Introduzione: i modelli matematici nella sperimentazione biologica

Andrea Onofri
Dipartimento di Scienze Agrarie ed Ambientali
Università degli Studi di Perugia

24 novembre 2011

Indice

1 Definizione di modello matematico

Per modello matematico si intende una rappresentazione, più o meno semplificata, di una parte della realtà, ottenuta attraverso il linguaggio della matematica.

$$E(Y) = f(X, \vartheta)$$

Con la notazione anzidetta (Laffelaar, 1993) si intende appunto che il risultato atteso per un certo fenomeno biologico (*E(Y)*) dipende da una o più variabili indipendenti (X), tramite una determinata funzione *f*, la cui forma algebrica è definita da uno o più parametri θ.

Come semplice esempio, possiamo considerare il seguente modello:

$$W_t = 0.15 + 2.5t$$

il quale esprime il concetto che il peso di una pianta al tempo t (W_t) è funzione lineare del tempo trascorso dall'emergenza, tramite i parametri $a = 0.15$ e $b = 2.5$.

Più in dettaglio, si può dire che la nostra pianta pesa 0.15 grammi al tempo 0 e cresce di 2.5 grammi al giorno, il che costrituisce un'ottima rappresentazione del fenomeno biologico in studio.

2 Cenni di storia dei modelli

La storia dei modelli convenzionalmente si fa iniziare con Galileo Galilei (1564-1642). Prima di allora, nel Medioevo, la concezione dominante prevedeva una rigida distinzione tra il mondo dei corpi celesti e il mondo della terra; il primo era assolutamente perfetto e di natura divina ed era l'unico ad essere degno di studio, mentre il secondo, in quanto imperfetto e umano, era continuamente soggetto a processi di generazione e corruzione, dei quali si tentava di dare una descrizione qualitativa, senza che però tutto ciò potesse essere elevato al rango di scienza. La vera scienza era invece nello studio del moto dei corpi celesti, che veniva fatto anche in modo quantitativo, ricorrendo, laddove necessario all'applicazione delle formule della matematica.

Nel cinquecento l'uso della matematica si va affermando anche in settori diversi dall'astronomia, ed essa viene impiegata anche in problemi di ingegneria e di artigianato (si pensi a Leonardo) e quindi a problemi legati al mondo terrestre. Il matematico del '500, tuttavia, non è ancora uno scienziato come lo intendiamo noi, ma è spesso anche un ingegnere, un costruttore, un artigiano e così via.

Si inizia quindi a rompere il diaframma tra le scienze del cielo e quelle della terra ed inoltre cominciano a venir meno alcune certezze scientifiche legate alla visione religiosa del Mondo. Ci si inizia insomma ad interrogare sulla natura nella sua interezza (cielo e terra) cercando di penetrarne i meccanismi di funzionamento e le sue leggi fondamentali verificando la validità delle ipotesi scientifiche, senza darne alcuna per scontata *a priori*. In questo quadro Galilei intuisce che è possibile fondere tra di loro le due facce della matematica, che era considerata fin dai tempi dei Greci un forma di pensiero e speculazione pura, ma che era anche già utilizzata come strumento per la soluzione di problemi pratici di natura ingegneristica.

L'intuizione di Galilei è che la natura è fondamentalmente semplice: l'apparente complessità dei fenomeni nasconde delle leggi che possono essere tradotte nel linguaggio semplice, ma logico e razionale della matematica. L'anzidetta scienza non è più quindi solo speculazione e non serve solo a risolvere problemi pratici: essa è il linguaggio per tradurre e comprendere le leggi con cui funziona la natura.

Il processo di conoscenza è quindi il seguente: (1) osservazione dei fenomeni; (2) formulazione di un ipotesi, nel linguaggio della matematica (3) verifica sperimentale dell'ipotesi.

In sostanza, la descrizione matematica dei fenomeni naturali sostituisce una descrizione qualitativa; nasce il modello matematico come rappresentazione della *verità* e delle leggi fisiche. Nel concreto, la verifica di un modello matematico astratto è fondamentale, ma deve essere fatta con cognizione di causa, *defalcando gli impedimenti della materia*.

Se Galilei ha posto le basi per l'uso dei modelli matematici nella scienza, è stato Isaac Newton (1642-1727) che è riuscito a completare la fusione del mondo celeste e quello terrestre, attraverso la legge della gravitazione universale, che è appunto uno schema descrittivo matematico (un modello) suggerito dall'esperienza, che spiega il moto dei corpi, siano essi celesti che terrestri. Ci è possibile anche grazie all'introduzione di strumenti matematici adeguati, come il calcolo differenziale.

