Sviluppo di reti probabilistiche di supporto alle decisioni in viticoltura ed enologia

- La linea B.4

Federico M. Stefanini

Dipartimento di Statistica 'G. Parenti',Università degli Studi di Firenze

Viale Morgagni 59, I-50134, Firenze, Italia

Tel: +39 055 4237266 Fax: +39 055 4223560

stefanini@ds.unifi.it http://www.ds.unifi.it/stefanini/

In collaborazione con: Ottorino-Luca Pantani

Università di Firenze, Dipartimento Scienza del Suolo e Nutrizione della Pianta

P.zle Cascine 28 50144 Firenze Italia

Tel 39 055 3288 202 (348 lab) Fax 39 055 333 273

6 dicembre 2010

Indice

1 La linea di ricerca B.4

La linea sperimentale B.4 del progetto *Tuscania* si propone di studiare e confrontare tecniche di vinificazione su uve Sangiovese per la produzione di vini da invecchiamento. In particolare, vengono considerate le capacità estrattive di due diverse macerazioni prefermentative, quali la *macerazione a freddo*, che prevede un raffreddamento piuttosto lento del pigiato con l'utilizzo di scambiatori di calore, e la *criomacerazione*, che sfrutta l'azione dell'azoto liquido in flusso continuo, pratica attualmente non diffusa in ambito industriale.
L'obiettivo dello studio è di valutare quantitativamente l'effetto di tali tecniche di vinificazione sul risultato enologico finale, in interazione con due diversi livelli di temperatura di fermentazione e con la pratica del salasso.
In particolare, questa relazione ha l'obiettivo di studiare le cinetiche di fermentazione di ciascun trattamento sperimentale della linea di ricerca B.4, combinazione dei seguenti fattori:

- **macerazione prefermentativa**: non trattato (*test*), macerazione a freddo (*mpf*) o crioestrazione (*crio*);
- **temperatura di fermentazione**: 20°o 30°;
- **salasso**: non effettuato (S0) o effettuato (S1).

Ogni trattamento è replicato tre volte. Nella Tabella 1 sono schematizzati i trattamenti previsti dall'esperimento e nella Tabella 2 sono elencate le variabili misurate in funzione del tempo.

Tabella 1: Trattamenti sperimentali della linea B.4

Tesi	Macerazione	Temperatura	Salasso
1	nessuna	20°	effettuato
2	a freddo	20°	effettuato
3	crioestrazione	20°	effettuato
4	nessuna	30°	effettuato
5	a freddo	30°	effettuato
6	crioestrazione	30°	effettuato
7	nessuna	20°	non effettuato
8	a freddo	20°	non effettuato
9	crioestrazione	20°	non effettuato
10	nessuna	30°	non effettuato
11	a freddo	30°	non effettuato
12	crioestrazione	30°	non effettuato

Tabella 2: Variabili misurate in funzione del tempo nella linea di ricerca B.4.

Label	Variabile
IC	Intensità Colorante
TON	Tonalità
ALTA	Antociani Liberi + Tannini
TAT	Complessi Tannino-Antociano-Tannino
FLAV	Flavonoidi
FNA	Flavonoidi Non Antocianici
ANTOT	Antociani Totali

2 Metodi

Al cuore della metodologia statistica Bayesiana troviamo la rappresentazione dell'incertezza, quindi dell'informazione, attraverso opportuni valori di probabilità assegnati agli eventi partizione di un conveniente spazio (Bernardo e Smith, 1994, Lindley, 2000).

In questo approccio all'inferenza i parametri sono trattati come variabili casuali, perciò hanno una distribuzione a priori (iniziale) prima che i dati siano osservati, ed una distribuzione a posteriori (finale) ottenuta integrando le evidenze fornite dai dati attraverso la funzione di verosimiglianza con l'informazione iniziale (O'Hagan, 1994, par. 1.15).

Per aderire pienamente al metodo Bayesiano si richiede l'adozione dell'interpretazione soggettiva della probabilità. Essa quantifica il grado di plausibilità (*belief*) di una proposizione, pertanto è una misura del convincimento raggiunto da una data persona che ha impiegato tutta l'informazione disponibile in maniera coerente (O'Hagan, 1994, par. 1.16, Lindley, 1987).

La nozione di probabilità soggettiva merita una trattazione ben più estesa (Berger, 1985, cap.3, O'Hagan, 1994, cap. 4, Bernardo, Smith, 1994, cap 2 e 3), in particolare in relazione alle altre possibili definizioni di probabilità ed alla loro efficacia nel campo statistico applicato (Barnett, 1982).

Nel seguito sono richiamati i soli elementi dell'approccio Bayesiano che sono direttamente richiesti per una piena comprensione di questo lavoro.

Il metodo Bayesiano può essere sintetizzato nei seguenti passi principali (O'Hagan 1994, par. 1.14):

1) definizione della funzione di verosimiglianza attraverso la specificazione del processo generatore dei dati $p(y \mid \theta)$ come funzione del parametro θ;

2) specificazione della distribuzione a priori $\pi(\theta)$ del parametro incognito (inteso come variabile casuale) la quale riassume esaustivamente quanto noto circa θ prima che siano osservati i dati; essa è detta anche distribuzione iniziale del parametro;

3) ottenimento della distribuzione a posteriori $p(\theta \mid y)$ del parametro attraverso l'applicazione del teorema di Bayes; la distribuzione così ottenuta, detta anche distribuzione finale del parametro, riassume esaustivamente l'informazione disponibile circa il parametro θ dopo avere osservato i dati;

4) costruzione di affermazioni inferenziali basate sulla distribuzione finale, ad esempio mediante il calcolo di statistiche che descrivono gli aspetti di $p(\theta \mid y)$ oggetto d'interesse.

