

UNIVERSITÀ DEGLI STUDI DI PADOVA
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in Scienze Statistiche

Analisi Statistica della crescita algale
nella produzione di biocombustibili tramite modelli non lineari

Relatore:

Ch.mo Prof.
Antonio Canale

Laureanda:

Giulia Poidomani
Matricola: 2110500

Indice

Introduzione	5
1 Inquadramento biologico e dati	7
1.1 Il cianobatterio Nostoc PCC 7120	7
1.2 Condizioni sperimentali	8
1.3 Il dataset	10
2 Modelli non lineari	15
2.1 Modello sulle singole repliche	15
2.2 Modello globale	21
2.3 Aspetti computazionali	22
3 Risultati	29
3.1 Risultati sulle singole repliche	29
3.2 Risultati del modello globale	32
3.3 Test congiunti sull'effetto globale delle variabili	36
Conclusioni	42
Bibliografia	43
A Codice R	45

Introduzione

L'elaborato nasce in seguito alla necessità di analizzare dati sperimentali relativi al cianobatterio azotofissatore *Nostoc PCC 7120*, una risorsa biotecnologica di notevole interesse per la produzione di biomasse di terza generazione, che permettono una produzione di biocombustibili indipendente dai fertilizzanti chimici, riducendo drasticamente l'impatto ambientale e i costi energetici del processo (Lucato et al., 2024b, 2025). Per rendere economicamente vantaggiosa e ottimizzare la produzione di tale biomassa è necessario comprendere come le condizioni ambientali e nutrizionali ne influenzino la resa e la velocità di crescita.

In quest'ottica, il presente lavoro di tesi si propone di modellare la crescita del cianobatterio *Nostoc PCC 7120* sotto diverse condizioni sperimentali, che comprendono l'intensità luminosa, il fotoperiodo (durata dell'esposizione giornaliera alla luce) e la concentrazione di fosforo nel mezzo di coltura.

L'analisi si svolge mediante modelli statistici non lineari, assumendo una funzione di crescita di tipo logistico, capace di catturare la tipica forma di crescita della biomassa.

L'approccio prevede inizialmente un'analisi esplorativa dove il modello viene stimato sulle singole condizioni sperimentali, per poi convergere verso una modellazione globale introducendo le condizioni come covariate, in modo da valutarne l'effetto congiuntamente, sia sulla velocità di crescita sia sulla resa finale. Una parte della trattazione è dedicata anche alle sfide computazionali emerse durante la fase di stima causate da problemi di ottimizzazione numerica.

Infine la ricerca prevede l'esecuzione di test di verifica d'ipotesi per valutare la significatività dei parametri stimati, sia singolarmente sia in blocchi. In questo modo si identificano le variabili che influenzano maggiormente la crescita del cianobatterio *Nostoc PCC 7120*.

Capitolo 1

Inquadramento biologico e dati

La ricerca di alternative sostenibili ai combustibili fossili rappresenta oggi una priorità assoluta per mitigare i cambiamenti climatici, riducendo le emissioni di CO_2 e ottimizzando l'uso delle risorse. In questo scenario, alcune microalghe emergono come una risorsa biotecnologica di notevole interesse per la produzione di biomasse di terza generazione, impiegabili come materia prima per la produzione di biocombustibili sostenibili. Si tratta di cianobatteri azotofissatori, capaci di convertire l'energia solare, l'acqua e l'anidride carbonica in biomassa attraverso la fotosintesi ossigenica (Lucato et al. (2024a,b)). Uno dei cianobatteri di maggiore interesse è Nostoc PCC 7120, oggetto di studio del presente elaborato.

1.1 Il cianobatterio Nostoc PCC 7120

Secondo quanto riportato da Lucato et al. (2024b), una delle caratteristiche che contraddistingue Nostoc PCC 7120 è la capacità di differenziare due tipi di cellule, ciascuno specializzato nello svolgimento di specifiche funzioni metaboliche e strettamente interdipendenti tra loro: le cellule vegetative, responsabili della fotosintesi ossigenica e della fissazione del carbonio, e le eterocisti, dedicate alla fissazione dell'azoto atmosferico (N_2). Dunque questo cianobatterio adotta una separazione dei due processi metabolici: la fotosintesi avviene nelle cellule vegetative, mentre la fissazione dell'azoto è confinata alle eterocisti. Tale

organizzazione cellulare consente lo svolgimento simultaneo dei due processi e favorisce una crescita della biomassa anche sotto illuminazione continua.

L'impiego di Nostoc PCC 7120 nella produzione di bioenergia presenta diversi vantaggi rispetto alle colture energetiche convenzionali. In primo luogo, la capacità di fissare l'azoto atmosferico trasformandolo in composti organici complessi, come proteine e biopolimeri, può ridurre significativamente, o in alcuni casi eliminare, il ricorso a fertilizzanti azotati di sintesi, la cui produzione richiede un elevato consumo energetico ed è responsabile di rilevanti emissioni di gas serra.

Inoltre, la biomassa prodotta presenta un potere calorifico inferiore (LHV) di circa 12,28 MJ/kg (Lucato et al., 2024a) e, a seconda delle condizioni di coltivazione, può accumulare elevate quantità di carboidrati, fino a circa il 40% della biomassa secca, rendendola una materia prima promettente per la produzione di bioetanolo e biogas.

Un ulteriore vantaggio è rappresentato dalla possibilità di coltivare questo microrganismo senza impiegare suoli agricoli fertili e utilizzando, in determinate condizioni, acque saline o reflue, limitando così la competizione con le produzioni destinate all'alimentazione umana (Lucato et al., 2024a, Lucato et al., 2024b).

1.2 Condizioni sperimentali

I dati analizzati in questo lavoro derivano da una campagna sperimentale, condotta per valutare l'influenza di alcuni fattori sulla resa e sul tasso di crescita di Nostoc PCC 7120. Come riportato in Lucato et al., 2025, la coltivazione ai fini sperimentali prevede diverse fasi:

1. **Mantenimento:**

Questa prima fase serve a mantenere il ceppo vitale e puro nel tempo. Il microrganismo viene mantenuto in condizioni asettiche, all'interno di flaconi contenenti il terreno di coltura *BG11₀* (privo di fonti di azoto combinato), incubati a 25°C, con un'intensità luminosa di $70 \mu\text{mol m}^{-2} \text{s}^{-1}$ di fotoni forniti sottoforma di luce bianca per 16 ore al giorno e sottoposti ad agitazione orbitale.

2. **Pre-coltura:**

Successivamente avviene la fase di pre-coltura, partendo dalle colture di mantenimento, per preparare la biomassa necessaria all'esperimento vero e proprio. In questa

fase le condizioni vengono intensificate per favorire una crescita più rapida e adattare le cellule prima del trasferimento nel sistema di coltivazione ad alta densità. Questa fase viene effettuata in fotobioreattori a colonna di bolle da 1 L, dove viene fornita un'illuminazione continua ($150 \mu\text{mol m}^{-2} \text{s}^{-1}$) mediante una combinazione di LED rossi e blu. Qui l'apporto di carbonio inorganico e la miscelazione sono garantiti dal gorgogliamento di aria atmosferica e da un agitatore magnetico. Il terreno *BG11*₀ utilizzato è formulato eliminando il nitrato di sodio e sostituendo il citrato di ammonio ferrico con cloruro di ferro (III). Inoltre il pH viene stabilizzato tra 7,5 e 8,0 tramite l'aggiunta di 250 mg L^{-1} di bicarbonato di sodio.

3. Sistema di coltivazione ad alta densità (HD):

Sebbene la produzione industriale di microalghe avvenga prevalentemente in vasche aperte (open ponds) o fotobioreattori chiusi convenzionali, questi sistemi soffrono spesso di limitazioni dovute all'attenuazione della luce e a scambi gassosi inefficienti. I sistemi di coltivazione ad alta densità (HD), come la piattaforma CellDEG, sono progettati per superare tali ostacoli.

La sperimentazione principale è condotta sulla piattaforma HDC 9.100 (CellDEG GmbH) utilizzando coltivatori HD100 da 100 mL. Il sistema si avvale di membrane idrofobiche non porose che permettono un approvvigionamento di CO_2 di tipo bubble-free (senza bolle), minimizzando le perdite di gas e l'accumulo di ossigeno fotosintetico. Per ottimizzare l'esposizione luminosa e ridurre la fotolimitazione, il reattore è progettato con un percorso ottico inferiore a 1 cm. La luce viene emessa in modo da arrivare perpendicolarmente verso le colture, tramite lampade LED bianche. L'apporto di anidride carbonica è gestito da un'unità di controllo della crescita (GCU) collegata alla piattaforma, che mantiene un rapporto CO_2 -aria costante all'1,5% all'interno di un circuito gassoso separato dall'atmosfera.

Le condizioni ambientali all'interno della piattaforma sono state variate sistematicamente nelle seguenti modalità:

- Intensità luminosa (I): testata a 90, 280 e $470 \mu\text{mol m}^{-2} \text{s}^{-1}$, tramite LED a luce bianca con emissione perpendicolare alle colture.
- Fotoperiodo (D): cicli giornalieri di luce di 12, 18 e 24 ore.

- Concentrazione di fosforo (P): livelli iniziali a 1, 25,5 e 50 $mg_P L^{-1}$.

La temperatura è stata mantenuta costante a 24 ± 2 °C durante l'intera durata dei test. Gli esperimenti sono stati eseguiti in modalità batch per una durata di 6 giorni, con un volume di lavoro di 110 mL e un inoculo iniziale di $0,35 \pm 0,10$ $g_X L^{-1}$. Il monitoraggio quotidiano della densità ottica (OD) è stato effettuato tramite misure a 750 nm (Lucato et al., 2025).

1.3 Il dataset

Il dataset oggetto di studio presenta 371 osservazioni relative alle rilevazioni di densità ottica (OD) al tempo 0 e ai successivi 6 istanti temporali. Le variabili di cui si dispone sono 5:

- Densità ottica (OD)
- Intensità luminosa (I)
- Fotoperiodo (D)
- Concentrazione di fosforo (P)
- Istante temporale (Tempo)

L'obiettivo d'analisi è quello di comprendere se, e che tipo, di influenza hanno I, D e P sulla crescita algale; la densità ottica (OD) è considerata la variabile risposta (misurata come descritto nella Sezione 1.2). Si tratta di una variabile quantitativa continua che nel dataset assume valori in un range tra 0,25 e 17,68. È possibile osservarne la distribuzione per ogni istante temporale nella Figura 1.1, in cui si nota come i valori di densità ottica tendono ad aumentare e ad aver maggiore variabilità all'avanzare del tempo.

Le variabili I, D e P assumono solo 3 valori, come definito nella Sezione 1.2. La loro combinazione identifica una specifica condizione sperimentale, ma non tutte le combinazioni possibili sono state sperimentate. Inoltre ogni condizione sperimentata è stata replicata più di una volta, ma non tutte lo stesso numero di volte. In totale si dispone di 17 condizioni sperimentali e 53 repliche, tutte indipendenti tra loro. Nella tabella 1.1 è possibile visionare nel dettaglio lo schema sperimentale.

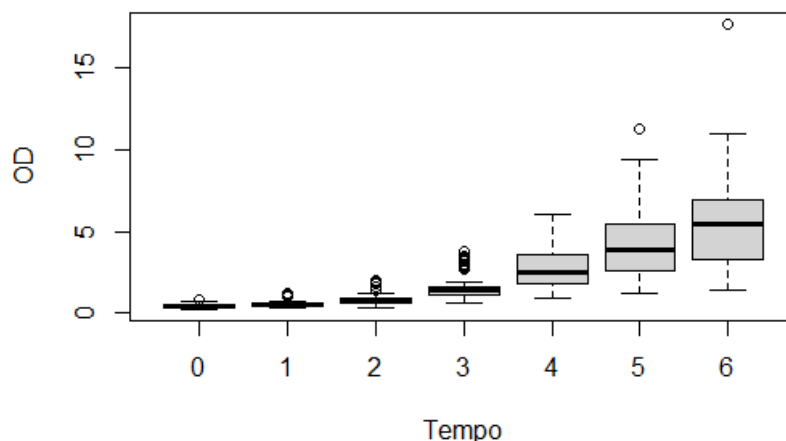


Figura 1.1: Boxplot della densità ottica (OD) al variare degli istanti temporali.

Si osservi che 2 osservazioni presentano valori mancanti di densità ottica (OD), entrambe relative all'ultimo istante temporale d'osservazione. Essendo mancante il valore relativo alla variabile risposta, queste due osservazioni vengono eliminate dal dataset. Inoltre vengono create le variabili che identificano la condizione sperimentale e il numero di replica. Alla luce di ciò il dataset finale presenta 369 osservazioni e 7 variabili.

Si effettua una prima analisi esplorativa in cui si osservano i valori di densità ottica (OD) al variare delle variabili I, D e P, escludendo le osservazioni al tempo 0. (Figura 1.2). La densità ottica sembra aumentare all'aumentare dell'intensità luminosa; mentre sembra non aumentare all'aumentare del fotoperiodo e della concentrazione di fosforo oltre i loro valori intermedi. Infine si visualizza l'andamento di crescita della biomassa per ogni replica (Figura 1.3).

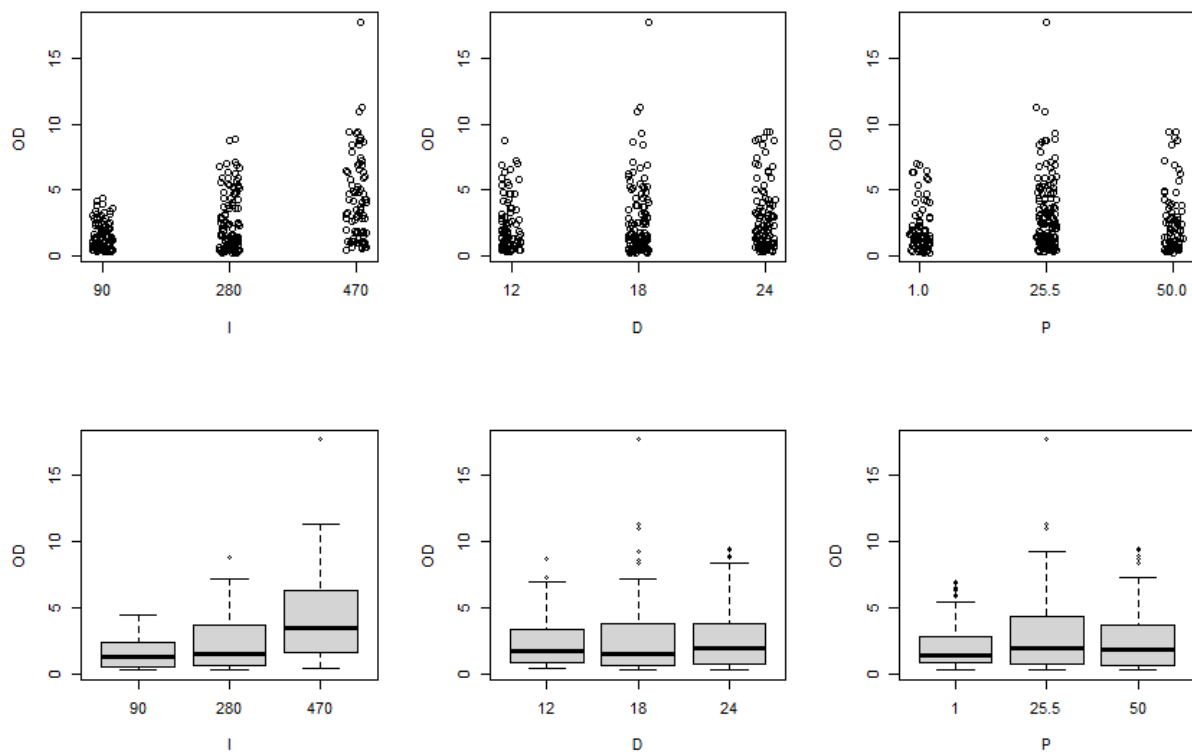


Figura 1.2: Grafici di dispersione e boxplot della densità ottica (OD), escluse le osservazioni al tempo 0, al variare rispettivamente delle variabili I,D e P.

Nota. I valori delle variabili I, D e P sono stati leggermente traslati sull'asse orizzontale tramite jittering per migliorarne la leggibilità.

Tabella 1.1: Numero di repliche per ciascuna condizione sperimentale.

