hanno provocato il dilavamento del terreno, causando il percolamento dell'azoto (risultato strettamente correlato alla sostanza organica) negli strati del terreno non raggiungibili dalle radici.

I due *split* immediatamente successivi, più che individuare variabili rilevanti, permettono di isolare unità statistiche portatrici di valori della variabile risposta nettamente inferiori alla media: rispettivamente 2 unità con semina dopo novembre (classi di semina maggiori del livello 3) e 2 unità con valori della sostanza organica minori di 1,025. A questo proposito si fa notare che lo sviluppo di un albero di regressione, essendo basato sui valori della somma dei residui quadratici (si veda la (5)), risulta molto sensibile alla presenza di valori anomali della variabile risposta (è invece abbastanza robusto rispetto a valori estremi delle variabili esplicative). Isolare anche pochi valori estremi di  $Y$  porta, infatti, ad una notevole diminuzione della funzione obiettivo. Per separare gli outliers un albero di regressione tende ad effettuare *split* con nodi di numerosità molto esigua e ad avere, quindi, alcuni nodi terminali composti da poche unità statistiche. Ciò permette di isolare i valori anomali in modo da segnalare la loro presenza e identificarli facilmente, minimizzandone al contempo gli effetti. Per tale ragione, gli outliers costituiscono un minor motivo di distorsione rispetto a quanto accade, ad esempio, nel modello di regressione lineare.

Si è inoltre osservato, tra le unità statistiche con valori superiori delle precipitazioni, l'effetto delle colture precedenti. Con riferimento alle precessioni più frequenti si è osservato un risultato in parte non atteso: il mais e la soia sono risultate precessioni migliori rispetto al sorgo e alla barbabietola da zucchero.

---

<sup>4</sup> Si è scelta la 10-fold *cross validation*.

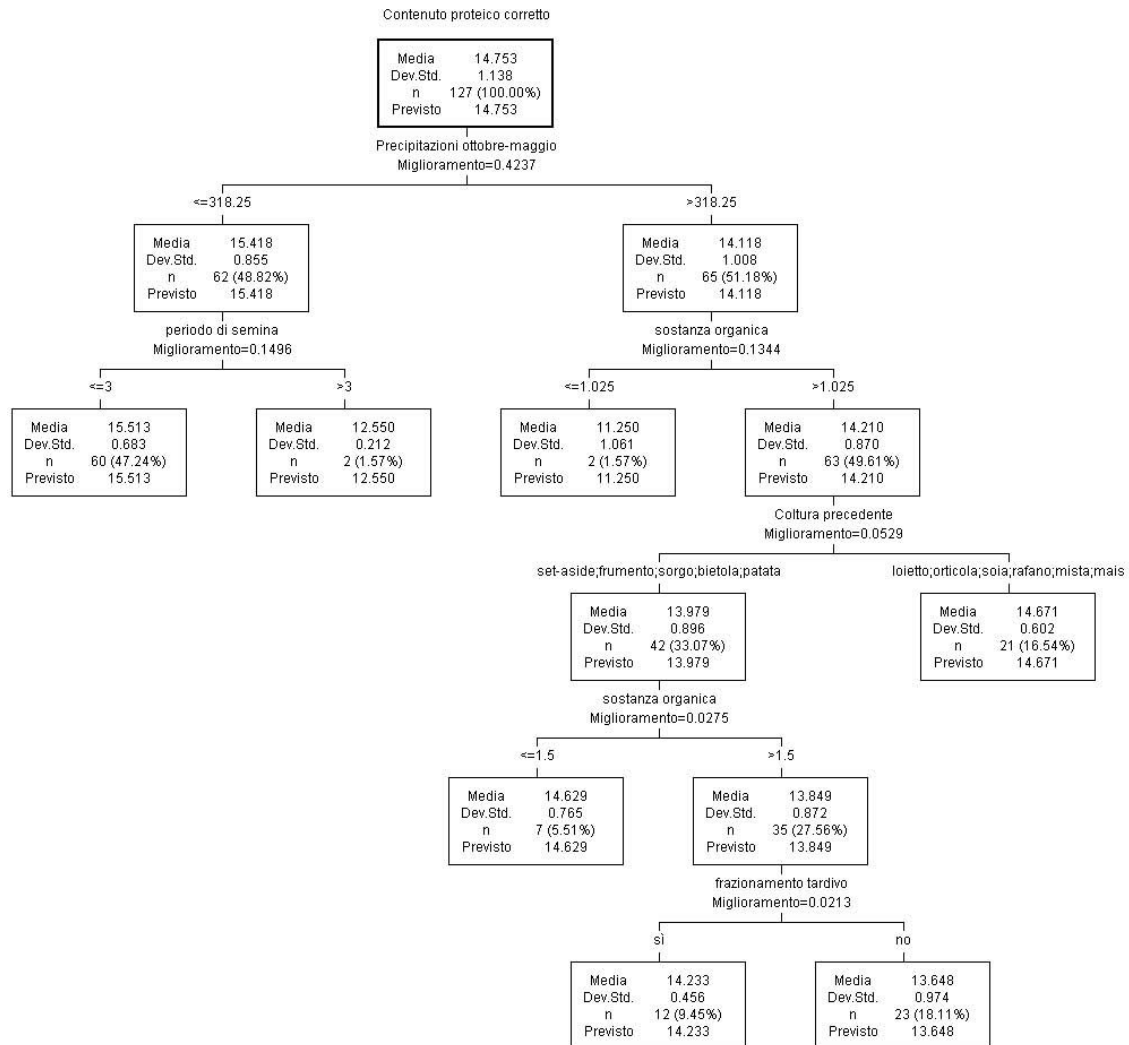


Figura 2 – L'albero di regressione ottenuto sui dati in esame.