Sia θ il parametro (eventualmente non scalare) della funzione di verosimiglianza $p(y \mid \theta)$, cioé della funzione che descrive la dipendenza dei valori osservati y dai valori del parametro θ. Sia inoltre $\pi(\theta)$ la distribuzione a priori del parametro θ. Applicando il teorema di Bayes si ha

$$p(\theta \mid y) = \frac{\pi(\theta) \cdot p(y \mid \theta)}{\int \pi(\theta) \cdot p(y \mid \theta) \cdot d\theta} = \frac{p(\theta, y)}{\int p(\theta, y) \cdot d\theta} \propto \pi(\theta) \cdot p(y \mid \theta)$$

Il termine alla estrema destra indica che la distribuzione a posteriori è proporzionale al prodotto della funzione di verosimiglianza per la distribuzione a priori del parametro. La forma della distribuzione a posteriori non risulta modificata se la verosimiglianza è moltiplicata per una costante o per una funzione $h(y)$ delle sole osservazioni (O'Hagan, 1994, par. 3.2). Pertanto i termini costanti o di tipo $h(y)$ che moltiplicano la funzione di verosimiglianza possono essere omessi nel calcolo di $p(\theta \mid y)$.

In alcune circostanze è conveniente fare riferimento alla quantità $\pi(\theta) \cdot p(y \mid \theta)$, detta distribuzione a posteriori non normalizzata. L'integrale che compare a denominatore nella formula di Bayes a volte non possiede una espressione in forma chiusa, ed in tali casi si deve ricorrere a metodi alternativi di valutazione, quali l'integrazione numerica ed i metodi di simulazione Monte Carlo.

Il teorema di Bayes può essere applicato in maniera sequenziale (O'Hagan, 1994, par. 3.5) a sottoinsiemi dei dati osservati che sono indipendenti, senza causare variazioni nella distribuzione finale del parametro.

Sia $y = (x_1, x_2)$ una delle possibili suddivisioni di y in parti indipendenti, e siano $\pi(\theta)$ e $\pi(\theta \mid x_1)$ rispettivamente la distribuzione a priori del parametro prima di avere osservato i dati e dopo avere osservato x_1. Allora la seguente espressione sottolinea l'equivalenza tra l'analisi effettuata in due fasi e quella effettuata senza suddividere i dati

$$\begin{gathered} p(\theta \mid y) \propto \pi(\theta) \cdot p(y \mid \theta) = \pi(\theta) \cdot p(x_1, x_2 \mid \theta) = \pi(\theta) \cdot p(x_1 \mid \theta) \cdot p(x_2 \mid \theta) \\ \Updownarrow \\ p(\theta \mid y) \propto \pi(\theta \mid x_1) \cdot p(x_2 \mid \theta) \\ \Updownarrow \\ p(\theta \mid y) \propto \pi(\theta \mid x_2) \cdot p(x_1 \mid \theta) \end{gathered}$$

È opportuno sottolineare che le distribuzioni

$$\pi(\theta \mid x_j), \quad j = 1, 2$$
$$\pi(\theta \mid x_j) \propto \pi(\theta) \cdot p(x_j \mid \theta)$$

sono utilizzate come distribuzioni a priori nella fase di analisi rimanente, cioé dopo che la prima parte dei dati è stata analizzata. Inoltre, la distribuzione a posteriori $p(\theta \mid y)$ diverrà la distribuzione a priori $\pi(\theta \mid y)$ nell'analisi di futuri dati sperimentali y_f. Per questo motivo la distribuzione a priori è a volte indicata con la stessa simbologia della distribuzione a posteriori, cioè $p(\theta)$.

Un parametro richiesto per la piena specificazione del modello ma che non riveste interesse inferenziale è detto parametro di disturbo (*nuisance parameter*, O'Hagan, 1994, par. 3.13). La distribuzione a posteriori di interesse è ottenuta attraverso la marginalizzazione della distribuzione a posteriori rispetto i parametri di disturbo.

Sia $\theta = (\theta_1, \theta_2)$ un parametro continuo bivariato, dove θ_2 è il parametro di disturbo. La distribuzione a posteriori richiesta per l'inferenza è

$$p(\theta_1 \mid y) = \int p(\theta_1, \theta_2 \mid y) \quad d\theta_2$$

Le distribuzioni $p(\theta)$ e $p(\theta \mid y)$ indicano la plausibilità associata ai valori che il parametro può assumere rispettivamente prima e dopo l'osservazione dei dati. Se i valori del parametro dotati di maggior plausibilità iniziale sono estremamente differenti da quelli che i dati osservati indicano come i più plausibili allora si incorre in una situazione di conflitto informativo (O'Hagan, 1994, par. 3.35).

Un caso semplice di conflitto informativo è dato da $p(\theta)$ e $p(y \mid \theta)$ entrambe di forma Gaussiana, con media e varianza tali che le due curve risultano possedere una piccolissima area di sovrapposizione (intuitivamente, le due distribuzioni sono "ben separate").

In tali circostanze, la distribuzione finale $p(\theta \mid y)$ è costituita da valori del parametro che risultano estremamente plausibili benché essi non lo siano secondo la sola distribuzione iniziale oppure la sola verosimiglianza. In altri termini, la forma distributiva di $p(\theta)$ e di $p(y \mid \theta)$ in regioni caratterizzate da bassa plausibilità determina le principali caratteristiche della distribuzione finale $p(\theta \mid y)$, ed un cambiamento anche minimo della forma che $p(\theta)$ e $p(y \mid \theta)$ assumono in tali regioni può comportare drastici cambiamenti nella forma della distribuzione finale $p(\theta \mid y)$.

In situazioni di conflitto informativo si richiede un'attenta revisione dell'intero modello, ed eventualmente l'effettuazione dell'analisi di sensibilità (spiegata nel seguito).