Condizione sperimentale	Intensità luminosa (I) ($\mu mol\ m^{-2}\ s^{-1}$)	Fotoperiodo (D) (h)	Concentrazione di fosforo (P) ($mg_P L^{-1}$)	Numero di repliche
A	90	12	1.0	3
B	90	12	50.0	3
C	90	18	25.5	3
D	90	24	1.0	3
E	90	24	25.5	3
F	90	24	50.0	3
G	280	12	25.5	4
H	280	18	1.0	3
I	280	18	25.5	8
J	280	18	50.0	3
K	280	24	25.5	3
L	470	12	1.0	3
M	470	12	50.0	2
N	470	18	25.5	3
O	470	24	1.0	2
P	470	24	25.5	2
Q	470	24	50.0	2

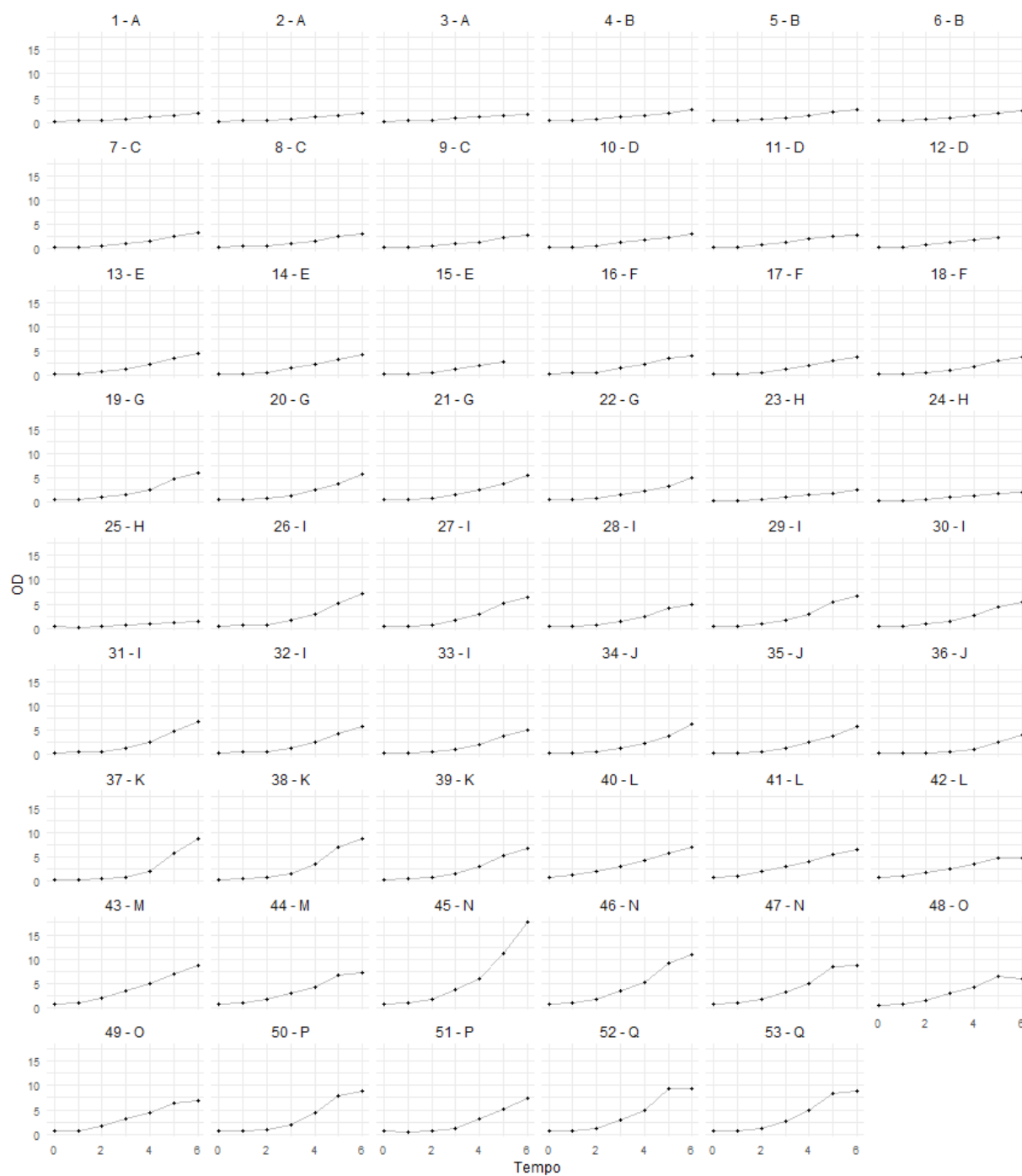


Figura 1.3: Crescita algale per ogni replica, il cui numero di riferimento e la corrispondente condizione sperimentale sono riportati sopra ogni grafico.

Nota. Le repliche 12 e 15 presentano un'osservazione mancante nell'ultimo istante temporale e sono pertanto costituite da sei osservazioni.

Capitolo 2

Modelli non lineari

L'obiettivo che ci si è posti è quello di capire come l'intensità luminosa (I), il fotoperiodo (D) e la concentrazione di fosforo (P) influenzino sia la velocità di crescita che la resa finale della densità ottica (OD). Per far ciò è stato utilizzato un modello di regressione non lineare basato sulla funzione logistica, in modo da descrivere la relazione tra le variabili indipendenti e la risposta quantitativa continua (OD). L'impiego di questo modello ha richiesto l'introduzione di alcune assunzioni riguardanti la struttura di dipendenza tra le osservazioni appartenenti alla stessa replica. In particolare, si è assunto che la dipendenza tra tali osservazioni sia attribuibile esclusivamente alla componente temporale, esplicitamente modellata attraverso la funzione scelta. Di conseguenza, una volta tenuto conto dell'effetto del tempo, anche le osservazioni relative alla stessa replica vengono considerate condizionatamente indipendenti. Mentre le repliche sono tutte indipendenti tra loro. Inizialmente è stata adattata ai dati una versione semplificata del modello, stimato su ogni replica. Successivamente le variabili I, D e P, relative alle condizioni sperimentali, sono state introdotte come covariate nel modello.

2.1 Modello sulle singole repliche

Prima di procedere alla costruzione di un modello complessivo che tenesse conto delle diverse condizioni sperimentali, si è ritenuto opportuno analizzare separatamente ciascuna replica. Questa fase ha consentito di verificare l'adeguatezza della funzione logistica nel

descrivere l'andamento della crescita algale e ha permesso di ottenere una prima caratterizzazione dei parametri del modello, evidenziando eventuali differenze tra le condizioni sperimentali e fornendo indicazioni utili per la successiva definizione del modello globale.

Per descrivere l'andamento temporale della densità ottica (OD), è stato utilizzato un modello di regressione non lineare basato su una funzione logistica. Le curve logistiche sono comunemente impiegate per rappresentare processi di crescita caratterizzati da un andamento sigmoide, nel quale vi è una fase iniziale di crescita accelerata, seguita da un progressivo rallentamento e dal raggiungimento di un valore asintotico (Seber e Wild, 2003).

Per ciascun istante temporale t , il valore osservato della densità ottica, indicato con OD_t , è stato modellato assumendo una distribuzione normale:

$$OD_t \sim \mathcal{N}(\mu_t, \sigma^2), \quad (2.1)$$

dove μ_t rappresenta il valore atteso della densità ottica al tempo t ed è descritto dalla funzione logistica

$$\mu_t = \gamma \frac{\exp(\beta t + \alpha)}{1 + \exp(\beta t + \alpha)}. \quad (2.2)$$

I parametri da stimare sono pertanto quattro:

- γ rappresenta l'asintoto superiore della curva.
Infatti, assumendo $\beta > 0$, si ha $\lim_{t \rightarrow +\infty} \mu_t = \gamma$. Dunque γ rappresenta il valore massimo teorico verso cui tende la densità ottica e può essere interpretato come una misura della resa finale, ovvero della densità ottica massima raggiungibile nelle condizioni sperimentali considerate;
- β è il parametro che determina la rapidità con cui la curva si avvicina all'asintoto superiore. Valori più elevati di β corrispondono a una crescita più rapida e a una maggiore pendenza della curva nella fase di incremento. Pertanto, tale parametro è associato alla velocità di crescita del cianobatterio;
- α è un parametro di posizione che influenza lo spostamento orizzontale della curva, dunque il momento temporale in cui si verifica la massima velocità di crescita;

- σ rappresenta la deviazione standard della componente casuale del modello e quantifica la variabilità dei valori osservati della densità ottica rispetto alla curva logistica stimata.

Il modello è stato adattato separatamente a ciascuna replica, costituita da sette osservazioni rilevate in corrispondenza dei diversi istanti temporali, ad eccezione di due repliche per le quali erano disponibili soltanto sei osservazioni (come descritto nella Sezione 1.3). Dunque è importante l'assunzione per cui la dipendenza tra tali osservazioni sia attribuibile esclusivamente alla componente temporale, esplicitamente modellata tramite la (2.2). Di conseguenza, una volta tenuto conto dell'effetto del tempo, anche le osservazioni relative alla stessa replica vengono considerate condizionatamente indipendenti.

Inoltre vista la limitata numerosità campionaria rispetto al numero di parametri da stimare, le stime ottenute devono essere interpretate con cautela; infatti per ciascuna replica sono stati stimati quattro parametri sulla base di sette osservazioni, o di sei nei due casi sopra menzionati.

La stima dei parametri è stata effettuata massimizzando la funzione di log-verosimiglianza, mediante l'algoritmo L-BFGS-B (*Limited-memory Broyden-Fletcher-Goldfarb-Shanno with Bounds*), un metodo quasi-Newton pensato per risolvere problemi di ottimizzazione non lineare e che consente anche di imporre vincoli inferiori e/o superiori ai parametri da stimare (Byrd et al., 1995). In particolare è stato posto il vincolo $\sigma > 0$, infatti, esso indica la deviazione standard della distribuzione normale assunta per la variabile risposta, per cui deve assumere esclusivamente valori positivi.

Successivamente si è svolta un'analisi di tipo esplorativo sulle stime dei parametri ottenute dall'adattamento del modello alle singole repliche, con l'obiettivo di ricavare indicazioni utili per la successiva definizione del modello globale. In particolare è possibile visualizzare i grafici di dispersione tra le varie coppie di parametri, insieme ai corrispondenti coefficienti di correlazione, in Figura 2.1. Si nota che le stime dei parametri $\hat{\gamma}$ e $\hat{\beta}$ non presentano una correlazione rilevante; al contrario di $\hat{\beta}$ e $\hat{\alpha}$ che risultano correlati come atteso sulla base dell'interpretazione dei due parametri. In Figura 2.2 si può invece osservare il valore delle stime dei parametri al variare delle condizioni sperimentali, dunque di intensità luminosa (I), fotoperiodo (D) e concentrazione di fosforo (P). Sembra che $\hat{\gamma}$ aumenti all'aumentare sia di I che di P. $\hat{\beta}$ sembra aumentare all'aumentare di tutte le covariate, eccetto P, per cui sembra esserci una diminuzione oltre un certo valore compreso tra 25,5 e

$50 \text{ mg}_P L^{-1}$. $\hat{\alpha}$ sembra diminuire all'aumentare di tutte e tre le covariate. Infine $\hat{\sigma}$ sembra diventare più variabile all'aumentare dei valori delle tre covariate.

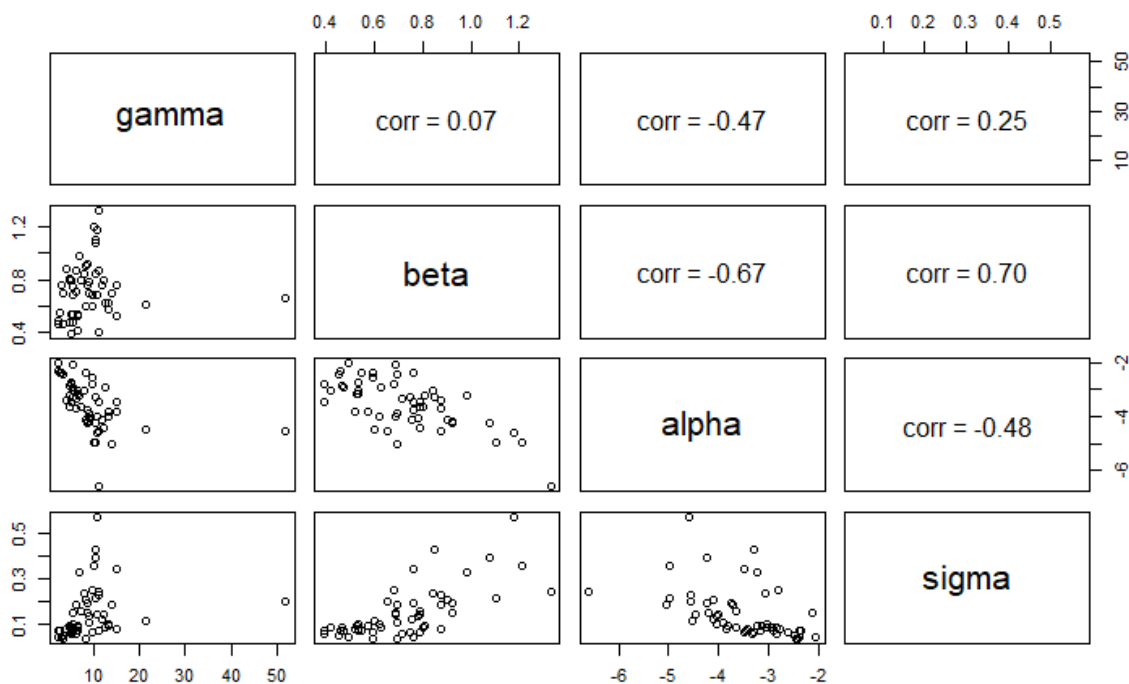


Figura 2.1: Grafici di dispersione tra coppie dei parametri $\hat{\gamma}$, $\hat{\beta}$, $\hat{\alpha}$ e $\hat{\sigma}^2$, con relativo valore di correlazione.

Sempre con fini esplorativi, è stato stimato un modello lineare multivariato (Grigoletto, Ventura e Pauli, 2017) per valutare congiuntamente la relazione tra i fattori sperimentali e i valori dei parametri stimati. A differenza del modello lineare univariato, nel quale viene considerata una sola variabile risposta, il modello lineare multivariato consente di analizzare simultaneamente più variabili risposta. Nel presente caso, le variabili risposta sono rappresentate dai parametri stimati $\hat{\gamma}$, $\hat{\beta}$, $\hat{\alpha}$ e la trasformata logaritmica di $\hat{\sigma}$ (le motivazioni dell'applicazione di tale trasformata saranno spiegate nella sezione successiva), mentre l'intensità luminosa (I), il fotoperiodo (D) e la concentrazione di fosforo (P) costituiscono le covariate esplicative.

Indicando con $\mathbf{Y}$ la matrice delle variabili risposta, con $\mathbf{X}$ la matrice delle covariate e con $\mathbf{B}$ la matrice dei coefficienti di regressione, il modello lineare multivariato può essere espresso nella forma:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \quad (2.3)$$

dove $\mathbf{E}$ rappresenta la matrice degli errori casuali. Si assume che le righe di $\mathbf{E}$ siano indipendenti e che ciascuna di esse segua una distribuzione normale multivariata con vettore medio nullo e matrice di covarianza $\mathbf{\Sigma}$:

$$\mathbf{E}_i \sim \mathcal{N}_q(\mathbf{0}, \mathbf{\Sigma}), \quad i = 1, \dots, n, \quad (2.4)$$

dove n rappresenta il numero di osservazioni e q il numero di variabili risposta. La matrice di covarianza $\mathbf{\Sigma}$ descrive sia la variabilità residua associata a ciascuna variabile risposta sia la correlazione residua tra le diverse variabili risposta.

Nella Tabella 2.1 sono riportati i valori dei coefficienti dei parametri del modello e, tra parentesi, i relativi standard error. Si nota che tutte le covariate sono significative per tutti i parametri, eccetto il fotoperiodo (D) per $\hat{\gamma}$.

Tabella 2.1: Stime dei coefficienti del modello lineare multivariato definito nella 2.3.