Che il sorgo sia una coltura precedente non auspicabile non stupisce, anzi conferma le aspettative se si ricorda (si veda la tabella 2) che sui terreni a precessione 'sorgo' si erano distribuite, in media, quantità di molto inferiori a quelle consigliate. Più insolito è invece il fatto che si sia ottenuto un valore proteico maggiore dopo il mais piuttosto che dopo la barbabietola da zucchero. Una probabile spiegazione è offerta dal confronto tra le unità (Kg/ha) di azoto distribuite in media dopo queste colture: il sospetto riguardo alle quantità di azoto consigliato (e apportato) dopo queste due colture induce a ritenere che probabilmente la differenza tra le dosi di concimazione è stata tale da ribaltare i risultati attesi.

Relativamente alle unità con alti valori delle precipitazioni e colture precedenti "peggiori" si è riscontrato un tenore proteico più elevato per le unità con valori inferiori della sostanza organica. Una possibile spiegazione si ha ricordando che i valori di tale variabile erano stati rilevati in settembre, cioè prima delle piogge autunnali e invernali, e che in base ad essi (oltre che in base alla precessione colturale) era stata pianificata la concimazione, che sarebbe stata effettuata da

gennaio in poi. Al momento in cui ci si apprestava a distribuire la prima dose di azoto, però, le differenze tra gli appezzamenti rispetto alla percentuale di sostanza organica erano state in parte livellate a causa delle precipitazioni che, relativamente alle unità statistiche cui si sta facendo riferimento, erano state molto abbondanti: così sui terreni su cui si era concimato di più – in base alle stime di settembre – si è ottenuto un risultato più elevato in termini proteici e si è in parte compensato l'effetto delle condizioni climatiche sfavorevoli.

Ciò non è in contraddizione con lo *split* precedente che faceva corrispondere percentuali basse di sostanza organica a valori proteici bassi. È infatti ragionevole ritenere che percentuali così basse di sostanza organica siano rimaste sensibilmente inferiori rispetto agli altri valori anche a seguito delle precipitazioni; è comunque possibile che quello *split* sia stato indotto dalla capacità dell'albero di isolare unità con valori di  $Y$  fortemente minus-varianti.

Infine, relativamente al nodo con le unità con valori più alti di sostanza organica, viene effettuato un ultimo *split*, in base al frazionamento tardivo. Coerentemente con le aspettative, si è ottenuta una media proteica più alta dove è stata effettuata una concimazione tardiva (successiva al 15 aprile), cioè dove è stata distribuita opportunamente una dose di azoto nel periodo in cui questa avrebbe inciso prevalentemente sulla quantità di proteine accumulata nella cariosside.

## 6. CONCLUSIONI

La ricerca fin qui descritta si proponeva di approfondire lo studio della relazione tra alcune variabili agronomiche e il contenuto proteico del frumento. I dati presi in esame, tuttavia, erano stati raccolti con altre finalità. Non è stato quindi possibile determinare la numerosità campionaria, o assicurare che la rilevazione avvenisse in maniera esaustiva. D'altronde la possibilità di predisporre un disegno sperimentale si scontra con l'impossibilità di controllare alcune componenti agronomiche (è il caso, ad esempio, della sostanza organica presente nel terreno o delle precipitazioni) e con la difficoltà di immaginare un contesto in cui centinaia di appezzamenti vengano coltivati in maniera da garantire le condizioni ottimali per uno studio scientifico anziché la massima resa in termini economici. A meno che non si tratti di prove parcellari, i cui risultati però sono positivamente condizionati dall'effetto perimetro<sup>5</sup> e mostrano una maggiore sensibilità a microvariazioni territoriali. L'analisi svolta può invece vantare un'altissima percentuale di dati a pieno campo. Di contro, la mancanza di omogeneità nella misurazione del contenuto proteico presso le varie agenzie dei consorzi, ha costretto ad apportare una "correzione" preliminare ai dati di questa variabile, dalla cui validità dipendono i risultati dell'intero studio.

L'analisi multivariata volta ad individuare le relazioni di dipendenza del tenore proteico dalle variabili agronomiche è stata effettuata con l'ausilio degli alberi di regressione, che si sono dimostrati un metodo molto potente, grazie alla loro

---

<sup>5</sup> A causa del maggiore rapporto tra perimetro ed area, vi è una maggiore percentuale di spighe a bordo campo, che godono di una maggiore traspirabilità.

flessibilità e alla capacità di cogliere e far emergere con chiarezza le interazioni tra le variabili esplicative.