Un modello Bayesiano può essere assimilato ad un metodo particolare (ad esempio, un metodo in cui i parametri sono variabili casuali) per specificare le caratteristiche del processo che ha generato i dati osservati. Dopo aver ottenuto la distribuzione finale $p(\theta \mid y)$, è possibile ricavare la distribuzione delle osservazioni future y_f condizionatamente alle osservazioni disponibili y. Questo tipo di inferenza, detta predittiva (O'Hagan, 1994, par. 3.58), si effettua ottenendo la distribuzione delle osservazioni future y_f condizionatamente alle evidenze relative ai dati disponibili y

$$p(y_f \mid y) = \int p(y_f \mid \theta, y) \cdot p(\theta \mid y) \quad d\theta$$

in cui la verosimiglianza $p(y_f \mid \theta, y) = p(y_f \mid \theta)$ se i dati sono indipendenti condizionatamente al valore del parametro θ.

Se il modello è una valida spiegazione del processo generatore dei dati, cioè complessivamente in accordo con i dati, i valori predetti y_f non risultano particolarmente differenti dai valori osservati y, quindi rispetto alla distribuzione $p(y_f \mid y)$ i valori osservati y sono plausibili (Gelman et al., 1995). In termini intuitivi, ammettendo che $p(y_f \mid y)$ sia gaussiana, i valori di y non appartengono alle code della distribuzione normale $p(y_f \mid y)$.

Una valutazione numerica approssimata di $p(y_f \mid y)$ può essere ottenuta estraendo a caso un campione di valori $(\theta_1, \theta_2, \ldots)$ dalla distribuzione $p(\theta \mid y)$, e campionando per ogni valore θ_i le osservazioni $(y_{\theta_i,1}, y_{\theta_i,2}, \ldots)$ da $p(y \mid \theta_i)$.

Se $p(y_f \mid y)$ assegna scarsa plausibilità ai dati osservati y allora è necessario un attento riesame del modello formulato, in ogni suo aspetto.

La formulazione della distribuzione iniziale $p(\theta)$ richiede l'espressione di un sistema di preferenze per i valori che il parametro può assumere (elicitazione). Tuttavia, l'assegnazione dei valori di probabilità (o di densità di probabilità) non può essere effettuato con precisione arbitraria, pertanto in alcuni casi l'elicitazione ha esito in una classe di distribuzioni iniziali, in luogo di una sola.

Nell'analisi di sensibilità di un modello (*sensitivity analysis,* O'Hagan, 1994, par. 4.33) si determinano i cambiamenti intervenuti nella distribuzione a posteriori per effetto del cambiamento della distribuzione a priori, della verosimiglianza o di entrambi. Un modello poco sensibile ai (plausibili)

cambiamenti di distribuzione iniziale, è detto robusto rispetto ad errori di formulazione della distribuzione a priori.

Nei metodi Bayesiani, la distribuzione finale $p(\theta \mid y)$ contiene tutta l'informazione che riguarda il parametro θ, pertanto si pone la necessità di comprenderne le principali caratteristiche (O'Hagan, 1994, capitolo 2).

Se il parametro θ è univariato oppure bivariato l'esame grafico della distribuzione $p(\theta \mid y)$ rivela le caratteristiche principali della distribuzione in maniera sintetica. L'esame grafico per parametri multivariati non è altrettanto informativo poiché deve essere effettuato su molteplici distribuzioni marginali bidimensionali derivate da $p(\theta \mid y)$, e nel processo di marginalizzazione parte dell'informazione contenuta in $p(\theta \mid y)$ è persa. In certi casi, l'esame grafico di particolari distribuzioni condizionali rispetto ad una o più componenti del vettore dei parametri θ costituisce uno strumento di comprensione in grado di fornire preziose informazioni di significato applicativo.

Le caratteristiche salienti della distribuzione finale possono essere ricavate da funzioni dei momenti (centrali e non) di $p(\theta \mid y)$ e da alcuni dei suoi percentili. La scelta del tipo di statistiche da impiegare per riassumere le caratteristiche principali della distribuzione a posteriori $p(\theta \mid y)$ dipende largamente dagli scopi dell'analisi.

2.1 I metodi Monte Carlo

I metodi Monte Carlo (MC) sono una famiglia di tecniche di simulazione stocastica, cioé sono basati sulla generazione di successioni di numeri pseudocasuali (Ripley, 1987). I metodi MC sono impiegati nella risoluzione numerica di espressioni integrali intrattabili (O'Hagan, 1994, par. 8.43 e seguenti), nell'analisi statistica non-Bayesiana, nella caratterizzazione numerica della funzione di verosimiglianza e nei calcoli richiesti in ambito Bayesiano (Smith, Gelfand, 1992, Tanner, 1993, Tierney, 1994, Besag et al., 1995).

Il campo di applicazione ed i principi teorici coinvolti sono assai vasti. Gilks, Richardson e Spiegelhalter (coordinatori dell'opera, 1996) forniscono un'introduzione di portata applicativa ad una specifica classe di metodi MC, le catene di Markov (*Markov Chain Monte Carlo*, MCMC). Essa include i recenti risultati ottenuti in questo campo ed un contributo introduttivo all'uso dei metodi MC in ambito genetico (Thomas, Gauderman, 1996).

Da un punto di vista intuitivo, è utile osservare che esiste una stretta relazione tra una generica funzione di densità di probabilità e l'istogramma di un campione di osservazioni tratte dalla medesima distribuzione. In altri termini, un campione sufficientemente grande fornisce un'informazione sulla distribuzione da cui sono estratte le osservazioni che è esaustiva a fini applicativi, con un livello di precisione nell'approssimazione che può essere stimato.

La dimensione del campione, in generale, dipende dal tipo di caratteristica della distribuzione che è oggetto d'interesse (ad esempio la media rispetto al percentile ε_p con $p = 0.001$), dal numero di dimensioni in cui è definita la distribuzione, dal grado di precisione stabilito.