	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\alpha}$	$\log(\hat{\sigma})$
Intercetta	3.241 (4.068)	-0.060 (0.082)	-1.317** (0.454)	-4.097*** (0.266)
I	0.021** (0.006)	0.001*** (0.000)	-0.002* (0.001)	0.003*** (0.000)
D	-0.123 (0.184)	0.030*** (0.004)	-0.065** (0.021)	0.044*** (0.012)
P	0.108* (0.050)	0.002* (0.001)	-0.023*** (0.006)	0.012*** (0.003)
R^2	0.256	0.666	0.407	0.651

Nota: gli errori standard sono riportati tra parentesi. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

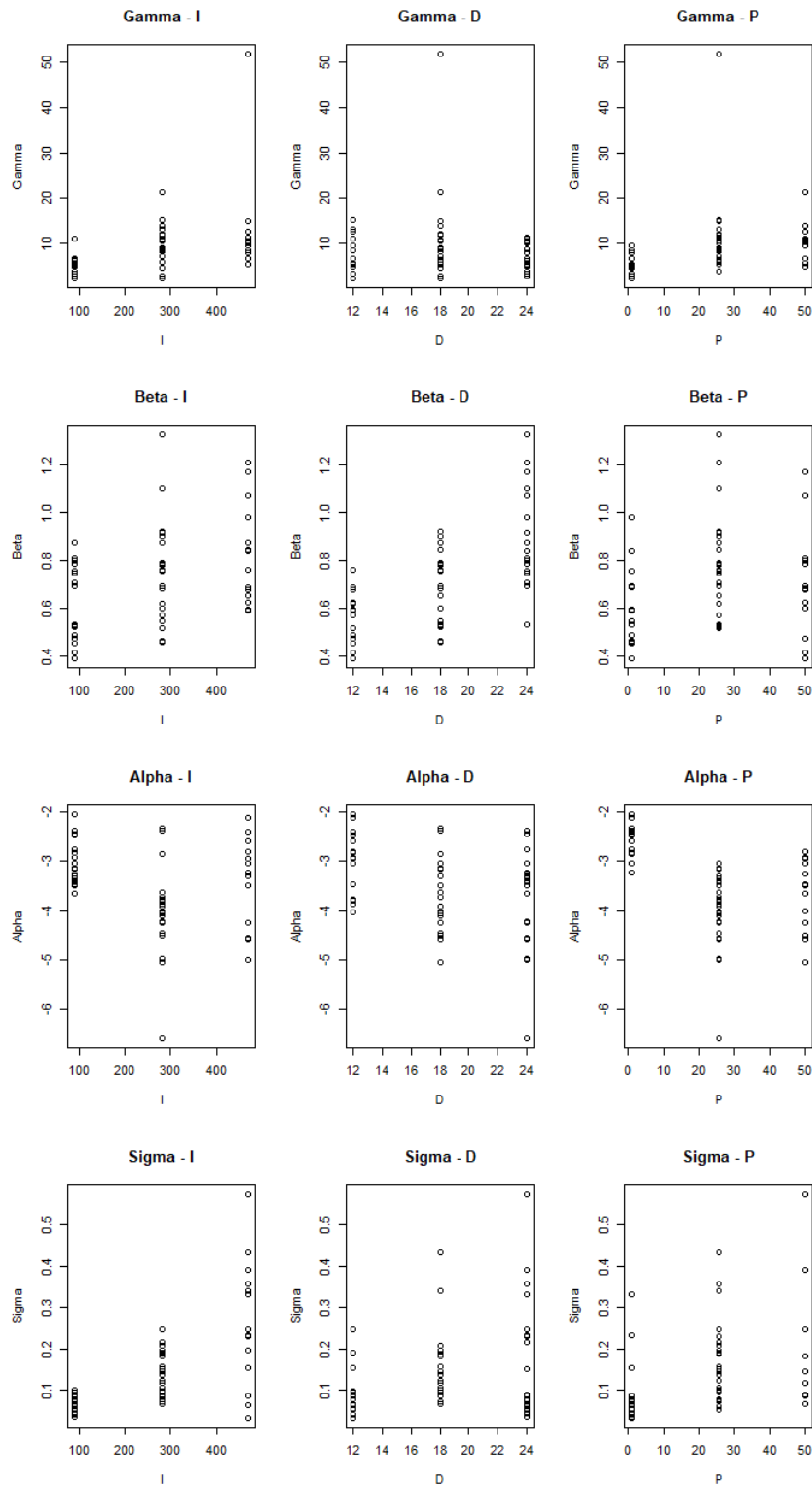


Figura 2.2: Grafici di dispersione delle stime dei parametri $\hat{\gamma}$, $\hat{\beta}$, $\hat{\alpha}$ e $\hat{\sigma}^2$ al variare dei fattori sperimentali: intensità luminosa (I), fotoperiodo (D) e concentrazione di fosforo (P).

2.2 Modello globale

Dopo aver analizzato separatamente le singole repliche e verificato l'adeguatezza della funzione logistica nel descrivere l'andamento della crescita della densità ottica, i fattori sperimentali intensità luminosa (I), fotoperiodo (D) e concentrazione di fosforo (P) sono stati introdotti come covariate nel modello descritto nelle equazioni 2.1 e 2.2, in modo da tenerne conto congiuntamente. Il modello è stato stimato sull'intero insieme delle osservazioni. L'assunzione di indipendenza condizionata delle osservazioni relative alla stessa replica, discussa nella sezione precedente, rimane valida, infatti si assume che esse dipendano tra loro solo attraverso la componente temporale, anche in questo caso inserita nella modellazione.

In particolare, per l' i -esima replica ($i = 1, \dots, n$), all'istante t -esimo, si assume che la densità ottica Y_{it} , segua una distribuzione normale:

$$Y_{it} \sim \mathcal{N}(\mu_{it}, \sigma_i^2), \quad (2.5)$$

dove μ_{it} rappresenta il valore atteso della densità ottica per l' i -esima osservazione al tempo t e

$$\mu_{it} = \gamma(I_i, D_i, P_i) \frac{\exp(\beta(I_i, D_i, P_i) t + \alpha(I_i, D_i, P_i))}{1 + \exp(\beta(I_i, D_i, P_i) t + \alpha(I_i, D_i, P_i))}. \quad (2.6)$$

Analogamente a quanto visto in precedenza i parametri $\gamma(I, D, P)$, $\beta(I, D, P)$ e $\alpha(I, D, P)$ indicano, rispettivamente, l'asintoto superiore della curva logistica che può essere interpretato come una misura della resa finale, il parametro associato alla velocità di crescita e il parametro di posizione della curva logistica. Tali parametri sono stati espressi come funzioni lineari delle condizioni sperimentali:

$$\begin{aligned} \gamma(I_i, D_i, P_i) &= \theta_{\gamma 0} + \theta_{\gamma 1} I_i + \theta_{\gamma 2} D_i + \theta_{\gamma 3} P_i, \\ \beta(I_i, D_i, P_i) &= \theta_{\beta 0} + \theta_{\beta 1} I_i + \theta_{\beta 2} D_i + \theta_{\beta 3} P_i, \\ \alpha(I_i, D_i, P_i) &= \theta_{\alpha 0} + \theta_{\alpha 1} I_i + \theta_{\alpha 2} D_i + \theta_{\alpha 3} P_i. \end{aligned} \quad (2.7)$$

Pertanto, l'intensità luminosa, il fotoperiodo e la concentrazione di fosforo influenzano indirettamente l'andamento della densità ottica attraverso il loro effetto sui parametri della funzione logistica.

Anche la varianza della densità ottica σ_i^2 è stata modellata in funzione delle condizioni sperimentali secondo la seguente espressione:

$$\sigma_i^2 = \exp(\theta_{\sigma 0} + \theta_{\sigma 1} I_i + \theta_{\sigma 2} D_i + \theta_{\sigma 3} P_i), \quad (2.8)$$

dove $\theta_{\sigma 0}$ rappresenta l'intercetta, mentre $\theta_{\sigma 1}$, $\theta_{\sigma 2}$ e $\theta_{\sigma 3}$ rappresentano gli effetti dell'intensità luminosa, del fotoperiodo e della concentrazione di fosforo sulla variabilità della densità ottica. La funzione esponenziale garantisce che σ_i^2 assuma valori strettamente positivi per ogni possibile combinazione delle covariate.

Il modello complessivo è pertanto caratterizzato da 16 parametri, stimati utilizzando le 369 osservazioni disponibili mediante la massimizzazione della funzione di log-verosimiglianza $l(\boldsymbol{\theta})$, dove $\boldsymbol{\theta}$ denota il vettore dei parametri del modello. La procedura di stima è stata effettuata mediante l'algoritmo L-BFGS-B (*Limited-memory Broyden-Fletcher-Goldfarb-Shanno with Bounds*), descritto nella Sezione 2.1. Tale algoritmo è adatto al contesto considerato, sia per la presenza di un numero relativamente elevato di parametri da stimare, sia per la possibilità di imporre vincoli inferiori e superiori ai valori dei parametri durante il processo di ottimizzazione (Byrd et al., 1995).

2.3 Aspetti computazionali

Durante la fase di stima sono emerse alcune criticità legate all'ottimizzazione numerica. La complessità della funzione di log-verosimiglianza e l'elevato numero di parametri da stimare hanno reso necessario l'impiego di diverse strategie volte ad aumentare l'affidabilità della procedura di ottimizzazione e a ridurre il rischio di convergenza verso massimi locali, anziché verso il massimo globale della funzione di log-verosimiglianza $l(\boldsymbol{\theta})$. Tra le strategie adottate rientrano l'imposizione di vincoli inferiori e superiori sui valori assumibili dai parametri e l'utilizzo di molteplici punti iniziali, a partire dai quali è stato avviato l'algoritmo di ottimizzazione. In particolare sono stati considerati i seguenti vettori di valori iniziali:

- 50 vettori di valori generati casualmente da una distribuzione Uniforme, di cui:
 - 25 vettori con valori estratti dall'intervallo $[-0.5, 0.5]$;
 - 25 vettori con valori estratti dall'intervallo $[-0.15, 0.15]$;

- 1 vettore con le stime esatte dei coefficienti del modello lineare multivariato, definito nella 2.3;
- 9 vettori di valori intorno le stime dei coefficienti del modello lineare multivariato, distanziati di 0.01.

Per quanto riguarda i vincoli dell'algoritmo L-BFGS-B imposti ai parametri, nelle diverse esecuzioni dell'algoritmo sono state considerate differenti combinazioni di limiti inferiori e superiori. In particolare, sono state alternate configurazioni prive di vincoli a configurazioni in cui i limiti sono definiti sulla base degli estremi degli intervalli di confidenza al livello del 99%, estesi di un fattore pari a 3 (Figura 2.3), dei rispettivi coefficienti del modello lineare multivariato definito nella Sezione 2.1 nell'Equazione 2.3.

Sia $\hat{\theta}$ la stima di un generico parametro e sia $[L, U]$ il relativo intervallo di confidenza, dove L e U rappresentano rispettivamente l'estremo inferiore e l'estremo superiore. I limiti inferiori e superiori utilizzati sono definiti come:

$$L^* = L - (U - L) = 2L - U;$$

$$U^* = U + (U - L) = 2U - L.$$

Tale approccio ha consentito di confrontare le soluzioni ottenute in corrispondenza di differenti regioni ammissibili e, in alcuni casi, di favorire il raggiungimento della convergenza, non ottenuta in assenza di vincoli. In particolare, sono stati utilizzati i seguenti limiti:

- $\theta_\gamma, \theta_\beta, \theta_\alpha \in [-\infty, +\infty]; \quad \theta_{\sigma 0} \in [-7, 1]; \quad \theta_{\sigma 1}, \theta_{\sigma 2}, \theta_{\sigma 3} \in [-0.5, 1];$
- $\theta_\gamma, \theta_\beta, \theta_\alpha \in [-\infty, U^*]; \quad \theta_{\sigma 0} \in [-7, U^*]; \quad \theta_{\sigma 1}, \theta_{\sigma 2}, \theta_{\sigma 3} \in [-0.5, U^*];$
- $\theta_\gamma, \theta_\beta, \theta_\alpha \in [L^*, +\infty]; \quad \theta_\sigma \in [L^*, 1];$
- $\theta_\gamma, \theta_\beta, \theta_\alpha, \theta_\sigma \in [L^*, U^*]$

Combinando i diversi vettori di valori iniziali con le varie configurazioni di vincoli imposte ai parametri, l'algoritmo di ottimizzazione è stato eseguito complessivamente 240 volte, arrivando a convergenza 139 volte. L'obiettivo di tale procedura è quello di selezionare, tra le soluzioni ottenute, quella associata al valore più elevato della funzione di log-verosimiglianza, in quanto la convergenza poteva essere associata a massimi locali.

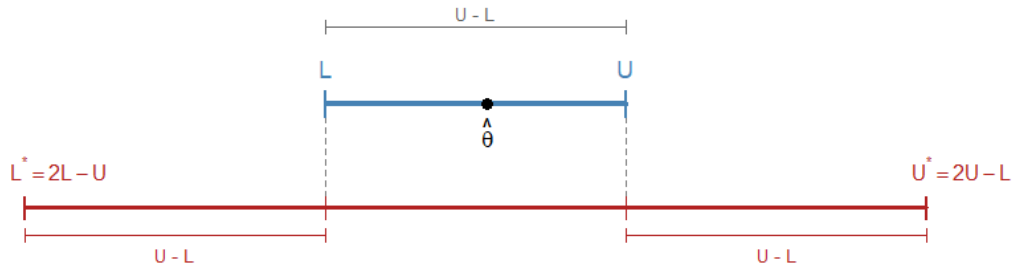


Figura 2.3: Rappresentazione grafica dell'estensione dell'intervallo di confidenza di un fattore pari a 3, dove $L^* = 2L - U$ e $U^* = 2U - L$ sono gli estremi estesi rispetto all'intervallo di confidenza originale $[L, U]$.

Alla luce delle difficoltà riscontrate durante la procedura di ottimizzazione, è stata condotta un'analisi della funzione di log-verosimiglianza, con l'obiettivo di comprenderne la struttura e di individuare le possibili cause delle criticità emerse nella stima dei parametri. Dato che la funzione dipende da 16 parametri, non è possibile rappresentarla graficamente, dunque l'analisi è stata svolta mediante un approccio numerico, volto a esaminare il comportamento della funzione intorno ai punti di massimo locale, individuati dalle 139 esecuzioni dell'algoritmo di ottimizzazione arrivate a convergenza. A tal fine, per ognuno dei 139 vettori di parametri ottimali (ottenuti come soluzione dall'algoritmo), sono stati selezionati casualmente 1000 punti, intesi come vettori di 16 valori, appartenenti a un intorno di ampiezza pari a 0.2 della rispettiva soluzione.

La scelta dell'ampiezza dell'intorno è stata supportata da un'analisi grafica, nella quale sono stati rappresentati i profili dei 139 vettori di parametri corrispondenti alle soluzioni convergenti e quelli dei 1000 punti nell'intorno di una di tali soluzioni. Nella Figura 2.4 si possono visualizzare, in grigio, i profili dei vettori di parametri associati alle 139 soluzioni convergenti e, in blu, i vettori generati nell'intorno della soluzione associata al valore più elevato della funzione di log-verosimiglianza. La rappresentazione grafica ha consentito di valutare visivamente l'ampiezza dell'intorno rispetto alla variabilità osservata tra le diverse soluzioni ottenute dall'algoritmo.

Successivamente, per ciascuno dei 139 massimi locali individuati dall'algoritmo, è stato calcolato il valore della funzione di log-verosimiglianza in corrispondenza dei 1000 punti

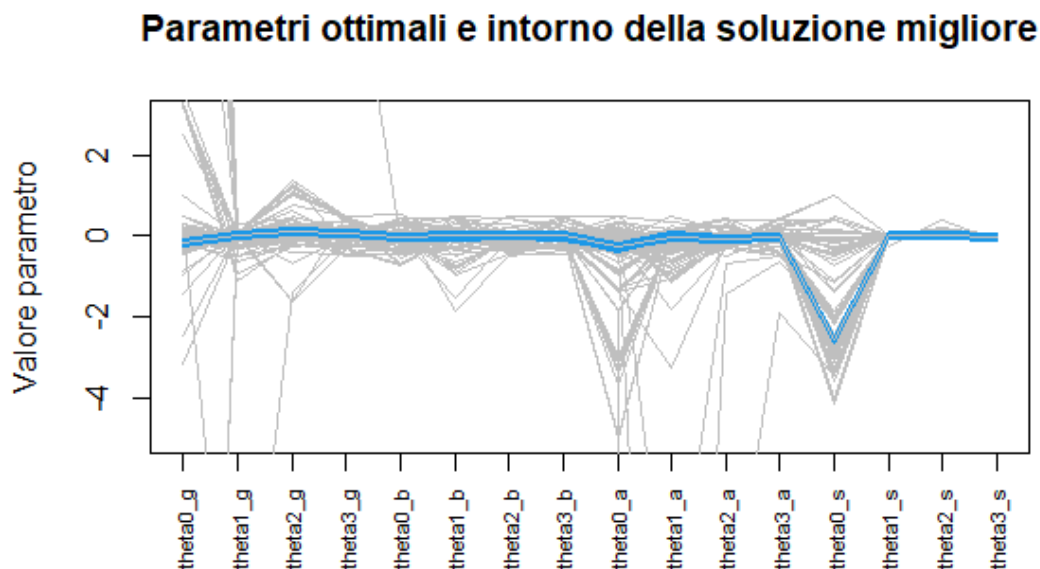


Figura 2.4: In grigio i profili dei vettori di parametri associati alle 139 soluzioni convergenti e, in blu, i 1000 vettori generati nell'intorno della soluzione associata al valore più elevato della funzione di log-verosimiglianza.

generati nel relativo intorno. Tale andamento è illustrato nella Figura 2.5, nella quale sono riportati, in ordine crescente, i valori della funzione di log-verosimiglianza calcolati nei 1000 punti appartenenti all'intorno di ciascuna delle 139 soluzioni convergenti. A ciascun insieme di valori è stato attribuito un colore differente, mentre in nero sono evidenziati i valori dell'intorno relativo alla soluzione migliore trovata dall'algoritmo. Il punto nero in alto a destra indica il valore della funzione di log-verosimiglianza corrispondente a tale soluzione. I valori ottenuti risultano sensibilmente differenti da quelli associati alle rispettive soluzioni individuate dall'algoritmo. Inoltre, si osserva che circa 500 dei 1000 valori calcolati sono inferiori a -15000 . Dunque ciò suggerisce che la funzione di log-verosimiglianza presenti variazioni rilevanti anche in prossimità delle soluzioni trovate dall'algoritmo. Questo comportamento può essere indicativo della presenza di regioni di ottimo caratterizzate da una forte curvatura o da picchi particolarmente stretti, contribuendo a spiegare le difficoltà di ottimizzazione della funzione.