La cosa interessante è che il costrutto formale che ne risulta è tale da permettere di prevedere quello che succederà in un certo sistema di corpi in quiete o in moto in un qualunque momento del futuro o che cosa è successo in qualunque momento del passato, semplicemente conoscendo lo stato iniziale del sistema e le forze che su di esso agiscono. Si va insomma facendo strada il principio del determinismo di cui sono intrisi tutti gli studi della meccanica classica: lo stato passato e futuro di un sistema è completamente determinato dal suo stato iniziale e dalle forze che su esso agiscono; non sono quindi possibili deviazioni di tipo casuale od aleatorio.

Questa idea di una legge matematica che governa la natura è molto affascinante e contagia gli scienziati dell'epoca, che con Laplace (1749-1827) arrivano ad estendere il principio del determinismo a tutti i fenomeni meccanici e addirittura a tutti i fenomeni naturali (meccanicismo o riduzionismo meccanicista). Se qualche legge non può essere trovata è solo per ignoranza umana, non perché essa non esista.

In realtà questa concezione è destinata a scontrarsi con la realtà e altri scienziati osservano che molti fenomeni fisici non possono essere spiegati con le leggi matematiche del moto dei corpi. Ma soprattutto subentrano altre teorie scientifiche (relatività, legge dei quanti, meccanica quantistica) che presentano fenomeni alla base dei quali non vi è più una certezza deterministica, ma addirittura una definizione probabilistica.

Tutto questo inizia a mettere in crisi il concetto classico di modello matematico come rappresentazione della realtà di un fenomeno scientifico. La crisi è ancora più evidente quando negli anno '20 e '30 si inizia ad applicare modelli matematici alla biologia e all'economia, due settori scientifici altamente complessi, che riguardano sistemi fondamentalmente aperti, con notevoli influenze esterne difficili da quantificare e conoscere con dettaglio.

Insomma il modello matematico non può più essere considerato come la spiegazione della verità sull'Universo; le leggi unitarie non sono più rintracciabili anche perchè la scienza diviene così complicata da richiedere la specializzazione dei ricercatori in settori sempre più ristretti.

A questo punto la modellistica matematica viene investita di un ruolo nuovo: essere l'immagine di un settore della realtà, del quale rappresenta uno schema ed una descrizione (non quindi lo schema o la descrizione, della realtà non si pretende di sapere nulla!) la cui validità va ricercata unicamente sulla sua capacità di funzionare in un intervallo di condizioni sufficientemente ampio. Quindi non ha nessuna importanza il fatto che un modello per analogia sia applicato a fenomeni totalmente diversi: non ci interessa la verità, ma solo come il modello funziona!

Inoltre, dato che non è più necessario che il modello rifletta l'armonia e la semplicità dell'Universo (che non è affatto semplice, come credeva Galilei!), è possibile ricorrere anche a nuove costruzioni matematiche meno semplici come le reti neurali, i frattali e i modelli stocastici.

3 Modelli deterministici e stocastici: variabilità casuale

Il modello indicato in precedenza nel primo paragrafo è fondamentalmente di tipo deterministico, in quanto descrive una relazione causa-effetto, in assenza di ogni tipo di disturbo o evento perturbativo. In realtà, i fenomeni biologici sono estrememente complessi e soggetti ad una serie innumerevole di perturbazioni, processi ignoti e/o variabilità casuale. Quest'ultima è fondamentalmente riconducibile ai seguenti aspetti.

- Errore di misura. Dipende dal fatto che gli apparecchi di misura non sono mai assolutamente precisi. Attenzione! Non bisogna confondere l'errore casuale (che tende a presentarsi sempre, ma con segni opposti) dall'errore sistematico, dovuto ad una cattiva taratura dello strumento, che tende a presentarsi sempre con lo stesso segno.
- Variabilità intrinseca. Si pensi al lancio della moneta: su 10 lanci posso ottenere 4, 5 o 6 teste, in modo casuale, anche se la moneta è sempre la stessa. In questo caso la variabilità del risultato è intrinsecamente legata ad un processo che è di natura stocastica.
- Variabilità o irregolarità degli individui da misurare.
- Variabilità ambientale.

Il modellista può avere l'esigenza di includere nel modello anche una componente stocastica (casuale), in genere rappresentata con ϵ. Il modello completo è quindi:

$$E(Y) = f(X, \vartheta) + \varepsilon$$

4 Prerogative di un modello, finalità e tipologie

Ogni modello matematico deve avere almeno due prerogative fondamentali:

- Correttezza. Con questo non si intende che il modello deve dare un risultato perfetto, ma semplicemente che le previsioni sono accettabili in relazione agli obiettivi prefissati (*All models are wrong, but some are useful.* George Box).
- Semplicità (*Entia non sunt multiplicanda praeter necessitatem.* Principio del rasoio di Occam).