In questo lavoro, l'obiettivo principale consiste nell'ottenimento della distribuzione a posteriori

$$p(\theta \mid y) = \frac{\pi(\theta) \cdot p(y \mid \theta)}{\int \pi(\theta) \cdot p(y \mid \theta) \cdot d\theta}$$

con $p(y \mid \theta)$ la verosimiglianza e $\pi(\theta)$ la distribuzione a priori del parametro θ.

Benché la distribuzione a posteriori non normalizzata sia immediatamente disponibile in seguito alla specificazione di un modello gerarchico, il compito di riassumere le caratteristiche di interesse della distribuzione finale $p(\theta \mid y)$ non è banale. L'integrale richiesto a denominatore può non essere calcolabile in forma chiusa, oppure essere arduo da valutare anche per via numerica.

Anche nel caso in cui tale costante (il denominatore) sia stata numericamente valutata, gli integrali richiesti per calcolare i principali riassunti numerici della distribuzione finale $p(\theta \mid y)$ (quali la media, la varianza od un certo percentile) possono rivelarsi analiticamente intrattabili.

O'Hagan (1994, par. 8.1) sottolinea che un'adeguata pratica del metodo Bayesiano richiede di specificare la distribuzione a priori in maniera che rifletta accuratamente l'informazione disponibile circa i parametri del modello, quale che sia la complessità che ne deriva. In altri termini, l'applicazione del paradigma Bayesiano non può essere ristretta ai modelli trattabili analiticamente (risolvibili in forma chiusa).

I metodi Monte Carlo impiegati in ambito Bayesiano costituiscono una classe di procedure utilizzate per ottenere un campione $(\theta^1, \ldots, \theta^n)$ di valori del parametro θ da impiegare quale approssimazione della distribuzione $p(\theta \mid y)$. Alcune di queste procedure, come l'algoritmo di Metropolis, richiedono la specificazione della sola distribuzione finale non normalizzata.

L'implementazione al calcolatore degli algoritmi MC è indispensabile e il livello di complessità trattabile nel modello dipende in gran parte dalla potenza computazionale disponibile.

Nei paragrafi successivi, l'attenzione è rivolta esclusivamente ai metodi impiegati in questo lavoro ed alle relative problematiche di impiego.

2.2 L'algoritmo di Metropolis

L'algoritmo di Metropolis (Metropolis et al., 1953, O'Hagan, 1994, par. 8.72, Gelman et al., 1995, par. 11.2, Chib e Greenberg, 1995) è un metodo per realizzare una passeggiata casuale (*random walk*) nello spazio dei parametri, a cui corrisponde la successione di valori del parametro $(\theta^0, \theta^1, \ldots, \theta^t, \ldots, \theta^n)$ che converge alla distribuzione finale $p(\theta \mid y)$ (*target distribution*) per $n \to \infty$.

Il campione di valori $(\theta^0, \theta^1, \ldots, \theta^t, \ldots, \theta^n)$ è ottenuto in maniera sequenziale, a partire dal valore θ^0 estratto da una distribuzione $p_0(\theta)$ (*starting)*, attraverso la transizione di stato probabilistica specificata dalla distribuzione *jumping* $J(\widetilde{\theta} \mid \theta^{t-1})$ la quale dipende dal valore θ^{t-1} generato all'ultima transizione.

La distribuzione *jumping* deve essere scelta in maniera tale che la catena di Markov risultante converga ad una sola distribuzione stazionaria, la distribuzione a posteriori $p(\theta \mid y)$.

L'algoritmo è definito nei passi seguenti (Gelman et al. 1995, par. 11.2, modificato):

1) estrazione di un valore θ^0 dalla distribuzione *starting* $p_0(\theta)$, in maniera tale che $p(\theta^0 \mid y) > 0$;

2) estrazione di $\widetilde{\theta}$ come punto candidato a divenire θ^t al passo t nella successione $(\theta^0, \theta^1, \ldots, \theta^{t-1})$; il punto candidato è ottenuto ricorrendo alla distribuzione *jumping* $J(\widetilde{\theta} \mid \theta^{t-1})$; la distribuzione *jumping* deve essere simmetrica, cioè $J(\theta_a \mid \theta_b) = J(\theta_b \mid \theta_a)$ per ogni θ_a, θ_b, t;

3) calcolo del rapporto r

$$r = \frac{p(\widetilde{\theta} \mid y).}{p(\theta^{t-1} \mid y).}$$

4) definizione di θ^t :

se $r \geq 1$ allora $\theta^t = \widetilde{\theta}$

altrimenti si estrae un valore ausiliario uniformemente distribuito $\alpha \sim uniforme(0, 1)$

se $\alpha < r$ allora $\theta^t = \widetilde{\theta}$

altrimenti $\theta^t = \theta^{t-1}$

5) ripetizione dei punti (2)-(4) fino ad ottenere una successione di n termini;

6) ripetizione dei punti (1)-(5) per un numero J di volte.

Il punto (6) consiste nella ripetizione della simulazione un numero J di volte, come richiesto nella fase di studio della convergenza.

Dato il punto θ^{t-1}, la distribuzione di transizione di stato della catena di Markov $T(\theta^t \mid \theta^{t-1})$ è una mistura della distribuzione $J(\theta^t \mid \theta^{t-1})$ e della massa situata nel punto $\theta^t = \theta^{t-1}$. Per costruzione, la catena di Markov descritta è irriducibile, aperiodica e non transiente, pertanto essa ha un'unica distribuzione stazionaria, che si può mostrare è proprio $p(\theta \mid y)$.