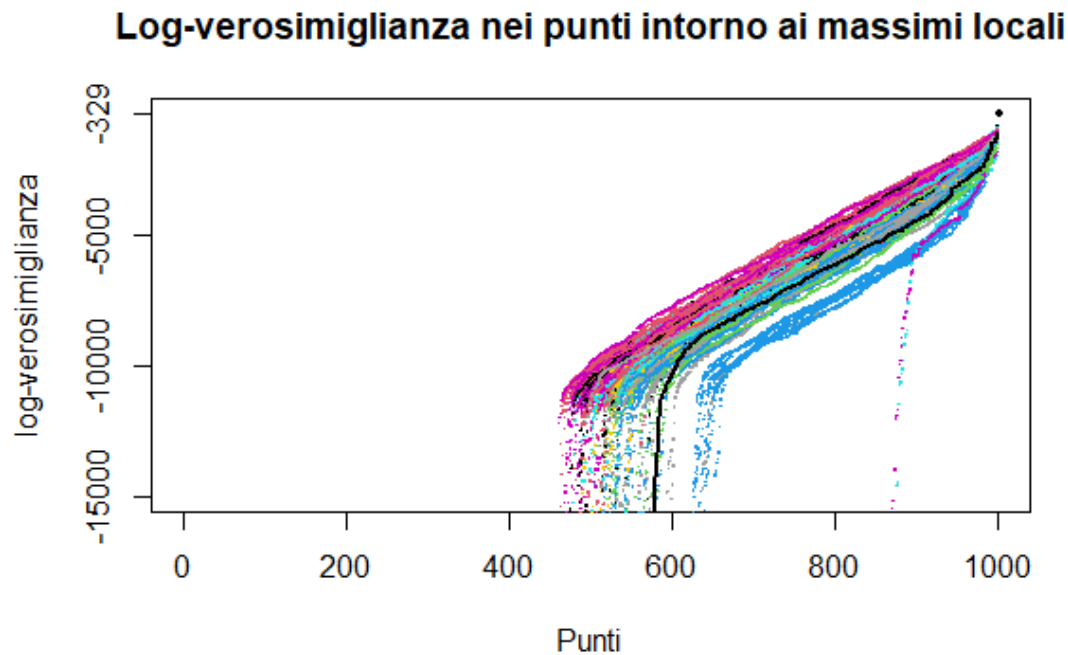


Figura 2.5: Valori della funzione di log-verosimiglianza, ordinati in senso crescente, nei 1000 punti generati nell'intorno di ciascuna delle 139 soluzioni convergenti. Ogni colore identifica un diverso intorno; in nero è evidenziato quello relativo alla soluzione associata al valore più elevato di log-verosimiglianza, che è indicato dal punto nero in alto a destra.

Considerata la struttura della funzione di log-verosimiglianza, è stata adottata un'ulteriore strategia per esplorare in modo più approfondito lo spazio dei parametri e individuare valori più elevati della funzione di log-verosimiglianza, in modo da evitare di rimanere intrappolati in massimi locali. In particolare si è cercato un vettore di valori da utilizzare come nuovo punto iniziale dell'algoritmo, selezionato con l'obiettivo di favorire la convergenza verso soluzioni associate a valori più elevati della funzione di log-verosimiglianza. La strategia adottata si basa sulla semplificazione temporanea della funzione di log-verosimiglianza. In particolare, è stato fissato a zero, uno alla volta, ciascuno dei 16 componenti del vettore dei parametri, ottenendo così 16 versioni ridotte del modello. Ognuna di esse è stata poi ottimizzata, mediante l'algoritmo L-BFGS-B, adottando le stesse impostazioni descritte all'inizio della presente sezione, ovvero utilizzando i 60 differenti vettori di valori iniziali e le diverse configurazioni di vincoli precedentemente considerate.

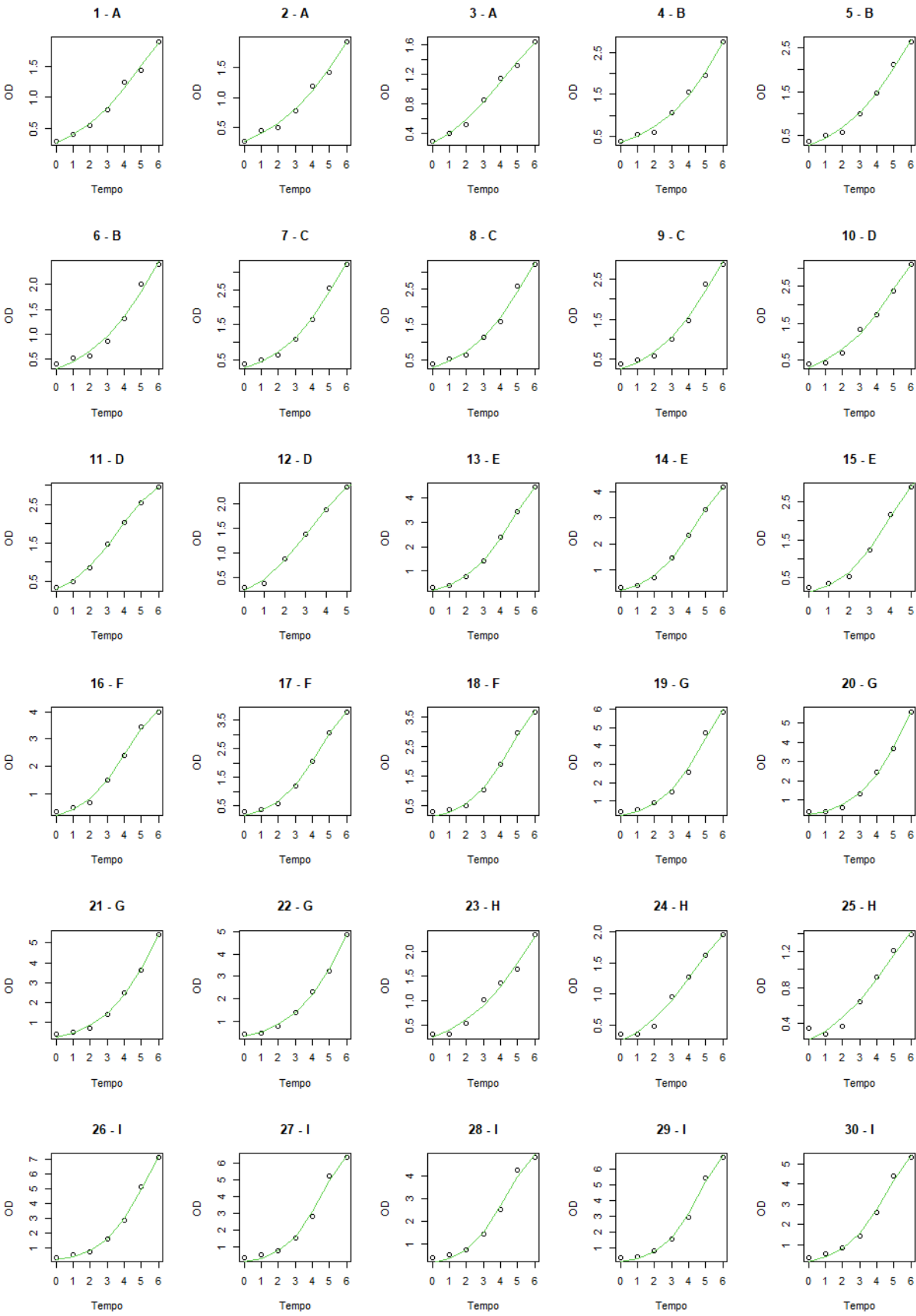
Successivamente è stata selezionata la versione del modello associata al valore di log-verosimiglianza più elevato ottenuto in tutte queste prove. In particolare, il vettore di parametri corrispondente è stato utilizzato come nuovo punto iniziale per l'ottimizzazione del modello completo, ponendo pari a zero il parametro mancante, corrispondente a $\theta_{\beta 3}$. L'applicazione di questa procedura ha consentito di ottenere un valore della funzione di log-verosimiglianza superiore a quello individuato nelle precedenti esecuzioni. La soluzione così ottenuta è stata pertanto selezionata come quella associata al valore più elevato della funzione di log-verosimiglianza tra tutte le soluzioni considerate.

Capitolo 3

Risultati

3.1 Risultati sulle singole repliche

L'adattamento del modello alle singole repliche ha permesso di valutare l'adeguatezza della funzione scelta nel descrivere l'andamento della densità ottica (OD) nel tempo e di ottenere, per ciascuna replica, le stime dei quattro parametri γ , β , α e σ , del modello definito nella 2.1 e 2.2. Nella Figura 3.1 sono riportate, in verde, le curve stimate per ciascuna replica e i dati osservati. Nel complesso, si osserva un buon adattamento del modello ai dati, infatti le curve stimate seguono in modo soddisfacente l'andamento della densità ottica osservata. Le stime dei parametri ottenute per le 53 repliche mostrano una certa variabilità, sia per quanto riguarda il livello massimo della densità ottica, sia per la rapidità con cui tale livello viene raggiunto. In particolare, le stime di γ variano da circa 2.3 a 21.5 (escludendo un valore pari a 51.8, relativo alla replica 45, che si discosta molto dagli altri), mentre quelle di β sono comprese tra circa 0.39 e 1.33. La variabilità osservata nelle stime di γ e β suggerisce quindi la presenza di differenze nelle dinamiche di crescita associate alle diverse condizioni sperimentali. Il parametro α assume valori compresi tra circa -6.59 e -2.05 . Tale parametro determina la posizione temporale della curva e, insieme a β , contribuisce a individuare il momento in cui viene raggiunta la fase di massima velocità di crescita. Le stime di σ , infine, variano tra circa 0.032 e 0.573, evidenziando una diversa variabilità delle osservazioni attorno alle rispettive curve stimate. Nel complesso, i risultati ottenuti in questa fase preliminare confermano l'adeguatezza della funzione logistica nel rappresentare l'andamento della crescita del cianobatterio nelle diverse repliche.



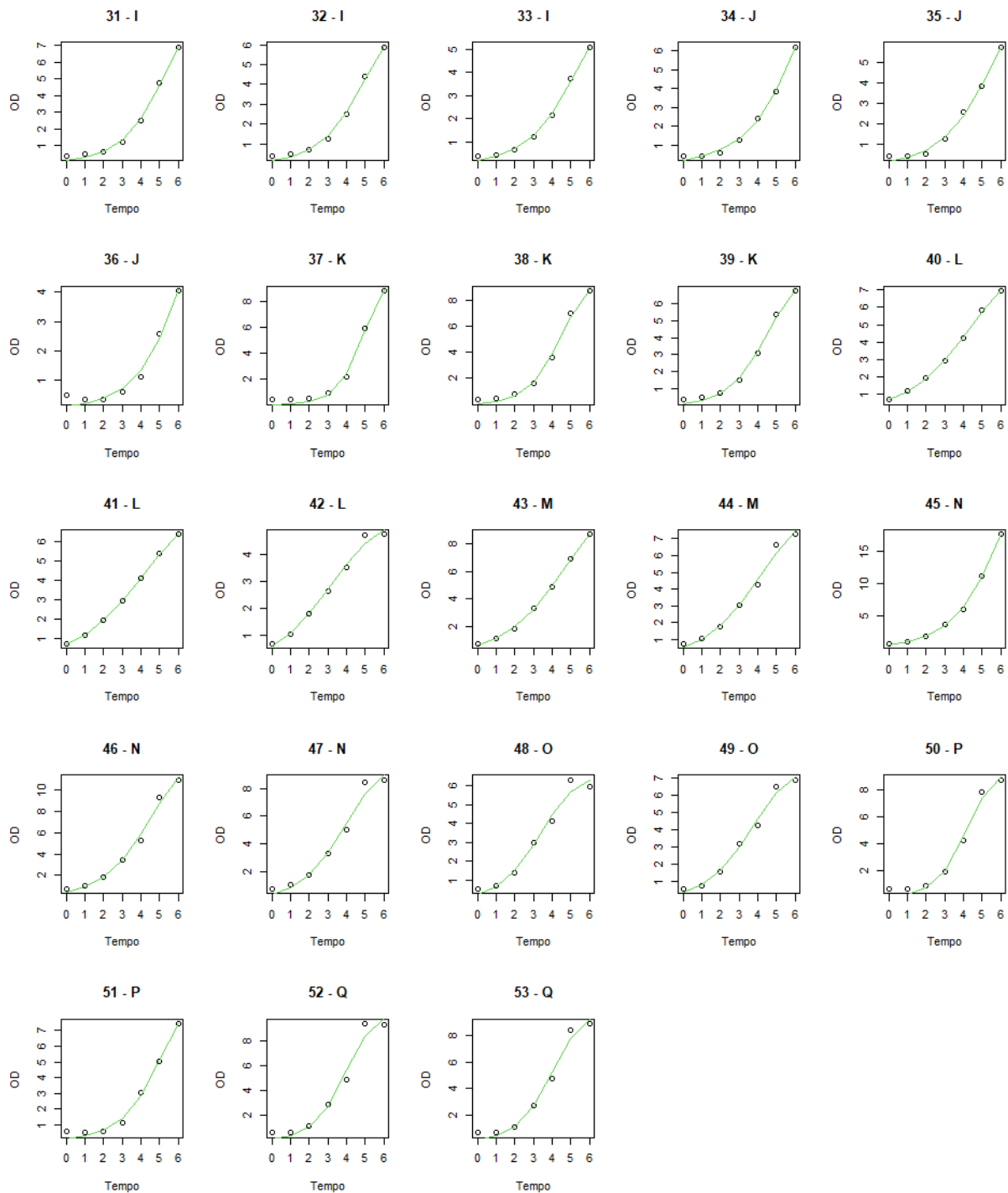


Figura 3.1: Curve stimate del modello, definito nella 2.1 e 2.2, per ogni replica.

Sopra ogni grafico sono riportati il rispettivo numero di replica e la rispettiva condizione sperimentale.

Nota. Le repliche 12 e 15 presentano un'osservazione mancante nell'ultimo istante temporale e sono pertanto costituite da sei osservazioni.