A parte le prorogative fondamentali prima espresse, è chiaro che i modelli sono molto diversi tra di loro, a seconda delle finalità con cui vengono costruiti. Distinguiamo Decoursey (1992):

- Ricerca e studio
  - Sappiamo abbastanza di un processo solo se lo sappiamo modellizzare
  - Se non lo sappiamo modellizzare, possiamo identificare quali parti del processo dobbiamo studiare meglio
- Screening
  - Possiamo produrre una quantità elevata di dati, a baso costo
- Decision making (Expert systems)
  - Organizzazione della conoscenza, che viene resa più fruibile

I modelli di ricerca sono in genere molto complessi e descrivono una porzione, talvolta molto ridotta, di un sistema. Permettono di prevedere, anche in modo molto preciso, l' effetto di perturbazioni esterne al sistema stesso. Siccome indagano su funzioni fisiologiche fondamentali, richiedono molti *inputs* e forniscono un *output* numerico (per es. la concentrazione di erbicida in un determinato istante e strato di terreno), ma richiedono supporti informatici molto complessi e tempi di elaborazione piuttosto lunghi. Per questo possono essere utilizzati solo da personale altamente qualificato per scopi di ricerca. Non necessariamente forniscono risultati di impiegabilità pratica e richiedono una attenta validazione in condizioni diverse da quelle in cui sono stati definiti. Inoltre, non bisogna dimenticare che ogni modello teorico è valido solo fintanto che la teoria su cui è basato è valida!

Al contrario, i modelli di screening sono utilizzati per prevedere l' andamento di un fenomeno in diversi siti sperimentali, classificare questi ultimi in base a quanto osservato in relazione al fenomeno in studio ed individuare, eventualmente, quali siano quelli in cui sono necessarie ulteriori ricerche o monitoraggi in pieno campo. Questi modelli sono molto semplici, richiedono pochi *inputs* e non forniscono generalmente un *output* numerico, ma solo una descrizione della probabile risposta (per esempio alta, media o bassa).

I modelli a supporto di sistemi esperti si collocano in posizione intermedia. Sono utilizzati per prevedere la risposta del sistema ad un cambiamento del suo stato o delle variabili che lo governano e in genere forniscono un output numerico di precisione sufficiente agli scopi per cui il monitoraggio viene eseguito. Il principio guida è che il modello non dovrebbe mai avere una complessità maggiore di quella richiesta. All' interno di questa categoria si trovano anche i modelli utilizzati o utilizzabili per l' assistenza tecnica, in un quadro di lotta integrata alle malerbe. La caratteristica fondamentale di questi ultimi dovrebbe essere non tanto la precisione, che deve essere peraltro sufficiente, ma soprattutto la facilità d' impiego e la possibilità di funzionare in base a parametri rilevabili facilmente e precisamente. Siccome debbono essere utilizzati in pieno campo, essi debbono essere stati oggetto di una attenta validazioni in numerose situazioni sperimentali diverse tra di loro (Donigian and Carsel, 1992).

Dalla classificazione precedente, risulta chiaro che, a seconda della finalità, ci troveremo a costruire modelli matematici molto diversi, comunque riconducibili a due tipologie fondamentali: empirici/fenomenologici e meccanicistici/causali (Decoursey, 1992; Hunt, 1982; Thornley, 1976).

I modelli empirici sono generalmente basati su semplici osservazioni della realtà: si individua una grandezza fisica la cui variazione provoca la modifica del livello di un'altra grandezza fisica, si deriva una legge matematica che descrive sufficientemente bene il fenomeno (cioè che ha la 'forma' giusta) e si utilizza questa funzione per la descrizione dei dati sperimentali. I modelli empirici possono essere **dinamici** o **statici** a seconda che la dimensione tempo sia presa in considerazione o no, ma, comunque, **essi risultano sempre descrittivi**: mostrano cioè l' esistenza di relazioni tra gli elementi del sistema, ma non forniscono informazioni sui processi che determinano l' effetto.