Una comune scelta delle distribuzioni richieste dal metodo Metropolis segue le indicazioni contenute in Gelman et al. (1995, pag. 334). Questi autori hanno proposto dei valori per i parametri della simulazione Metropolis che dovrebbero essere ottimali nel caso in cui la distribuzione finale e la distribuzione *jumping* siano normali multivariate; in particolare:

1) $p_0(\theta)$ è una distribuzione normale multivariata con vettore delle medie pari ai valori della distribuzione finale $p(\theta \mid y)$ nel punto modale e con matrice di covarianza pari a quella di $p(\theta \mid y)$, stimata nel punto modale;

2) $J(\widetilde{\theta} \mid \theta^{t-1})$ è una distribuzione normale multivariata con vettore delle medie θ^{t-1} e matrice di covarianza $c^2\Sigma$, con Σ uguale alla matrice di covarianza della distribuzione finale $p(\theta \mid y)$ stimata nel punto modale e $c = 2.4/\sqrt{d}$, con d il numero di parametri nel modello.

Il grado di convergenza può essere valutato con il metodo di Gelman e Rubin (1992a, 1992b).

2.3 Modelli semiparametrici Bayesiani

In questa sezione sono brevemente descritti i modelli semiparametrici impiegati nell'analisi del 2008 e 2009.

I dati del 2008 sono stati analizzati con un modello a basi troncate di secondo grado dopo trasformazione del tempo in modo che abbia valore compreso tra 0 e 1. La trasformazione dell'asse temporale è stata necessaria per poter confrontare trattamenti con diversa durata della fase di macerazione: infatti il trattamento MPF (18 giorni nel 2008, 15 giorni nel 2009) dura sensibilmente più a lungo degli altri due (14 giorni nel 2008, 12 giorni nel 2009). Con questa trasformazione il tempo viene espresso come una percentuale della macerazione ed ha un valore compreso tra 0 e 100%.

Sia $\ell = 1, \ldots, L$, l'indice di vasca dentro trattamento, allora la risposta all'i-esima osservazione è scomposta come segue:

$$y_{i,l} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \sum_{k=1}^{K} b_k (x_i - \kappa_k)_+^2 + \sum_{\ell=2}^{L} z_{i,\ell}(\gamma_{0,\ell} + \gamma_{1,\ell}\, x_i + \gamma_{2,\ell}\, x_i^2) + \sum_{\ell=1}^{L} z_{i,\ell} \left\{ \sum_{k=1}^{K} c_k^{\ell}\, (x_i - \kappa_k)_+^2 \right\} + \varepsilon_i \quad (1)$$

in cui $b_k \sim N(0, \sigma_b^2)$, $c_k^{\ell} \sim N(0, \sigma_{c,\ell}^2)$, con varianza σ_b^2 e con varianze degli effetti casuali $\sigma_{c,\ell}^2$.

Brumback, Ruppert and Wand (1999) mostrano che tale equazione può essere scritta in forma matriciale

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta}\, + \mathbf{Z}\mathbf{u} + \,\boldsymbol{\varepsilon} \quad (2)$$

dove

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim N\left(\mathbf{0}\,, \begin{matrix} \boldsymbol{G} & \mathbf{0} \\ \mathbf{0} & \sigma^2_\varepsilon\, \boldsymbol{I} \end{matrix} \right)$$

e la matrice $\boldsymbol{G} = diag(\sigma^2_b 1_K, \sigma^2_{c,\ell} 1_K, \ldots, \sigma^2_{c,L} 1_K)$. Il vettore unario è 1_K cioè un vettore di uno di dimensione $K \times 1$.

I dati del 2009 sono stati analizzati con un modello splines a basi radiali dopo trasformazione del tempo in modo che abbia valore compreso tra 0 e 1. La trasformazione dell'asse temporale è stata necessaria per poter confrontare trattamenti con diversa durata della fase di macerazione. Con questa trasformazione il tempo viene espresso come una percentuale della macerazione ed ha un valore compreso tra 0 e 100%.

Il modello ottenuto via ottimizzazione mostra una miglior flessibilità in accordo a quanto richiesto dalle caratteristiche dei dati del 2009. Formalmente il modello lineare di una vasca è definito da:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \sum_{k=1}^{K} B_k(x_i, \kappa_k) + \varepsilon_i \tag{3}$$

in cui la base generalizzata dipende da un iperparametro Σ, ovvero

$$B_k(x_i, \kappa_k) = \frac{1}{\sqrt(2\pi\Sigma)} \exp\left\{ -0.5 \frac{(x_i - \kappa_k)^2}{\Sigma} \right\}$$

e dove $\{\kappa_k\}$ è la successione dei nodi. La varianza è specifica per il trattamento considerato.

La costruzione di opportuni contrasti consente di ottenere la media di trattamento, la quale in genere dipende da tre vasche ma che in due casi è calcolata in 2 (ovvero 4) vasche.

La distribuzione predittiva della media di ogni trattamento è stata diagrammata come cinetica attesa, ed in seguito sono stati calcolati i contrasti funzionali tra trattamento di interesse e trattamento di riferimento (vedere grafici in sezioni successive).

La convergenza delle simulazioni MC sono state valutate con gli usuali metodi diagnostici.

3 Grafici del 2008

In questa sezione sono riportati i grafici relativi ai risultati del 2008. In primo luogo sono riportate le cinetiche, quindi i contrasti funzionali.

I grafici dei contrasti vanno letti come segue: la linea superiore del titolo di ogni grafico è il trattamento sotto esame, mentre la riga sotto è il trattamento di riferimento. L'area grigia indica i tempi in cui i due trattamenti sono diversi con un intervallo di credibilità del 95 %.