3.2 Risultati del modello globale

Il modello globale, definito nella 2.5 e 2.6, consente di valutare congiuntamente il contributo dell'intensità luminosa (I), del fotoperiodo (D) e della concentrazione di fosforo (P) sulla velocità di crescita (β), sulla resa finale (γ) e sulla posizione della curva (α), oltre che sulla variabilità (σ) della densità ottica. Tramite la procedura descritta nella Sezione 2.3, sono state ottenute le stime dei parametri di tale modello, riportate in Tabella 3.1 insieme alla relativa significatività e statistica test, che è stata calcolata mediante test del log-rapporto di verosimiglianza, confrontando il modello completo con il modello ridotto, ottenuto rimuovendo uno alla volta il singolo parametro in esame. La metodologia adottata per la costruzione di tali test, sarà illustrata nella sezione successiva. Dai risultati emerge che la quasi totalità dei coefficienti risulta significativa a un livello dell'1%, indicando un effetto rilevante di tutte e tre le variabili (I, D e P) sui parametri della curva logistica (γ , β e α) e sulla varianza (σ^2). L'unica eccezione è rappresentata dal coefficiente $\theta_{\beta 3}$, associato all'effetto della concentrazione di fosforo (P) sulla velocità di crescita (β), che risulta significativo al livello del 10% ($p = 0.099$), seppur con un p -value prossimo alla soglia di non significatività, mentre risulta non significativo al livello del 5%; indice del fatto che la concentrazione di fosforo sembra non influenzare la velocità di crescita della biomassa. In particolare si nota che all'aumentare di intensità luminosa (I), fotoperiodo (D) e concentrazione di fosforo (P) aumenta e si anticipa il livello di massima crescita della densità ottica (rispettivamente rappresentati da γ e α). Per quanto riguarda la velocità di crescita, essa aumenta solo all'aumentare di intensità luminosa (I) e fotoperiodo (D). Poiché α determina la posizione della curva lungo l'asse temporale, tale effetto deve tuttavia essere interpretato congiuntamente a quello di β . In particolare, il tempo corrispondente al punto di flesso è dato da

$$t_{\text{flesso}} = -\frac{\alpha}{\beta}. \quad (3.1)$$

Pertanto, la variazione di α non può essere interpretata isolatamente come un anticipo o un ritardo della crescita, ma deve essere valutata considerando anche la variazione di β . Anche la varianza aumenta all'aumentare di intensità luminosa (I) e fotoperiodo (D), ma diminuisce all'aumentare della concentrazione di fosforo (P), indicando una maggior variabilità nelle osservazioni esposte a una luce di maggior intensità e per un periodo di maggior durata.

Tabella 3.1: Stime dei coefficienti del modello globale (2.5 e 2.6), statistica test e p-value del test del log-rapporto di verosimiglianza.

Interpretazione	Parametro	Stima	Statistica LRV	p-value	
<i>Asintoto superiore (γ)</i>					
Intercetta	$\theta_{\gamma 0}$	0.8858	65.90	< 0.001	***
Intensità luminosa (I)	$\theta_{\gamma 1}$	0.0071	47.66	< 0.001	***
Fotoperiodo (D)	$\theta_{\gamma 2}$	0.0883	73.16	< 0.001	***
Fosforo (P)	$\theta_{\gamma 3}$	0.1508	88.19	< 0.001	***
<i>Velocità di crescita (β)</i>					
Intercetta	$\theta_{\beta 0}$	-0.0316	55.48	< 0.001	***
Intensità luminosa (I)	$\theta_{\beta 1}$	0.0016	87.27	< 0.001	***
Fotoperiodo (D)	$\theta_{\beta 2}$	0.0211	94.53	< 0.001	***
Fosforo (P)	$\theta_{\beta 3}$	-0.0003	2.72	0.099	.
<i>Posizione della curva (α)</i>					
Intercetta	$\theta_{\alpha 0}$	-0.4525	123.87	< 0.001	***
Intensità luminosa (I)	$\theta_{\alpha 1}$	-0.0047	143.05	< 0.001	***
Fotoperiodo (D)	$\theta_{\alpha 2}$	-0.0749	30.15	< 0.001	***
Fosforo (P)	$\theta_{\alpha 3}$	-0.0290	89.87	< 0.001	***
<i>Varianza ($\log \sigma^2$)</i>					
Intercetta	$\theta_{\sigma 0}$	-3.2075	163.16	< 0.001	***
Intensità luminosa (I)	$\theta_{\sigma 1}$	0.0063	358.26	< 0.001	***
Fotoperiodo (D)	$\theta_{\sigma 2}$	0.0545	39.78	< 0.001	***
Fosforo (P)	$\theta_{\sigma 3}$	-0.0028	72.09	< 0.001	***

Livelli di significatività: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; . $p < 0.1$.

Per visualizzare graficamente la stima del modello globale, le curve stimate sono state rappresentate insieme ai valori osservati di densità ottica per ciascuna delle 17 condizioni sperimentali, come è possibile vedere in Figura 3.2, dove oltre alla curva stimata, in blu, è stata rappresentata la sua proiezione, in rosso, nei 2 istanti temporali successivi all'ultimo osservato. Nella maggior parte dei casi il modello coglie adeguatamente l'andamento sigmoidale della crescita, ma si osservano alcune eccezioni, in particolare per le condizioni sperimentali H (I pari a $280 \mu\text{mol m}^{-2} \text{s}^{-1}$, D pari a 18 h, P pari a $1 \text{ mg}_P \text{L}^{-1}$) e N (I pari a $470 \mu\text{mol m}^{-2} \text{s}^{-1}$, D pari a 18 h, P pari a $25.5 \text{ mg}_P \text{L}^{-1}$) in cui la curva stimata si discosta dai valori osservati; in particolare nella condizione sperimentale H la curva tende a sovrastimare sistematicamente la densità ottica a partire dall'istante temporale 2, mentre nella condizione N si osserva il comportamento opposto, con una generale sottostima dei valori osservati. Graficamente si nota che nelle condizioni caratterizzate da valori più elevati di intensità luminosa e fotoperiodo si osservano generalmente curve caratterizzate da una crescita più rapida e da valori finali di densità ottica maggiori, raggiunti anche in corrispondenza di elevati valori di concentrazione di fosforo. Questo andamento è coerente con le stime dei coefficienti del modello (2.5 e 2.6), riportate nella Tabella 3.1. Come accennato in precedenza in Figura 3.2 è riportata anche la proiezione della rispettiva curva nei 2 istanti temporali successivi all'ultimo osservato; così si evidenzia come in alcune condizioni sperimentali la densità ottica abbia già raggiunto, o quasi raggiunto, il suo valore di resa finale, mentre in altre sia ancora in fase di crescita. Inoltre in Figura 3.2 si notano alcune differenze tra valori osservati e curve stimate in alcune condizioni sperimentali, soprattutto nelle fasi finali della crescita, dove si osserva una maggiore dispersione dei punti attorno alla curva. Tali differenze sono coerenti con la variabilità descritta dal parametro σ , che a sua volta dipende dalle condizioni sperimentali attraverso il modello (2.5 e 2.6). Nel complesso, la rappresentazione grafica conferma un buon adattamento del modello ai dati e la capacità di riprodurre la diversa dinamica di crescita osservata nelle 17 condizioni sperimentali, evidenziando come la variazione di I , D e P determini modifiche sia nella velocità di crescita sia nel livello finale di densità ottica.

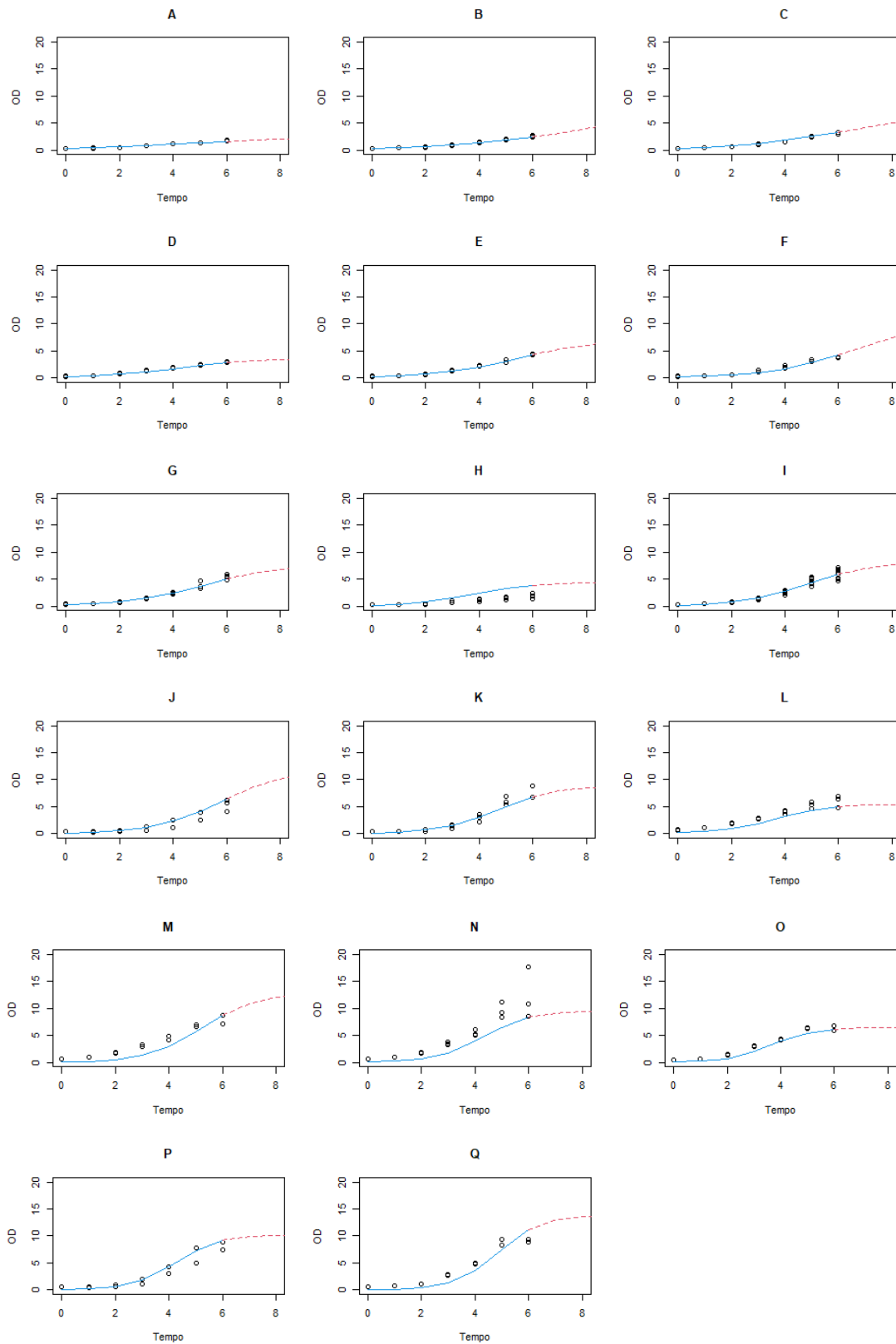


Figura 3.2: Curve stimate, in blu, e proiezione nei 2 istanti temporali successivi all'ultimo osservato, in rosso tratteggiato, del modello globale (2.5 e 2.6) rappresentate per ogni condizione sperimentale, indicata sopra il rispettivo grafico.

I punti rappresentano i dati osservati.

3.3 Test congiunti sull'effetto globale delle variabili

Come anticipato nella sezione precedente, per valutare la significatività dei coefficienti del modello globale (2.5 e 2.6) è stato utilizzato il test del log-rapporto di verosimiglianza (Azzalini, 2001). Il modello è caratterizzato da un vettore di parametri $\boldsymbol{\theta}$ di dimensione 16; si supponga di voler testare la significatività di un sottoinsieme di parametri per cui si partiziona $\boldsymbol{\theta}$ in 2 parti $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$, dove $\boldsymbol{\psi}$ è di dimensione q (con $1 \leq q \leq 16$) e rappresenta il parametro di interesse, mentre le restanti $16 - q$ componenti costituiscono $\boldsymbol{\lambda}$ che rappresenta il parametro di disturbo (Azzalini, 2001). Il sistema di ipotesi generale per la verifica di nullità dei parametri di interesse è espresso come:

$$\begin{cases} H_0 : \boldsymbol{\psi} = \mathbf{0}, \\ H_1 : \boldsymbol{\psi} \neq \mathbf{0}. \end{cases} \quad (3.2)$$

dove H_0 è l'ipotesi nulla associata al modello ridotto, mentre H_1 è l'ipotesi alternativa associata al modello completo. Si definisce poi la funzione di log-verosimiglianza profilo

$$\ell^*(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}), \quad (3.3)$$

dove $\hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}$ indica il valore di $\boldsymbol{\lambda}$ che massimizza la log-verosimiglianza per ogni valore fissato di $\boldsymbol{\psi}$. La statistica test del log-rapporto di verosimiglianza (LRV) può quindi essere definita tramite la log-verosimiglianza profilo:

$$LRV = 2 \left[\ell^*(\hat{\boldsymbol{\psi}}) - \ell^*(\mathbf{0}) \right] = 2 \left[\ell(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\lambda}}) - \ell(\mathbf{0}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}_{H_0}}) \right], \quad (3.4)$$

dove $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\lambda}})$ è la stima di massima verosimiglianza nel modello completo e $\hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}_{H_0}}$ la stima di massima verosimiglianza sotto vincolo dell'ipotesi nulla, quindi del modello ridotto. Sotto l'ipotesi nulla H_0 , e assumendo valide le consuete condizioni di regolarità, la statistica LRV segue asintoticamente una distribuzione χ^2 con q gradi di libertà, dove q corrisponde alla dimensione del parametro di interesse. Il p -value è pertanto calcolato come:

$$p = P(\chi_q^2 \geq LRV). \quad (3.5)$$

Dal punto di vista computazionale, il valore di massima verosimiglianza del modello ridotto è stato calcolato attraverso la procedura di ottimizzazione descritta nella Sezione 2.3. Nello specifico, per ogni configurazione di parametri vincolati a zero, il corrispondente

modello è stato stimato eseguendo 240 volte l'algoritmo L-BFGS-B, alternando le diverse combinazioni di valori iniziali e limiti sui parametri, da cui sono stati esclusi i valori relativi ai parametri posti pari a 0 nell'ipotesi nulla H_0 .

Il test è stato eseguito sia sui singoli coefficienti per valutarne la significatività, sia su diverse configurazioni di parametri per testarne congiuntamente la nullità. Nel primo caso, per ogni parametro $\theta_j \in \boldsymbol{\theta}$, $j = 1, \dots, 16$, è stato definito il seguente sistema d'ipotesi, coerentemente alla notazione adottata precedentemente:

$$\begin{cases} H_0 : \boldsymbol{\psi} \equiv \theta_j = 0, \\ H_1 : \boldsymbol{\psi} \equiv \theta_j \neq 0. \end{cases} \quad (3.6)$$

Qui la dimensione q di $\boldsymbol{\psi}$ è pari a 1, quindi la statistica LRV si distribuisce approssimativamente come una χ_1^2 . I risultati dei test svolti sui singoli coefficienti sono riportati nella Tabella 3.1 e discussi nella sezione precedente.

Successivamente, il test è stato esteso a diversi sottoinsiemi di parametri, al fine di valutarne congiuntamente la significatività. I sottoinsiemi definiti sono di due tipi: uno per verificare l'effetto complessivo di ciascuna delle variabili sperimentali (intensità luminosa I, fotoperiodo D e concentrazione di fosforo P) sull'intero andamento di crescita, l'altro per verificare l'effetto congiunto di tutte le 3 variabili sperimentali sulla resa finale (γ), velocità di crescita (β), posizione della curva (α) e variabilità delle osservazioni (σ). Nel primo caso sono stati considerati separatamente quattro sistemi di ipotesi, ciascuno dei quali pone simultaneamente uguali a zero i quattro coefficienti associati rispettivamente all'intercetta e alle variabili I, D e P:

$$\begin{cases} H_0 : \boldsymbol{\psi} \equiv (\theta_{\gamma 0}, \theta_{\beta 0}, \theta_{\alpha 0}, \theta_{\sigma 0}) = \mathbf{0}, \\ H_1 : \boldsymbol{\psi} \equiv (\theta_{\gamma 0}, \theta_{\beta 0}, \theta_{\alpha 0}, \theta_{\sigma 0}) \neq \mathbf{0}. \end{cases} \quad (3.7)$$

$$\begin{cases} H_0 : \boldsymbol{\psi} \equiv (\theta_{\gamma 1}, \theta_{\beta 1}, \theta_{\alpha 1}, \theta_{\sigma 1}) = \mathbf{0}, \\ H_1 : \boldsymbol{\psi} \equiv (\theta_{\gamma 1}, \theta_{\beta 1}, \theta_{\alpha 1}, \theta_{\sigma 1}) \neq \mathbf{0}. \end{cases} \quad (3.8)$$

$$\begin{cases} H_0 : \boldsymbol{\psi} \equiv (\theta_{\gamma 2}, \theta_{\beta 2}, \theta_{\alpha 2}, \theta_{\sigma 2}) = \mathbf{0}, \\ H_1 : \boldsymbol{\psi} \equiv (\theta_{\gamma 2}, \theta_{\beta 2}, \theta_{\alpha 2}, \theta_{\sigma 2}) \neq \mathbf{0}. \end{cases} \quad (3.9)$$

$$\begin{cases} H_0 : \boldsymbol{\psi} \equiv (\theta_{\gamma 3}, \theta_{\beta 3}, \theta_{\alpha 3}, \theta_{\sigma 3}) = \mathbf{0}, \\ H_1 : \boldsymbol{\psi} \equiv (\theta_{\gamma 3}, \theta_{\beta 3}, \theta_{\alpha 3}, \theta_{\sigma 3}) \neq \mathbf{0}. \end{cases} \quad (3.10)$$

Il sistema di ipotesi nella 3.7 consente di valutare congiuntamente la nullità dei coefficienti relativi all'intercetta che spiega γ, β, α e σ , definiti nella 2.7 e 2.8. Il sistema nella 3.8 considera invece i coefficienti associati all'intensità luminosa (I), quello nella 3.9 i coefficienti associati al fotoperiodo (D) e infine il sistema nella 3.10 riguarda i coefficienti associati alla concentrazione di fosforo (P). In ciascun caso, l'ipotesi nulla corrisponde al modello ridotto ottenuto imponendo simultaneamente a zero i quattro coefficienti considerati.