Un semplice modello empirico è la funzione matematica esponenziale che descrive l' incremento del peso secco di una pianta nel tempo. Questa funzione è basata sulla semplice osservazione che, fino ad una certa fase del ciclo biologico, il tasso relativo di crescita assume un valore costante nel tempo; essa descrive il processo (accrescimento), ma non spiega alcuno dei meccanismi fisiologici fondamentali che lo governano. I modelli empirici sono molto semplici, possono essere sviluppati senza grosse conoscenze scientifiche e, se utilizzati a fini predittivi, richiedono generalmente pochi *inputs* risultando quindi di facile impiego. Gli svantaggi risiedono nel fatto che le predizioni basate su di essi possono lasciare ampi margini di incertezza, in quanto trattandosi di modelli descrittivi, i parametri hanno poco significato se non nelle condizioni sperimentali in cui sono stati determinati.

I modelli meccanicistici, come dice il nome stesso, descrivono e spiegano un determinato fenomeno in base ai meccanismi fondamentali che governano il funzionamento del sistema. Per esempio, l' incremento del peso secco di una pianta può essere descritto con una serie di funzioni più complesse, ognuna delle quali tiene conto di sottoprocessi più piccoli, come l' influenza delle caratteristiche meteorologiche ed ecofisiologiche della specie vegetale sull' intercettazione luminosa, sull' andamento del processo fotosintetico, sulla produzione di assimilati e, quindi, sull' incremento del peso secco. Il modello che ne risulta è piuttosto complesso, ma, almeno teoricamente in grado di prevedere la crescita di una pianta qualunque siano le condizioni ambientali (si veda, per esempio, Ittersum et al., 2003).

Anche i modelli meccanicistici possono essere statici o dinamici; un esempio di modello statico è quello in cui la distribuzione della luce sul canopy vegetale viene calcolata in dipendenza dell' architettura del canopy stesso, della radiazione riflessa e trasmessa dalle foglie, dalla posizione del sole e dalle condizioni luminose in genere. I modelli meccanicistici dinamici considerano invece la dimensione temporale e tendono a simulare l' andamento di un intero sistema. Per questo motivo sono spesso definiti modelli di simulazione. Un esempio tipico è DAISY (Abrahamsen and Hansen, 2000).

Il dibattito tra gli autori sull' impiego di modelli empirici o meccanicistici è estremamente vivace, anche se è necessario menzionare il fatto che una distinzione netta non è sempre possibile. Per decidere quale sia il modello più adeguato da utilizzare in una determinata situazione è necessario individuare quali sono le finalità che il modello stesso si propone.

5 Lavorare con i modelli matematici

Per lavorare con un modello matematico è richiesto uno protocollo abbastanza chiaro, che dovrebbe essere sempre tenuto presente:

- Identificare con esattezza la 'biological question'.
- Identificare il modello deterministico
  - Scelta delle variabili esplicative.
  - Scelta del modello matematico (funzione).
- Identificare il modello stocastico (per ϵ). In genere si utilizza un'opportuna distribuzione di probabilità, molto spesso gaussiana.
- Parametrizzazione del modello

- Verifica della bontà descrittiva del modello.
- Validazione della capacità predittiva del modello (se necessario).

Vediamo un semplice esempio:

Immaginiamo di voler modellizzare la degradazione di un erbicida nel terreno, per determinare la sua semivita (questione biologica). Per modellizzare questo fenomeno biologico scegliamo di utilizzare la concentrazione dell'erbicida come variabile dipendente e il tempo come variabile esplicativa. Per quanto riguarda la forma della funzione, assumiamo di poter utilizzare una cinetica lineare sul logaritmo della concentrazione (cinetica del primo ordine) e valutiamo che eventuali deviazioni da questa cinetica siano perfettamente casuali e gaussiane (errori positivi e negativi si alternano con ugual frequenza e segno opposto intorno al valore medio atteso):

$$W_t = a + bt + \epsilon$$

dove

$$\epsilon \sim N(0, \sigma)$$

6 Parametrizzazione di un modello

Parametrizzare il modello significa individuare i valori per a, b e σ, il che può essere fatto utilizzando:

- informazioni in letteratura;
- mediante esperimenti 'ad hoc'
- secondo l'esperienza locale

Il secondo aspetto è fondamentale. Infatti possiamo organizzare un esperimento nel quale contaminiamo un campione di terreno con l'erbicida in studio e ad intervalli regolari valutiamo la sua concentrazione residua, derivando quindi il valore dei parametri del modello direttamente da questo impianto sperimentale.

Ciò richiede metodiche di 'fitting', con le quali il modello proposto viene 'adattato' ai dati, individuando i valori dei parametri che permettono il minimo scostamento tra l'output del modello e i valori osservati (metodiche dei 'minimi quadrati') o meglio che massimizzano la verosimiglianza del modello in relazione ai dati osservati (metodiche di 'massima verosimiglianza'). Le due metodiche sono perfettamente analoghe solo quando la componente stocastica è perfettamente normale con varianza omogenee ed indipendenti dal livello della variabile esplicativa.