3.1 Cinetiche del 2008

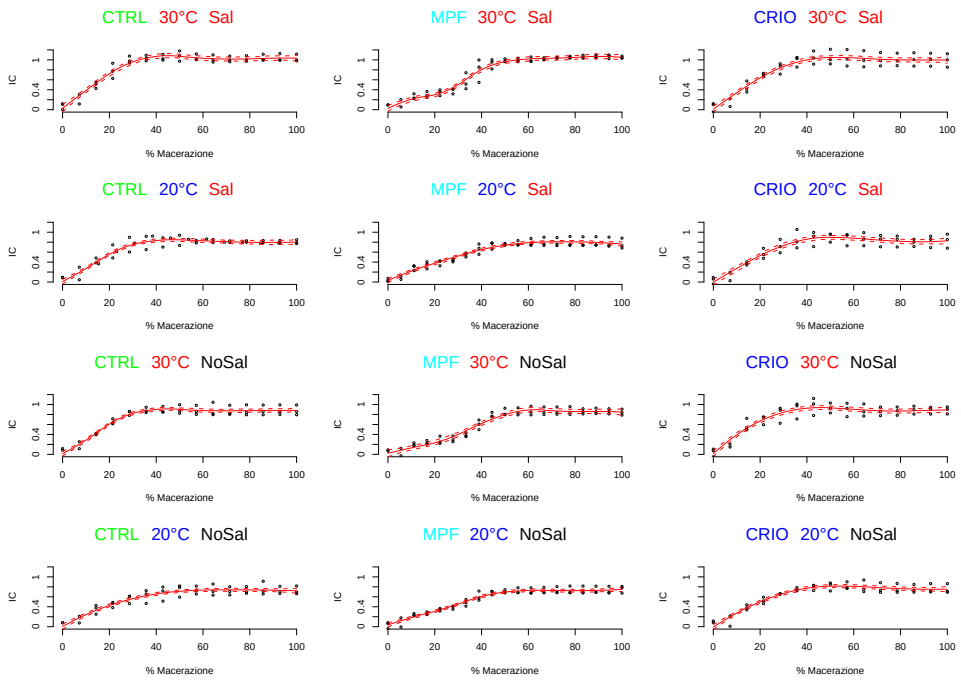

Figura 1: Cinetica di IC (2008)

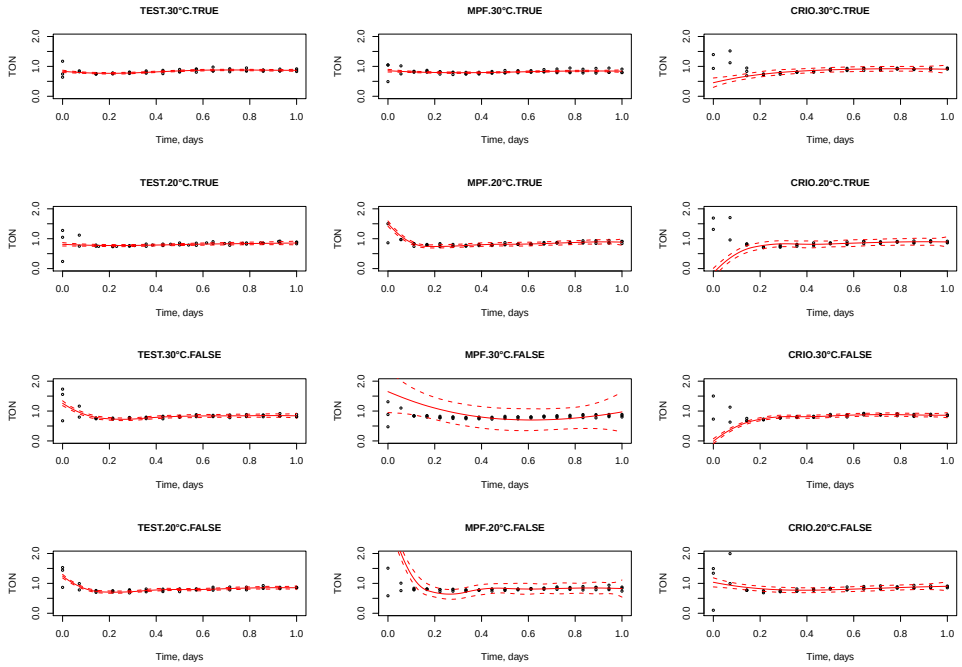

Figura 2: Cinetica di TON (2008)

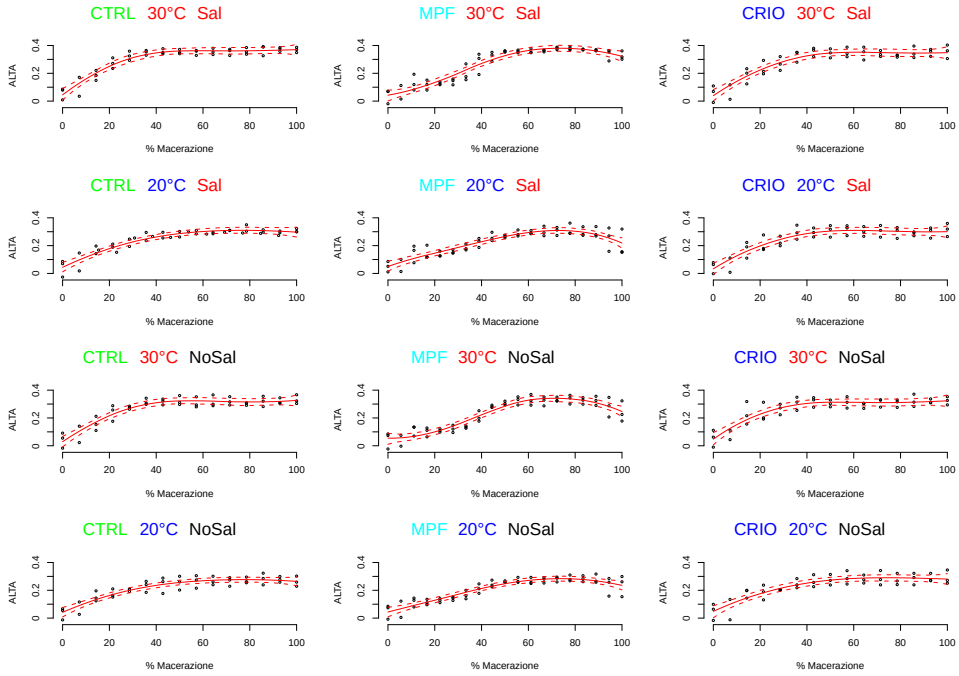

Figura 3: Cinetica di ALTA (2008)