Anche nel secondo caso sono stati considerati separatamente quattro sistemi di ipotesi, ciascuno dei quali pone simultaneamente uguali a zero i tre coefficienti associati alle variabili I, D e P per ogni parametro che definisce la curva di crescita (γ, β, α) e la variabilità delle osservazioni (σ):

$$\begin{cases} H_0 : \boldsymbol{\psi} \equiv (\theta_{\gamma 1}, \theta_{\gamma 2}, \theta_{\gamma 3}) = \mathbf{0}, \\ H_1 : \boldsymbol{\psi} \equiv (\theta_{\gamma 1}, \theta_{\gamma 2}, \theta_{\gamma 3}) \neq \mathbf{0}. \end{cases} \quad (3.11)$$

$$\begin{cases} H_0 : \boldsymbol{\psi} \equiv (\theta_{\beta 1}, \theta_{\beta 2}, \theta_{\beta 3}) = \mathbf{0}, \\ H_1 : \boldsymbol{\psi} \equiv (\theta_{\beta 1}, \theta_{\beta 2}, \theta_{\beta 3}) \neq \mathbf{0}. \end{cases} \quad (3.12)$$

$$\begin{cases} H_0 : \boldsymbol{\psi} \equiv (\theta_{\alpha 1}, \theta_{\alpha 2}, \theta_{\alpha 3}) = \mathbf{0}, \\ H_1 : \boldsymbol{\psi} \equiv (\theta_{\alpha 1}, \theta_{\alpha 2}, \theta_{\alpha 3}) \neq \mathbf{0}. \end{cases} \quad (3.13)$$

$$\begin{cases} H_0 : \boldsymbol{\psi} \equiv (\theta_{\sigma 1}, \theta_{\sigma 2}, \theta_{\sigma 3}) = \mathbf{0}, \\ H_1 : \boldsymbol{\psi} \equiv (\theta_{\sigma 1}, \theta_{\sigma 2}, \theta_{\sigma 3}) \neq \mathbf{0}. \end{cases} \quad (3.14)$$

Il sistema di ipotesi nella 3.11 consente di valutare congiuntamente l'effetto dei 3 fattori sperimentali sulla resa finale della densità ottica, infatti nell'ipotesi nulla H_0 vengono annullati i coefficienti relativi a γ . Il sistema nella 3.12 valuta invece l'effetto dei tre fattori sulla velocità di crescita, infatti considera i coefficienti associati a β ; quello nella 3.13 ne valuta l'effetto sullo spostamento orizzontale della curva, infatti sono considerati i coefficienti associati ad α ; e infine il sistema nella 3.14 ne valuta l'effetto sulla variabilità delle osservazioni, infatti sono considerati i coefficienti associati a σ .

I risultati dei test sono riportati nella Tabella 3.2, dalla quale si osserva che tutti i sottoinsiemi di parametri considerati risultano significativi a un livello di significatività dell'1%. Questo risultato evidenzia la rilevanza complessiva sia dell'intercetta e dei tre

fattori sperimentali, intensità luminosa (I), fotoperiodo (D) e concentrazione di fosforo (P), singolarmente nel modello globale, sia congiuntamente dei tre fattori sui singoli parametri che definiscono la curva di crescita e la variabilità delle osservazioni (γ, β, α e σ).

Tabella 3.2: Risultati dei test LRV sui sottoinsiemi di parametri del modello globale, relativi alle condizioni sperimentali (definiti rispettivamente nella 3.7, 3.8, 3.9, 3.10) e ai parametri della curva di crescita e variabilità delle osservazioni (definiti rispettivamente nella 3.11, 3.12, 3.13, 3.14).

Sottoinsieme parametri nulli	Statistica test <i>LRV</i>	<i>p</i> -value
Intercetta	147.6682	< 0.001
Intensità luminosa (<i>I</i>)	695.8242	< 0.001
Fotoperiodo (<i>D</i>)	267.1819	< 0.001
Fosforo (<i>P</i>)	178.7079	< 0.001
I, D e P su γ	93.4759	< 0.001
I, D e P su β	83.0212	< 0.001
I, D e P su α	140.5473	< 0.001
I, D e P su σ	707.6262	< 0.001

Conclusioni

Il presente lavoro di tesi ha avuto come obiettivo l'analisi statistica della crescita di *Nostoc PCC 7120*, con particolare attenzione alla valutazione dell'influenza delle condizioni sperimentali sulla dinamica di crescita e, di conseguenza, sull'ottimizzazione della produzione di biomassa. L'analisi è stata condotta sui dati ottenuti mediante la piattaforma CellDEG, considerando differenti combinazioni dei fattori sperimentali relativi all'intensità luminosa, al fotoperiodo e alla concentrazione di fosforo.

Un primo obiettivo è stato quello di individuare un modello in grado di descrivere adeguatamente l'andamento temporale della densità ottica. L'analisi preliminare delle singole repliche ha evidenziato come un modello non lineare basato sulla funzione logistica rappresenti una scelta appropriata per descrivere la crescita osservata.

Successivamente, è stato sviluppato un modello globale nel quale i parametri della funzione logistica sono stati espressi in funzione dei fattori sperimentali. In questo modo è stato possibile valutare non solo l'andamento della crescita nel tempo, ma anche come le diverse condizioni di coltivazione influenzino la velocità di crescita, la resa finale e la variabilità delle osservazioni. Il modello è stato stimato mediante massimizzazione della funzione di log-verosimiglianza, assumendo una distribuzione normale per gli errori e utilizzando l'algoritmo L-BFGS-B. Nella fase di stima sono stati affrontati alcuni problemi numerici, causati dalla presenza di molteplici massimi locali e dalla sensibilità della funzione di log-verosimiglianza rispetto ai valori iniziali dei parametri. Per questo motivo sono state considerate numerose inizializzazioni e differenti configurazioni di vincoli, selezionando infine la soluzione associata al valore massimo della log-verosimiglianza tra quelle ottenute.

I risultati evidenziano un effetto complessivamente significativo delle tre variabili sperimentali sui parametri della curva logistica e sulla variabilità della risposta. In particolare, all'aumentare dell'intensità luminosa, del fotoperiodo e della concentrazione di fosforo au-

menta il livello massimo di densità ottica raggiunto dalla coltura, mentre la velocità di crescita aumenta con l'intensità luminosa e il fotoperiodo, senza evidenze di un effetto significativo della concentrazione di fosforo. Per quanto riguarda la variabilità, la varianza aumenta all'aumentare dell'intensità luminosa e del fotoperiodo, mentre diminuisce con l'aumentare della concentrazione di fosforo.

La valutazione della significatività dei parametri è stata svolta attraverso il test del log-rapporto di verosimiglianza, applicato sia ai singoli parametri sia a loro sottoinsiemi. Questo approccio ha permesso di verificare l'effettiva necessità dei diversi termini presenti nel modello e di fornire un'ulteriore valutazione dell'importanza dei fattori sperimentali. I risultati ottenuti hanno complessivamente confermato la rilevanza dei fattori considerati nella specificazione del modello globale.

Nel complesso, il lavoro ha quindi permesso di costruire un modello capace di integrare la dimensione temporale della crescita con le informazioni relative alle condizioni sperimentali. Questo rappresenta un aspetto particolarmente importante nel contesto della produzione di biomassa, poiché l'interesse non è rivolto esclusivamente alla quantità finale prodotta, ma anche alla velocità con cui essa viene raggiunta e alla variabilità associata al processo.

In conclusione i risultati ottenuti forniscono una base quantitativa per comprendere l'effetto delle condizioni di coltivazione sulla crescita di *Nostoc PCC 7120* e rappresentano un punto di partenza per successive analisi finalizzate all'ottimizzazione delle condizioni di produzione di biomassa a fini biotecnologici.

Bibliografia

- Azzalini, Adelchi (2001). *Inferenza statistica: una presentazione basata sul concetto di verosimiglianza*. ita. 2. ed. Springer collana di statistica. [Heidelberg] Milano: Springer. ISBN: 978-88-470-0130-5.
- Byrd, Richard H. et al. (set. 1995). «A Limited Memory Algorithm for Bound Constrained Optimization». en. In: *SIAM Journal on Scientific Computing* 16.5, pp. 1190–1208. ISSN: 1064-8275, 1095-7197. DOI: 10.1137/0916069. URL: <http://epubs.siam.org/doi/10.1137/0916069> (visitato il 01/08/2026).
- Grigoletto, Matteo, Laura Ventura e Francesco Pauli (2017). *Modello Lineare: Teoria e Applicazioni con R*. ita. Turin: G. Giappichelli. ISBN: 978-88-921-6531-1.
- Lucato, Veronica et al. (mag. 2024a). «From Nitrogen to Protein: Harnessing the Power of Nitrogen-Fixing Cyanobacteria for Protein-Rich Biomass Production». en. In: *ACS Sustainable Chemistry & Engineering* 12.20, pp. 7693–7702. ISSN: 2168-0485, 2168-0485. DOI: 10.1021/acssuschemeng.3c07548. URL: <https://pubs.acs.org/doi/10.1021/acssuschemeng.3c07548> (visitato il 14/07/2026).
- Lucato, Veronica et al. (giu. 2024b). «Modeling Biomass Growth and Nitrogen Fixation by Heterocystous Cyanobacteria Cultivated in Continuous Photobioreactors». en. In: *Industrial & Engineering Chemistry Research* 63.23, pp. 10113–10125. ISSN: 0888-5885, 1520-5045. DOI: 10.1021/acs.iecr.4c01193. URL: <https://pubs.acs.org/doi/10.1021/acs.iecr.4c01193> (visitato il 14/07/2026).
- Lucato, Veronica et al. (lug. 2025). «Optimization of Nostoc sp. biomass and protein composition in high-density cultivator based on DoE». en. In: *Algal Research* 89, p. 104088. ISSN: 22119264. DOI: 10.1016/j.algal.2025.104088. URL: <https://linkinghub.elsevier.com/retrieve/pii/S2211926425001973> (visitato il 14/07/2026).

Seber, George A. F. e Christopher J. Wild (2003). *Nonlinear regression*. eng. Wiley series in probability and mathematical statistics. New York: Wiley Interscience. ISBN: 978-0-471-61760-0 978-0-471-72531-2. DOI: 10.1002/0471725315.

Appendice A

Codice R

Librerie

```
1 library(ggplot2)
  library(tidyverse)
  library(future)
  library(future.apply)
```

Pulizia del dataset e analisi esplorativa

```
1 str(dati)
  for (i in 1:6) {
    print(table(dati[,i]))
  }
  dati$OD[dati$OD=="—"] <- NA
6 sum(is.na(dati$OD))
  dati$OD <- as.numeric(dati$OD)

  # Rinomino variabile tempo
  names(dati)[5] <- "Tempo"
11
  # Creazione variabile di replica
  replica <- rep(0, nrow(dati))
  i=1
  p=1
```

```
16 while (i <= nrow(dati)) {  
    replica[i:(i+6)]<- rep(p,7)  
    i = i+7  
    p = p+1  
}  
21 dati[,7] <- replica  
    names(dati)[7] <- "replica"  
  
    # Creazione variabile di condizione sperimentale  
    dati$condizioni <- dati$I + dati$D + dati$P  
26 for (i in 1:length(unique(dati$condizioni))) {  
    dati$condizioni[which(dati$condizioni==unique(dati$condizioni)[i])]   
        <- i  
    }  
    dati$condizioni lettere <- LETTERS[dati$condizioni]  
  
31 # Schema sperimentale  
    attach(dati)  
    table(I,D,P)/7  
    detach(dati)  
  
36 # Distribuzione di OD al variare del tempo  
    boxplot(OD~Tempo, data=dati.input)  
  
    # Distribuzione di OD al variare di I,D e P (Grafici dispersione)  
    # escludendo le osservazioni al tempo 0  
41 par(mfrow=c(2,3))  
    plot(OD~jitter(I,0.4), data=dati.input[dati.input$Tempo!=0,], xlab="I"  
        , xaxt="n")  
    axis(1, at=unique(dati.input$I))  
    plot(OD~jitter(D,0.4), data=dati.input[dati.input$Tempo!=0,], xlab="D"  
        , xaxt="n")  
    axis(1, at=unique(dati.input$D))  
46 plot(OD~jitter(P,0.4), data=dati.input[dati.input$Tempo!=0,], xlab="P"  
        , xaxt="n")
```

```

axis(1, at=unique(dati.input$P))

# Distribuzione di OD al variare di I,D e P (Boxplot)
# escludendo le osservazioni al tempo 0
51 boxplot(OD~I, data=dati.input[dati.input$Tempo!=0,])
   boxplot(OD~D, data=dati.input[dati.input$Tempo!=0,])
   boxplot(OD~P, data=dati.input[dati.input$Tempo!=0,])
   par(mfrow=c(1,1))

56 # Grafici di crescita algale per ogni replica
   dati$facet_label <- paste0(dati$replica, " — ", dati$condizioni_
     lettere)
   dati$facet_label <- factor(
     dati$facet_label,
     levels = unique(dati$facet_label[order(as.numeric(dati$replica))])
61 )
   ggplot(dati, aes(x = Tempo, y = OD)) +
     geom_line(color = "gray") +
     geom_point(size = 0.9) +
     facet_wrap(~ facet_label, ncol=6) +
66 theme_minimal(base_size = 9) +
     theme(
       strip.text = element_text(size = 9),
       axis.text = element_text(size = 7)
     )

```

Modello sulle singole repliche

```

# Funzione di log-verosimiglianza (negativa) del modello semplice
loglik_neg <- function(dati.f, par, t=0:6) {
  gamma <- par[1]
  beta <- par[2]
5  alpha <- par[3]
   sigma <- par[4]

  mu <- gamma*(exp(beta*t+alpha)/(1+exp(beta*t+alpha)))

```

```

# log-verosimiglianza negativa
10 -sum(dnorm(dati.f, mean = mu, sd = sigma, log = TRUE))
}

# Funzione per rappresentare graficamente i modelli
grafico <- function(OD, parametri, t=0:6, colore = 1, aggiungere = F,
  title = "", col.title=1, yLim=NULL, prev=F, prev_lim=NULL){
15 mu <- parametri[1] * (exp(parametri[2] * c(0:10) + parametri[3])/(1+
  exp(parametri[2] * c(0:10) + parametri[3])))
  if (aggiungere == T) {
    lines(0:6, mu[1:7], col = colore)
    if (prev==T) {
      lines(6:8, mu[7:9], col = 2, lty=2)
20 }
  } else {
    plot(t, OD, xlab = "Tempo", ylab = "OD", main = title,
      col.main=col.title, ylim=yLim, xlim = prev_lim)
    lines(0:6, mu[1:7], col = colore)
25 if (prev==T) {
      lines(6:8, mu[7:9], col = 2, lty=2)
    }
  }
}

30

# STIMA MODELLI PER OGNI REPLICA
# vincolo che sigma sia > 0.
par(mfrow=c(3,5))
35 i = 1
j = 1
start <- c(gamma = 1, beta = 1, alpha = -1, sigma = 1)
stime <- matrix(NA, nrow = 53, ncol = 4)
colnames(stime) <- c("gamma", "beta", "alpha", "sigma")
40 rownames(stime) <- 1:53
while (i<nrow(dati)) {