Dato che gli esperimenti sono comunque soggetti ad errore, le stime dei parametri saranno comunque caratterizzate da un certo grado di incertezza che dovrà necessariamente essere rappresentato. In particolare, di queste stime ci interessano in particolare due aspetti:

- Precisione
- Accuratezza

La precisione misura quanto è affidabile la stima di un parametri (appunto precisa), mentre la seconda misura quanto la stima si avvicina al valore vero ed incognito di d. L'accuratezza è molto più importante della precisione: infatti una misura accurata, ma imprecisa riflette bene la realtà, anche se in modo vago, mentre una misura precisa ma inaccurata ci porta completamente fuori strada, perchè non riflette la realtà!

Accuratezza e precisione concorrono a definire il 'bias' di uno stimatore, così come quello di un modello (che è anche esso uno stimatore della realtà).

7 Valutazione del modello

La valutazione di un modello viene effettuata confrontando le sue stime con i valori osservati nella realtà. Questa valutazione si esende ad entrambe le componenti di un modello (deterministica e stocastica) e può essere grafica oppure numerica (ci torneremo in seguito). E' bene anticipare subito che la valutazione del modello può essere semplicemente sulla sua capacità descrittiva oppure anche sulla suo valore predittivo.

La valutazione descrittiva del modello viene fatta contro i dati utilizzati per la sua parametrizzazione. Nel caso dell'esempio precedente, se io valuto la mia cinetica degradativa contro i dati dello stesso esperimento che mi è servito per parametrizzare il modello, esprimo solo un giudizio su come il modello rappresenta un 'passato' specifico, ma non sono in grado di dire nulla su come il modello prevede il 'futuro'.

A quest'ultimo fine è fondamentale valutare il modello contro dati indipendenti da quelli utilizzati per la sua parametrizzazione (validazione di un modello). E' necessario diffidare di qualunque modello non sia stato attentamente validato nelle condizioni nelle quali esso viene utilizzato per generare previsioni!

8 Modelli matematici e metodologia sperimentale

Come vediamo, abbiamo introdotto il legame tra modellazione matematica e metodologia sperimentale: anche se il modello matematico può essere visto come un fatto puramente concettuale, il ruolo degli esperimenti è fondamentale sia per la parametrizzazione del modello, che per la sua valutazione. E' però fondamentale che la raccolta dei dati sia precisa ed accurata, per evitare che la valutazione di un modello sia compromessa da una scarsa qualità dei dati ('*garbage in, garbage out*'). La metodologia sperimentale viene quindi inquadrata da questo punto di vista, come la scienza che consente di organizzare esperimenti corretti da un punto di vista scientifico, che massimizzino precisione ed accuratezza delle stime.

9 Domande e spunti di verifica

1. Quali sono le fonti di variabilità (errore sperimentale)? Quali sono quelle più 'pericolose'?
2. Dare una definizione di modello matematico e fare un esempio. Avete più fiducia di un modello stocastico o di uno deterministico?
3. Illustrare le prerogative fondamentali di un modello.
4. Quale è la procudura esatta di lavoro, quando utilizziamo un modello?
5. Illustrare la definizione di precisione ed accuratezza di uno stimatore. E'meglio uno stimatore solo accurato o uno solo preciso?
6. Che cosa si intende per *bias*?

Riferimenti bibliografici

Abrahamsen, P., Hansen, S., 2000. Daisy: an open soil-crop-atmosphere system model. Environmental modelling & software, 313–330.

Decoursey, D., 1992. Developing models with more detail: do more algorithms give more thruth? Weed Tecnology 6, 709–715.

Donigian, A., Carsel, R., 1992. Developing computer simulation models for estimating risks of pesticide use: research vs. user needs. Weed Technology, 6, 677-682.

Hunt, R., 1982. Plant growth curves. the functional approach to plant growth analysis. Edwaed Arnold Limited.

Ittersum, M., Laffelaar, P., van Keulen, H., Kropff, M., Bastiaans, L., Goudrian, J., 2003. On approaches and applications of the wageningen crop models. European Journal of Agronomy 18, 201–234.

Laffelaar, P., 1993. Basic elements of dynamic simulation. Kluwer Academic Publishers, Doordrecht, The Netherlands, pp. 11–27.

Thornley, J., 1976. Mathematical models in plant physiology: a quantitative approach to problems in plant and crop physiology. Academic Press, London.