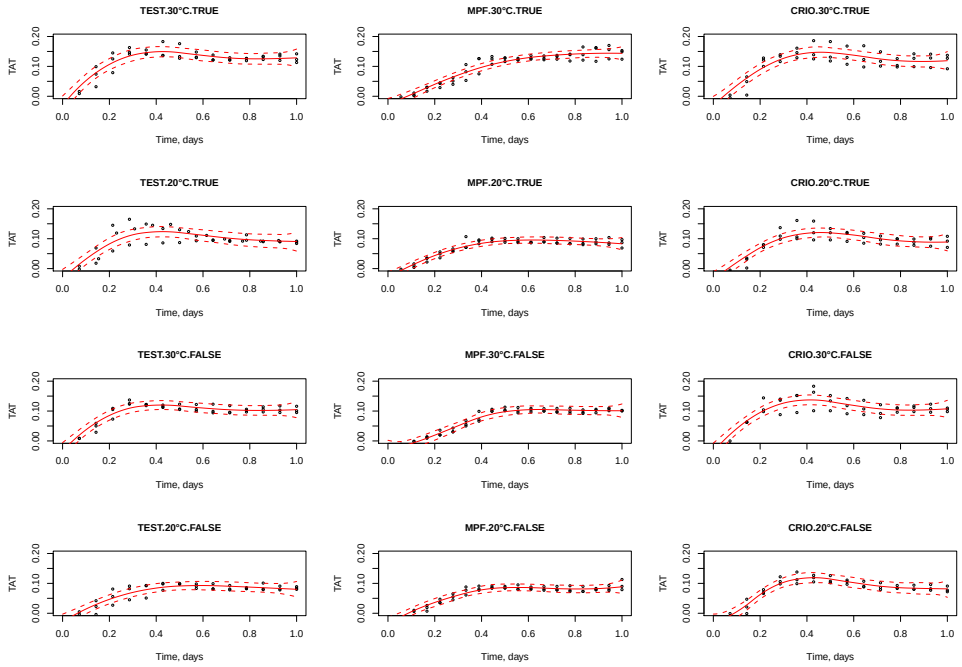

Figura 4: Cinetica di TAT (2008)

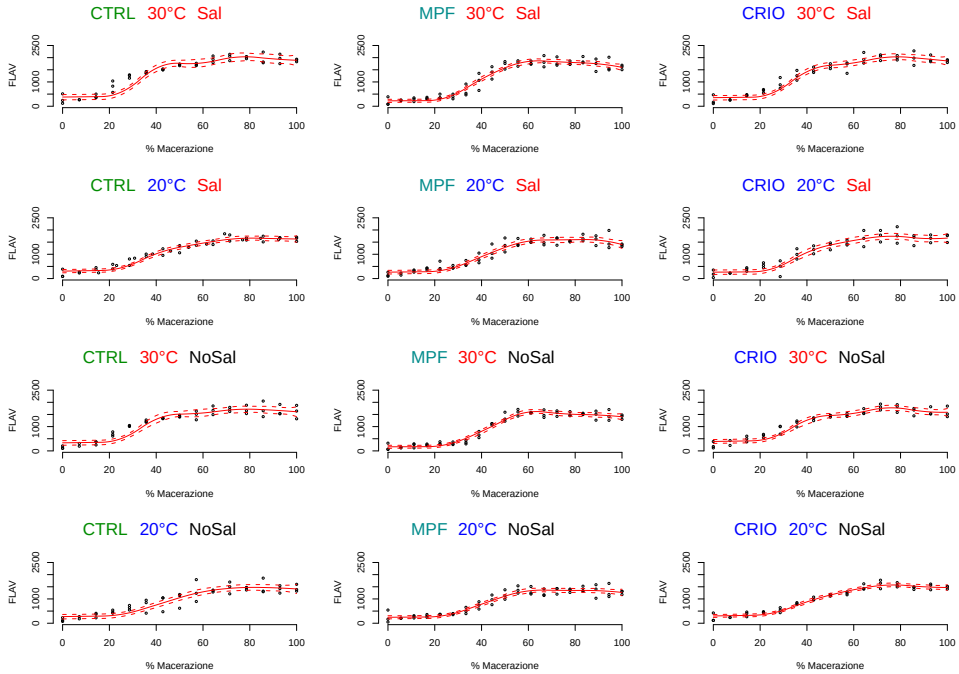

Figura 5: Cinetica di FLAV (2008)