```

```

if (is.na(dati$OD[(i+6)])) {
  fit <- optim(start, loglik_neg, dati.f = dati$OD[i:(i+5)],
              t = 0:5, method = "L-BFGS-B",
45         lower = c(gamma = -Inf, beta = -Inf, alpha = -Inf,
                    sigma = 1e-6))
  grafico(OD = dati$OD[i:(i+5)], fit$par, t = 0:5, colore = 3,
          aggiungere = F, title = dati[i,]$facet_label)
  stime[j,] <- fit$par
50 } else {
  fit <- optim(start, loglik_neg, dati.f = dati$OD[i:(i+6)],
              method = "L-BFGS-B",
              lower = c(gamma = -Inf, beta = -Inf, alpha = -Inf,
                        sigma = 1e-6))
55   grafico(OD = dati$OD[i:(i+6)], fit$par, colore = 3,
           aggiungere = F, title = dati[i,]$facet_label)
  stime[j,] <- fit$par
}
i = i+7
60 j = j+1
}
par(mfrow=c(1,1))

65 # ANALISI ESPLORATIVA STIME DEI PARAMETRI
stime <- as.data.frame(stime)

# Grafici di dispersione e correlazione dei parametri
panel.cor <- function(x, y, digits = 2, prefix = "corr = ", cex.cor,
70   ...)
{
  par(usr = c(0, 1, 0, 1))
  r <- cor(x, y)
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste0(prefix, txt)
75   if(missing(cex.cor)) cex.cor <- 1.5

```

```
    text(0.5, 0.5, txt, cex = cex.cor)
  }
pairs(stime, upper.panel = panel.cor)

80 # Grafici di dispersione tra parametri e condizioni sperimentali
par(mfrow=c(2,3))
plot(dati2$I, dati2$gamma, main = "Gamma - I", xlab = "I", ylab = "
  Gamma")
plot(dati2$D, dati2$gamma, main = "Gamma - D", xlab = "D", ylab = "
  Gamma")
plot(dati2$P, dati2$gamma, main = "Gamma - P", xlab = "P", ylab = "
  Gamma")

85 plot(dati2$I, dati2$beta, main = "Beta - I", xlab = "I", ylab ="Beta")
plot(dati2$D, dati2$beta, main = "Beta - D", xlab = "D", ylab ="Beta")
plot(dati2$P, dati2$beta, main = "Beta - P", xlab = "P", ylab ="Beta")

90 plot(dati2$I, dati2$alpha, main = "Alpha - I", xlab = "I", ylab = "
  Alpha")
plot(dati2$D, dati2$alpha, main = "Alpha - D", xlab = "D", ylab = "
  Alpha")
plot(dati2$P, dati2$alpha, main = "Alpha - P", xlab = "P", ylab = "
  Alpha")

plot(dati2$I, dati2$sigma, main = "Sigma - I", xlab = "I", ylab = "
  Sigma")
95 plot(dati2$D, dati2$sigma, main = "Sigma - D", xlab = "D", ylab = "
  Sigma")
plot(dati2$P, dati2$sigma, main = "Sigma - P", xlab = "P", ylab = "
  Sigma")
par(mfrow=c(1,1))

# MODELLO LINEARE MULTIVARIATO CON LE STIME
100 fit.mult <- lm(cbind(gamma, beta, alpha, log(sigma)) ~ I + D + P, data
  = dati2)
```

```
summary(fit.mult)
```

Modello globale

```
# Funzione di log-verosimiglianza (negativa) del modello globale
loglik_neg_gen <- function(dati.y_tempo_IDP, par, fixed=rep(T,16)) {
  t <- dati.y_tempo_IDP[,2]
  I <- dati.y_tempo_IDP[,3]
  D <- dati.y_tempo_IDP[,4]
  P <- dati.y_tempo_IDP[,5]

  name <- c("theta0_g", "theta1_g", "theta2_g", "theta3_g",
            "theta0_b", "theta1_b", "theta2_b", "theta3_b",
            "theta0_a", "theta1_a", "theta2_a", "theta3_a",
            "theta0_s", "theta1_s", "theta2_s", "theta3_s")

  j = 1
  for (i in 1:length(fixed)) {
    if (fixed[i]) {
      assign(name[i],par[j])
      j = j+1
    } else {
      assign(name[i],0)
    }
  }

  gamma <- theta0_g + theta1_g*I + theta2_g*D + theta3_g*P
  beta <- theta0_b + theta1_b*I + theta2_b*D + theta3_b*P
  alpha <- theta0_a + theta1_a*I + theta2_a*D + theta3_a*P
  sigma <- exp(theta0_s + theta1_s*I + theta2_s*D + theta3_s*P)

  # trasformazione logistica numericamente stabile
  logistic <- plogis(beta * t + alpha)

  # mu
  mu <- gamma * logistic
```

```

# log-verosimiglianza negativa
34 -sum(dnorm(dati.y_tempo_IDP[,1], mean = mu, sd = sigma, log = TRUE))
}

# Funzione per calcolare diverse quantita' (modello)
fit.quantita <- function(fit, fixed=rep(T,16)){
39
  I <- dati2$I
  D <- dati2$D
  P <- dati2$P

44  I_long <- dati.input$I
  D_long <- dati.input$D
  P_long <- dati.input$P
  Tempo_long <- dati.input$Tempo

49  j=1
  Theta <- rep(0, 16)
  for (i in 1:length(fixed)) {
    if (fixed[i]) {
      Theta[i]<-fit$par[j]
54     j=j+1
    }
  }

  gamma_long <- Theta[1] + Theta[2]*I_long + Theta[3]*D_long + Theta
    [4]*P_long
59  beta_long <- Theta[5] + Theta[6]*I_long + Theta[7]*D_long + Theta
    [8]*P_long
  alpha_long <- Theta[10] + Theta[11]*I_long + Theta[12]*D_long +
    Theta[13]*P_long
  sigma_long <- exp(Theta[14] + Theta[15]*I_long + Theta[16]*D_long +
    Theta[17]*P_long)

  gamma <- Theta[1] + Theta[2]*I + Theta[3]*D + Theta[4]*P

```

```

64  beta <- Theta[5] + Theta[6]*I + Theta[7]*D + Theta[8]*P
    alpha <- Theta[10] + Theta[11]*I + Theta[12]*D + Theta[13]*P
    sigma <- exp(Theta[14] + Theta[15]*I + Theta[16]*D + Theta[17]*P)

    y_fitted <- gamma_long * (exp(beta_long * Tempo_long + alpha_long)/
      (1+exp(beta_long * Tempo_long + alpha_long)))
69  y <- dati.input$OD

    loglik <- -fit$value
    k <- length(fit$par)
    AIC <- 2 * k - 2 * loglik
74

    return(list(EQM = mean((y_fitted-y)^2),
      fitted.values = y_fitted,
      Theta = Theta,
      param = cbind(gamma,beta,alpha,sigma),
79      param_long = cbind(gamma_long,beta_long,alpha_long,
        sigma_long),

      AIC = AIC,
      logLik = loglik))
}

84
# Dati per input (senza dati mancanti)
dati.input <- cbind(dati[,6:5], dati[,2:4])
dati.input <- dati.input[!is.na(dati.input$OD),]

89
# FASE DI STIMA
# La funzione stima calcola tutte le stime del modello che si
  ottengono partendo da punti iniziali diversi e con limiti per i
  parametri diversi.
# In modello va inserita la logverosimiglianza negativa relativa al
  modello che si vuole stimare.
stima <- function(modello, fixed = rep(T,16)){
94   fixed.values <- fixed

```

```

coef <- t(fit.mult$coefficients)
rownames(coef)[4] <- c("sigma")

bound <- confint(fit.mult, level = 0.99)

99  l_bound <- bound[,2] - bound[,1]

n <- 50
r <- 25
104 start_matrix <- matrix(0, nrow = n, ncol = 16)
for (k in 1:r) {
  set.seed(k)
  start_matrix[k,] <- runif(16, -0.5, 0.5)
}
109 for (t in (r+1):n) {
  set.seed(t+250)
  start_matrix[t,] <- runif(16, -0.15, 0.15)
}

114 m <- 10
start_matrix2 <- matrix(0, nrow = m, ncol = 16)
r <- 0.01
start_matrix2[1,] <- c(theta0_g=coef[1,1]-r*(m/2),
  theta1_g=coef[1,2]-r*(m/2), theta2_g=coef[1,3]-r*(m/2),
119  theta3_g=coef[1,4]-r*(m/2),
  theta0_b=coef[2,1]-r*(m/2), theta1_b=coef[2,2]-r*(m/2),
  theta2_b=coef[2,3]-r*(m/2), theta3_b=coef[2,4]-r*(m/2),
  theta0_a=coef[3,1]-r*(m/2), theta1_a=coef[3,2]-r*(m/2),
  theta2_a=coef[3,3]-r*(m/2), theta3_a=coef[3,4]-r*(m/2),
124  theta0_s=coef[4,1]-r*(m/2), theta1_s=coef[4,2]-r*(m/2),
  theta2_s=coef[4,3]-r*(m/2), theta3_s=coef[4,4]-r*(m/2))

for (h in 2:m) {
  start_matrix2[h,] <- c(theta0_g=start_matrix2[1,1]+r,
129  theta1_g=start_matrix2[1,2]+r,
```

```

theta2_g=start_matrix2[1,3]+r, theta3_g=start_matrix2[1,4]+r,
theta0_b=start_matrix2[1,5]+r, theta1_b=start_matrix2[1,6]+r,
theta2_b=start_matrix2[1,7]+r, theta3_b=start_matrix2[1,8]+r,
134 theta0_a=start_matrix2[1,9]+r, theta1_a=start_matrix2[1,10]+r,
theta2_a=start_matrix2[1,11]+r, theta3_a=start_matrix2[1,12]+r,
theta0_s=start_matrix2[1,13]+r, theta1_s=start_matrix2[1,14]+r,
theta2_s=start_matrix2[1,15]+r, theta3_s=start_matrix2[1,16]+r)
r <- r+0.01
}
139
start <- rbind(start_matrix, start_matrix2)

lower.values <- rbind(
  c(theta0_g = -Inf, theta1_g = -Inf, theta2_g = -Inf, theta3_g = -Inf,
144   theta0_b = -Inf, theta1_b = -Inf, theta2_b = -Inf, theta3_b = -Inf,
   theta0_a = -Inf, theta1_a = -Inf, theta2_a = -Inf, theta3_a = -Inf,
   theta0_s = -7, theta1_s = -0.5, theta2_s = -0.5, theta3_s = -0.5),
  c(theta0_g = bound[1,1]-l_bound[1], theta1_g = bound[2,1]-l_bound[2],
   theta2_g = bound[3,1]-l_bound[3], theta3_g = bound[4,1]-l_bound[4],
   theta0_b = bound[5,1]-l_bound[5], theta1_b = bound[6,1]-l_bound[6],
   theta2_b = bound[7,1]-l_bound[7], theta3_b = bound[8,1]-l_bound[8],
149   theta0_a = bound[9,1]-l_bound[9], theta1_a = bound[10,1]-l_bound
   [10], theta2_a = bound[11,1]-l_bound[11], theta3_a = bound[12,1]-l_
   bound[12],
   theta0_s = bound[13,1]-l_bound[13], theta1_s = bound[14,1]-l_bound
   [14], theta2_s = bound[15,1]-l_bound[15], theta3_s = bound[16,1]-l_
   bound[16])
)
upper.values <- rbind(c(theta0_g = Inf, theta1_g = Inf, theta2_g = Inf,
154   theta3_g = Inf,
   theta0_b = Inf, theta1_b = Inf, theta2_b = Inf, theta3_b = Inf,
   theta0_a = Inf, theta1_a = Inf, theta2_a = Inf, theta3_a = Inf,
   theta0_s = 1, theta1_s = 1, theta2_s = 1, theta3_s = 1),
  c(theta0_g = bound[1,2]+l_bound[1], theta1_g = bound[2,2]+l_bound[2],
   theta2_g = bound[3,2]+l_bound[3], theta3_g = bound[4,2]+l_bound[4],

```

```

    theta0_b=bound[5,2]+l_bound[5], theta1_b=bound[6,2]+l_bound[6],
    theta2_b=bound[7,2]+l_bound[7], theta3_b=bound[8,2]+l_bound[8],
159   theta0_a=bound[9,2]+l_bound[9], theta1_a=bound[10,2]+l_bound[10],
    theta2_a=bound[11,2]+l_bound[11], theta3_a=bound[12,2]+l_bound[12],
    theta0_s=bound[13,2]+l_bound[13], theta1_s=bound[14,2]+l_bound
    [14], theta2_s=bound[15,2]+l_bound[15], theta3_s=bound[16,2]+l_
    bound[16])
  )
  stime <- list()
  valori.input <- list()
164  z <- 1
  for (i in 1:(n+m)) {
    for (j in 1:dim(lower.values)[1]) {
      for (l in 1:dim(upper.values)[1]) {
        stime[[z]] <- tryCatch(optim(start[i,][fixed.values], modello,
169         dati.y_tempo_IDP = dati.input, fixed=fixed.values,
        lower = lower.values[j,][fixed.values],
        upper = upper.values[l,][fixed.values], method = "L-BFGS-B",
        control = list(maxit = 4500)), error = function(e) NULL
      )
174     valori.input[[z]] <- list(start = start[i,][fixed.values],
                                lower.values = lower.values[j,][fixed.values],
                                upper.values = upper.values[l,][fixed.values])

    print(z)
    z = z+1
179   }
  }
}
return(list(stime=stime, valori.input=valori.input))
}
184
# Stima
fixed.values <- c(T,T,T,T, # gamma
                  T,T,T,T, # beta
                  T,T,T,T, # alpha

```

```

189         T,T,T,T) # sigma
stime_gen <- stima(loglik_neg_gen, fixed = fixed.values)

logVS_gen <- c()
for (i in 1:length(stime_gen$stime)) {
194   if(is.null(stime_gen$stime[[i]]$value)){
       logVS_gen[i] <- NA
   } else {
       logVS_gen[i] <- -stime_gen$stime[[i]]$value
   }
199 }
table(is.na(logVS_gen))

# Ranking log-verosimiglianza
logVS_gen <- cbind(logVS_gen, seq(1,length(logVS_gen),1))
204 logVS_gen <- logVS_gen[order(logVS_gen[,1], decreasing = T),]
logVS_gen[1:10,]

# Verifica convergenze
messages <- sapply(stime_gen$stime, function(x) x$message)
209 messages <- unlist(messages)
table(messages)
convergence <- sapply(stime_gen$stime, function(x) x$convergence)
convergence <- unlist(convergence)
table(convergence)
214 convergence240 <- convergence[1:16]
convergence240[17]<- 100
convergence240[18:28] <- convergence[17:27]
convergence240[29:30]<- 100
convergence240[31:76] <- convergence[28:73]
219 convergence240[77]<- 100
convergence240[78:80] <- convergence[74:76]
convergence240[81]<- 100
convergence240[82:240] <- convergence[77:235]

```

```
224 # Calcolo delle quantita' del modello
modello1 <- fit.quantita_P2(stime_gen$stime[[120]], fixed.values)

# ANALISI DEGLI INTORNI
229 # Definizione seeds per gli intorni dei parametri ottimi
seeds.intorni_par <- matrix(NA, ncol = 1000, nrow = 240)
for (i in 1:240) {
  for (t in 1:1000) {
    seeds.intorni_par[i,t] <- ((t^2)+i)
234   }
}

# Grafico parametri ottimali e intorno della soluzione migliore
name_param <- c("theta0_g", "theta1_g", "theta2_g", "theta3_g",
239   "theta0_b", "theta1_b", "theta2_b", "theta3_b",
   "theta0_a", "theta1_a", "theta2_a", "theta3_a",
   "theta0_s", "theta1_s", "theta2_s", "theta3_s")
plot(stime_gen_1$stime[[1]]$par, type = "l", ylim=c(-5,3),
     col="gray", ylab = "Valore parametro", xlab = "", xaxt = "n",
244   main = "Parametri ottimali e intorno della soluzione migliore")
axis(1, at = 1:16, labels = name_param, las=2, cex.axis=0.7)
for (i in which(convergence240==0)) {
  lines(stime_gen_1$stime[[i]]$par, col="gray")
}
249 n.par <- 120
for (i in 1:1000) {
  set.seed(seeds.intorni_par[n.par,i])
  lines(stime_gen_1$stime[[n.par]]$par+runif(16, -0.1,0.1), col=4,
        lty=1)
254 }
lines(modello1$Theta[-9], type = "l", col="gray")
```

```

259 # Funzione per il calcolo delle logVerosimiglianze di tutti gli
    intorni dei parametri ottimi
logL_intorni_par <- function(loglik_Neg, fixed = rep(T,16), stime_gen)
{
  fixed.values <- fixed

  intorni <- array(NA, dim = c(1000,sum(fixed.values),240))
264 for (i in 1:240) {
    for (t in 1:1000) {
      set.seed((t^2)+i)
      if (is.null(stime_gen$stime[[i]])) {
        intorni[t, ,i] <- rep(NA,sum(fixed.values))
269      } else {
        intorni[t, ,i] <- stime_gen$stime[[i]]$par +
          runif(sum(fixed.values), -0.1,0.1 )
      }
    }
  }
274 }

  logL <- matrix(NA, nrow = 240, ncol = 1000)
  for (i in 1:nrow(logL)) {
    for (j in 1:ncol(logL)) {
279      logL[i,j] <- loglik_Neg(dati.y_tempo_IDP = dati.input,
        par = intorni[j,,i], fixed = fixed.values)
    }
    #print(i)
  }
284 return(logL)
}

l_intorni_par <- logL_intorni_par(loglik_Neg = loglik_neg_gen,
  stime_gen = stime_gen, fixed = fixed.values)
289 l_intorni_par <- -l_intorni_par

lVS_par_mean <- apply(l_intorni_par, 1, mean)