Le precipitazioni sono risultate il fattore che ha influito maggiormente sulla resa proteica. Per quanto perfettamente plausibile, questo risultato era tutt'altro che scontato. Difficilmente, infatti, si conosce l'ammontare delle precipitazioni annuali, per cui non è facile notare una variazione nel contenuto proteico in funzione di una maggiore o minore piovosità.

La variabile principale non è quindi controllabile in alcun modo dagli agricoltori, ma tale risultato non deve portare ad un atteggiamento fatalistico. Infatti si è visto che, probabilmente, concimazioni più abbondanti hanno portato a compensare in gran parte l'effetto della dilavazione del terreno. Analogamente, è stato possibile ottenere un ulteriore miglioramento proteico effettuando una concimazione tardiva, distribuita nel momento in cui incide prevalentemente sulla quantità di proteine accumulata nella cariosside. Un piano di concimazione adeguato, quindi, può smorzare gli effetti negativi dovuti ad una condizione climatica sfavorevole e, se correttamente collocato nel tempo, può produrre un miglioramento proteico senza che aumenti la quantità di azoto distribuita, con vantaggi da un punto di vista economico e ambientale.

Si è poi osservato che la concimazione consigliata in base alla precessione colturale dell'appezzamento non era, probabilmente, calcolata in maniera ottimale. In particolare, si è notato che la quantità di azoto prevista per i terreni a precessione colturale barbabietola risultava più bassa di quella ottimale, dal momento che, non venendo apportate abbondanti quantità di fertilizzante, si otteneva una media proteica inferiore a quella realizzata dopo precessioni colturali abitualmente considerate peggiori, ma in seguito alle quali si era concimato maggiormente.

Non si è invece potuto stabilire con certezza l'effetto della data di semina, anche a causa dell'elevato numero di valori mancanti di questa variabile.

*Dipartimento di Scienze statistiche "Paolo Fortunati"*  
*Università di Bologna*

DANIELA GIOVANNA CALÒ

*ARPA Emilia Romagna*

FABIANA SCOTTO

*Società Italiana Sementi*

STEFANO RAVAGLIA

#### RIFERIMENTI BIBLIOGRAFICI

- L. BARBANTI, C. INVERNIZZI, P. MERIGGI, F. ROSSO, B. SAVIOTTI (1995), *Fumento tenero: varietà e tecnica di impianto*, "Agronomica", 5, pp. 4-17.
- L. BREIMAN, J. FRIEDMAN, R. OLSHEN, C. STONE (1983), *Classification and regression trees*, Wadsworth and Brooks/Cole, Monterey.
- P. CRESCENTINI, C. GOVERNATORI (2000), *La buona pratica del frumento duro di qualità*, Agenzia Servizi Settore Agroalimentare delle Marche.
- J. FRIEDMAN (1991), *Multivariate adaptive regression splines (with discussion)*, "Annals of Statistics", 19, pp. 1-141.

- M. POLI, L. MAZZA, M. BARAVELLI, C. BELLONI, L. CASTELLO, S. CORNALE, S. CARUSO, A. INNOCENTI, D. TASSI, F. ZINONI (2000), *Emilia-Romagna, risultati della 27<sup>a</sup> sperimentazione nazionale*, "L'informatore agrario", 33, pp. 23-25.
- F. SCOTTO (2001), *Contenuto proteico del grano e sue determinanti agronomiche: un'analisi statistica*, Tesi di Laurea in Scienze Statistiche ed Economiche, Facoltà di Scienze Statistiche, Università di Bologna.

## RIASSUNTO

### *Contenuto proteico del grano e variabili agronomiche: un'analisi statistica*

Questo lavoro rappresenta un contributo allo studio degli effetti delle componenti agronomiche, e delle loro interazioni, sul tenore proteico del grano. Lo studio ha preso le mosse dall'osservazione di alcuni fattori agronomici sulle coltivazioni di frumento delle province di Bologna e Modena nell'annata agraria 1999-2000. L'impiego del metodo di regressione non parametrica noto con il nome di alberi di regressione ha prodotto risultati interessanti, che suggeriscono spunti di riflessione sull'interazione tra fattori climatici, caratteristiche pedologiche del terreno e interventi colturali. In particolare, si è osservato come un piano di concimazione adeguato, nelle dosi e nei tempi di somministrazione, può smorzare gli effetti negativi di precipitazioni abbondanti o di precessioni colturali sfavorevoli.

## SUMMARY

### *Wheat protein and agronomic variables: a statistical analysis*

In this paper the effect of agronomic factors, and that of their interactions, on wheat protein percentage has been studied. The analysis is based on the data collected on the wheat cultivation of Bologna and Modena provinces in the year 1999-2000 and considers the main agronomic variables. Interesting conclusions about interactions of climatic, soil and cultivation aspects have been drawn by employing the method of nonparametric regression known as regression trees. In particular, the importance of manuring has emerged: a right fertilizing plan, with respect both to the dosage time and to the dosage intensity, seems to damp the consequences of heavy rainfalls or bad previous cultivations.