```

```

max(l_intorni_par, na.rm = T)

294 # Grafico dei valori della funzione di log-verosimiglianza attorno ai
    139 parametri ottimi (arrivati a convergenza)
plot(sort(l_intorni_par[1,]), ylim = c(-15000, -320), pch=".",
      cex=1.5, col=2, xlab = "Punti", ylab = "log-verosimiglianza",
      main = "Log-verosimiglianza nei punti intorno ai massimi locali",
      yaxt = "n")
299 axis(2, at = c(-15000, -10000, -5000, -329), labels = c(-15000,
    -10000, -5000, -329))
for (i in which(convergence240==0)) {
  points(sort(l_intorni_par[i,]), col=i, pch=".", cex=1.5)
}
points(c(rep(-30000,1001),modello1$logLik), pch=16, cex=0.6)
304 lines(sort(l_intorni_par[120,]), lwd=2)

# STRATEGIA SEMPLIFICAZIONE FUNZIONE DI VEROSIMIGLIANZA
# Parallelizzazione
parallel::detectCores()
309 plan(multisession, workers = 3)

fixed.list <- list(
  c(F,T,T,T, T,T,T,T, T,T,T,T, T,T,T,T),
  c(T,F,T,T, T,T,T,T, T,T,T,T, T,T,T,T),
314 c(T,T,F,T, T,T,T,T, T,T,T,T, T,T,T,T),
  c(T,T,T,F, T,T,T,T, T,T,T,T, T,T,T,T),
  c(T,T,T,T, F,T,T,T, T,T,T,T, T,T,T,T),
  c(T,T,T,T, T,F,T,T, T,T,T,T, T,T,T,T),
  c(T,T,T,T, T,T,F,T, T,T,T,T, T,T,T,T),
319 c(T,T,T,T, T,T,T,F, T,T,T,T, T,T,T,T),
  c(T,T,T,T, T,T,T,T, F,T,T,T, T,T,T,T),
  c(T,T,T,T, T,T,T,T, T,F,T,T, T,T,T,T),
  c(T,T,T,T, T,T,T,T, T,T,F,T, T,T,T,T),
  c(T,T,T,T, T,T,T,T, T,T,T,F, T,T,T,T),
324 c(T,T,T,T, T,T,T,T, T,T,T,T, F,T,T,T),

```

```

    c(T,T,T,T, T,T,T,T, T,T,T,T, T,F,T,T),
    c(T,T,T,T, T,T,T,T, T,T,T,T, T,T,F,T),
    c(T,T,T,T, T,T,T,T, T,T,T,T, T,T,T,F)
  )
329 stime_pvalue <- future_lapply(seq_along(fixed.list), function(i){

    stima(
      loglik_neg_gen,
      fixed = fixed.list[[i]]
334   )

  })

# Estrarre le logvs per ogni modello
339 name_lvs.H <- paste0("logVS_", name_param)
lvs.H_list <- list()
for (i in 1:16) {
  lvs.H_list[[i]] <- rep(NA,length(stime_pvalue[[i]]$stime))
  for (j in 1:length(stime_pvalue[[i]]$stime)) {
344    if(!is.null(stime_pvalue[[i]]$stime[[j]]$value)){
      lvs.H_list[[i]][j] <- -stime_pvalue[[i]]$stime[[j]]$value
    }
  }
}
349 names(lvs.H_list) <- name_lvs.H

for (i in 1:16) {
  print(table(is.na(lvs.H_list[[i]])))
}
354

# Ranking verosimiglianze
for (i in 1:16) {
  cat(name_param[i], lvs.H_list[[i]][order(lvs.H_list[[i]],
    decreasing = T)][1:4], "\n")
359 }

```

```

# verifica convergenza
for (i in 1:16) {
  print(stime_pvalue[[i]]$stime[[which.max(lvs.H_list[[i]])]]$message)
364 }

# Calcolo le quantita per i modelli
modelli_pvalue <- list()
for (i in 1:16) {
369   modelli_pvalue[[i]] <- fit.quantita(stime_pvalue[[i]]$stime[[which.
     max(lvs.H_list[[i]])]], fixed = fixed.list[[i]])
}
names(modelli_pvalue) <- paste0("modello_", name_param)

# RICERCA MASSIMA VEROSIMIGLIANZA
374 # parametri ottimi da usare come start
modelli_pvalue$modello_theta3_b$Theta

cercaMax <- function(modello, fixed = rep(T,16)){
  fixed.values <- fixed
379   coef <- t(fit.mult$coefficients)
  rownames(coef)[4] <- c("sigma")

  bound <- confint(fit.mult, level = 0.99)

384   l_bound <- bound[,2] - bound[,1]

  lower.values <- rbind(
    c(theta0_g = -Inf, theta1_g = -Inf, theta2_g = -Inf, theta3_g = -Inf,
      theta0_b = -Inf, theta1_b = -Inf, theta2_b = -Inf, theta3_b = -Inf,
389     theta0_a = -Inf, theta1_a = -Inf, theta2_a = -Inf, theta3_a = -Inf,
      theta0_s = -7, theta1_s = -0.5, theta2_s = -0.5, theta3_s = -0.5),
    c(theta0_g = bound[1,1] - l_bound[1], theta1_g = bound[2,1] - l_bound[2],
      theta2_g = bound[3,1] - l_bound[3], theta3_g = bound[4,1] - l_bound[4],
      theta0_b = bound[5,1] - l_bound[5], theta1_b = bound[6,1] - l_bound[6],

```

```

394   theta2_b=bound[7,1]-l_bound[7], theta3_b=bound[8,1]-l_bound[8],
      theta0_a=bound[9,1]-l_bound[9], theta1_a=bound[10,1]-l_bound[10],
      theta2_a=bound[11,1]-l_bound[11], theta3_a=bound[12,1]-l_bound[12],
      theta0_s=bound[13,1]-l_bound[13], theta1_s=bound[14,1]-l_bound[14],
      theta2_s=bound[15,1]-l_bound[15], theta3_s=bound[16,1]-l_bound[16])
399 )
      upper.values <- rbind(c(theta0_g=Inf, theta1_g=Inf, theta2_g=Inf,
                             theta3_g=Inf,
                             theta0_b=Inf, theta1_b=Inf, theta2_b=Inf, theta3_b=Inf,
                             theta0_a=Inf, theta1_a=Inf, theta2_a=Inf, theta3_a=Inf,
404     theta0_s=1, theta1_s=1, theta2_s=1, theta3_s=1),
      c(theta0_g=bound[1,2]+l_bound[1], theta1_g=bound[2,2]+l_bound[2],
        theta2_g=bound[3,2]+l_bound[3], theta3_g=bound[4,2]+l_bound[4],
        theta0_b=bound[5,2]+l_bound[5], theta1_b=bound[6,2]+l_bound[6],
        theta2_b=bound[7,2]+l_bound[7], theta3_b=bound[8,2]+l_bound[8],
        theta0_a=bound[9,2]+l_bound[9], theta1_a=bound[10,2]+l_bound[10],
        theta2_a=bound[11,2]+l_bound[11], theta3_a=bound[12,2]+l_bound[12],
        theta0_s=bound[13,2]+l_bound[13], theta1_s=bound[14,2]+l_bound
        [14], theta2_s=bound[15,2]+l_bound[15], theta3_s=bound[16,2]+l_
        bound[16])
409 )
      stime <- list()
      valori.input <- list()
      z <- 1
      for (j in 1:dim(lower.values)[1]) {
414         for (l in 1:dim(upper.values)[1]) {
            stime[[z]] <- tryCatch(optim(modelli_pvalue$modello_theta3_b$
              Theta[fixed.values], modello,
              dati.y_tempo_IDP = dati.input,
              fixed=fixed.values,
419             lower = lower.values[j,][fixed.values],
              upper = upper.values[l,][fixed.values],
              method = "L-BFGS-B",
              control = list(maxit = 4500)),
              error = function(e) NULL

```

```

424     )
        valori.input[[z]] <- list(start = modelli_pvalue$
                                modello_theta3_b$Theta[fixed.values],
                                lower.values = lower.values[j,][fixed.values],
                                upper.values = upper.values[1,][fixed.values])
429     print(z)
        z = z+1
    }
}
return(list(stime=stime, valori.input=valori.input))
434 }

stime_MAXvs <- cercaMax(loglik_neg_gen, fixed = fixed.values)

modello1.new <- fit.quantita(stime_MAXvs$stime[[4]],
439     fixed = fixed.values)
modello1.new$logLik

# GRAFICO per ogni condizione
444 titolo_condizioniSperimentali <- LETTERS[1:17]
par(mfrow=c(2,3))
for (i in 1:length(unique(dati$condizioni))) {
    grafico(OD = dati$OD[dati$condizioni==i], unique(modello1.new$param)
            [i,], t = dati$Tempo[dati$condizioni==i], colore = 4, aggiungere =
            F, title = titolo_condizioniSperimentali[i], yLim = c(0,20))
}
449 par(mfrow=c(1,1))

# Grafici del modello globale con previsione
par(mfrow=c(2,3))
454 for (i in 1:length(unique(dati$condizioni))) {
    grafico(OD = dati$OD[dati$condizioni==i], unique(modello1.new$param)
            [i,], t = dati$Tempo[dati$condizioni==i], colore = 4, aggiungere =

```

```

    F, title = titolo_condizioniSperimentali[i], yLim = c(0,20), prev =
      T, prev_lim = c(0,8))
}
par(mfrow=c(1,1))

```

Test congiunti LRV

```

# TEST LOG-RAPPORTO DI VEROSIMIGLIANZA
# Inserire in input il risultato di fit.quantita (chiamati modello...)
  in mod.completo e mod.ridotto
3 testLRV <- function(mod.completo, mod.ridotto){
  # numero di parametri stimati
  k1 <- length(mod.completo$Theta[mod.completo$Theta!=0])
  k0 <- length(mod.ridotto$Theta[mod.ridotto$Theta!=0])

8  # test statistic
  LR <- 2 * (mod.completo$logLik - mod.ridotto$logLik)
  #print(LR)

  # gradi di liberta'
13 df <- k1 - k0
  #print(df)

  # p-value
  return(1 - pchisq(LR, df))
18 }

# P-VALUE
p_value <- c()
for (i in 1:16) {
23   p_value[i] <- testLRV(modello1.new, modelli_pvalue[[i]])
}

# modello1.new con p-value
output <- cbind(modello1.new$Theta, p_value)
28 row.names(output) <- name_param

```

```

colnames(output)[1] <- "coef"
output

# Inserimento statistica test negli output
33 calculate.stat.test <- function(mod.completo, mod.ridotto){
  LR <- 2 * (mod.completo$logLik - mod.ridotto$logLik)
  return(LR)
}

38 stat.test <- c()
for (i in 1:16) {
  stat.test[i] <- calculate.stat.test(modello1.new,
                                     modelli_pvalue[[i]])
}

43 output[,2] <- stat.test
colnames(output)[2] <- "stat.test"
output <- cbind(output, p_value)
output

48
# TEST CONGIUNTI
# modelli senza sottoinsiemi di parametri
fixed.list_blocchi <- list(
  c(F,T,T,T, # gamma
53   F,T,T,T, # beta
    F,T,T,T, # alpha
    F,T,T,T), # sigma
  c(T,F,T,T, # gamma
    T,F,T,T, # beta
58   T,F,T,T, # alpha
    T,F,T,T), # sigma
  c(T,T,F,T, # gamma
    T,T,F,T, # beta
    T,T,F,T, # alpha
63   T,T,F,T), # sigma

```

```

c(T,T,T,F, # gamma
  T,T,T,F, # beta
  T,T,T,F, # alpha
  T,T,T,F), # sigma
68 c(T,F,F,F, # gamma
    T,T,T,T,F, # beta
    T,T,T,T, # alpha
    T,T,T,T), # sigma
73 c(T,T,T,T, # gamma
    T,F,F,F,F, # beta
    T,T,T,T, # alpha
    T,T,T,T), # sigma
c(T,T,T,T, # gamma
  T,T,T,T,F, # beta
78 T,F,F,F, # alpha
  T,T,T,T), # sigma
c(T,T,T,T, # gamma
  T,T,T,T,F, # beta
  T,T,T,T, # alpha
83 T,F,F,F) # sigma
)

# Parallelizzazione
plan(multisession, workers = 3)
88
stime_blocchi.par <- future_lapply(seq_along(fixed.list_blocchi),
  function(i){
    stima(
      loglik_neg_gen,
      fixed = fixed.list_blocchi[[i]]
93 )
  })

name_stime_blocchi <- paste0("stime_H.", c("intercetta", "I", "D",
                                           "P", "gamma", "beta", "alpha", "sigma"))

```

```

98 names(stime_blocchi.par) <- name_stime_blocchi

# Estrarre le logvs per ogni modello
name_lvs.H_blocchi <- paste0("logVS_", c("intercetta", "I", "D",
                                         "P", "gamma", "beta", "alpha", "sigma"))
103 lvs.H_list_blocchi <- list()
for (i in 1:8) {
  lvs.H_list_blocchi[[i]] <- rep(NA, length(stime_blocchi.par[[i]]$
    stime))
  for (j in 1:length(stime_blocchi.par[[i]]$stime)) {
    if(!is.null(stime_blocchi.par[[i]]$stime[[j]]$value)){
108     lvs.H_list_blocchi[[i]][j] <- -stime_blocchi.par[[i]]$stime[[j]]
      $value
    }
  }
}
names(lvs.H_list_blocchi) <- name_lvs.H_blocchi

113 # Ranking verosimiglianze
for (i in 1:8) {
  cat(c("intercetta", "I", "D", "P", "gamma", "beta", "alpha", "sigma"
    )[i], lvs.H_list_blocchi[[i]][order(lvs.H_list_blocchi[[i]],
    decreasing = T)][1:4], "\n")
}

118 # Verifica convergenza
for (i in 1:8) {
  print(stime_blocchi.par[[i]]$stime[[which.max(lvs.H_list_blocchi[[i]]
    )]]$message)
}

123 # Massima verosimiglianza con algoritmo che converge
logVS_H.sigma <- lvs.H_list_blocchi_conIntercetta[[4]]
logVS_H.sigma <- cbind(logVS_H.sigma, seq(1, length(logVS_H.sigma), 1))
logVS_H.sigma <- logVS_H.sigma[order(logVS_H.sigma[,1], decreasing = T

```

```

    ),]
128 logVS_H.sigma[1:10,]
    stime_blocchi.par_conIntercetta[[4]]$stime[[103]]$message
    # l indice 103 e' relativo a quello che converge

    # Calcolo le quantita per i modelli senza sottoinsiemi di parametri
133 modelli_blocchi.par <- list()
    for (i in 1:7) {
        modelli_blocchi.par[[i]] <- fit.quantita(stime_blocchi.par[[i]]$
            stime[[which.max(lvs.H_list_blocchi[[i]])]],
            fixed = fixed.list_blocchi[[i]])
    }
138 modelli_blocchi.par[[8]] <- fit.quantita(stime_blocchi.par[[8]]$stime
    [[103]], fixed = fixed.list_blocchi[[8]])

    names.blocchipar <- paste0("modello_", c("intercetta", "I", "D", "P",
        "gamma", "beta", "alpha", "sigma"))
    names(modelli_blocchi.par) <- names.blocchipar

143 # Test LRV per modelli senza sottoinsiemi di parametri
    p_value_blocchi <- c()
    for (i in 1:8) {
        p_value_blocchi[i] <- testLRV(modello1.new,
            modelli_blocchi.par[[i]])
148 }

    names(p_value_blocchi) <- names.blocchipar
    output.blocchi <- t(t(p_value_blocchi))
    colnames(output.blocchi) <- "p-value"
153 output.blocchi

    stat.test_blocchi <- c()
    for (i in 1:8) {
        stat.test_blocchi[i] <- calculate.stat.test(modello1.new,
158             modelli_blocchi.par[[i]])
    }

```

```
}  
  
output.blocchi[,1] <- stat.test_blocchi  
colnames(output.blocchi)[1] <- "stat.test"  
163 output.blocchi <- cbind(output.blocchi, p_value_blocchi)  
colnames(output.blocchi)[2] <- "p-value"  
output.blocchi
```