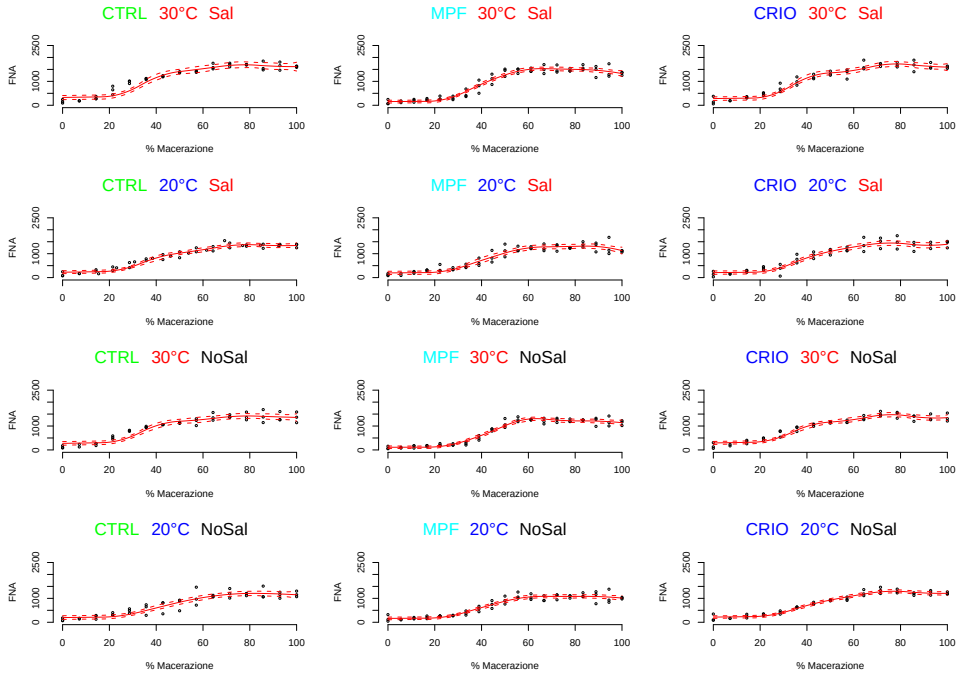

Figura 6: Cinetica di FNA (2008)

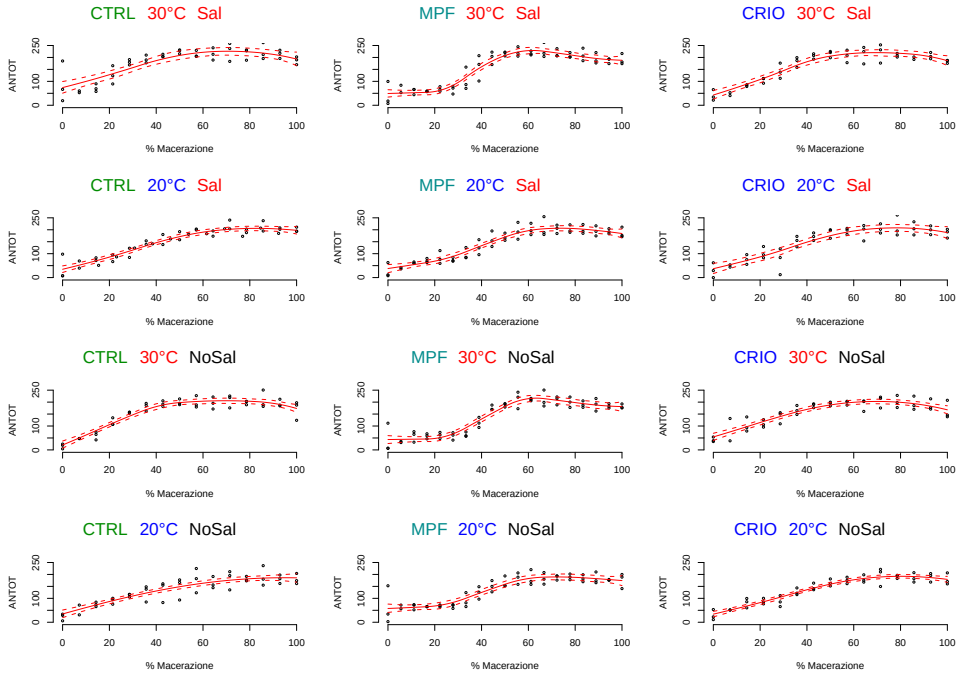

Figura 7: Cinetica di ANTOT (2008)

3.2 Contrasti del 2008

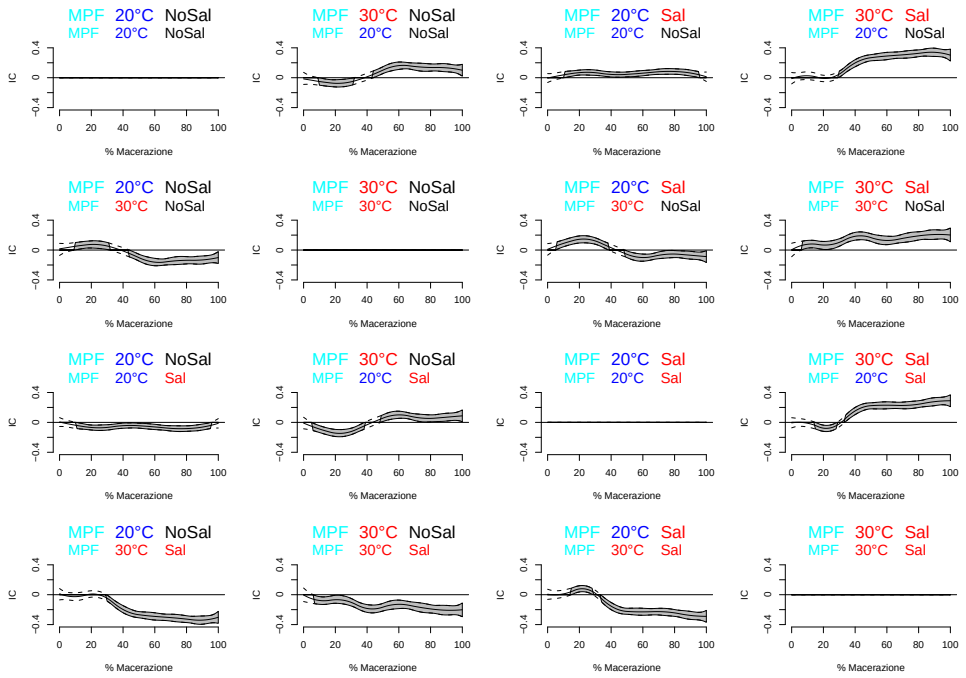

Figura 8: Contrasti per IC (2008)

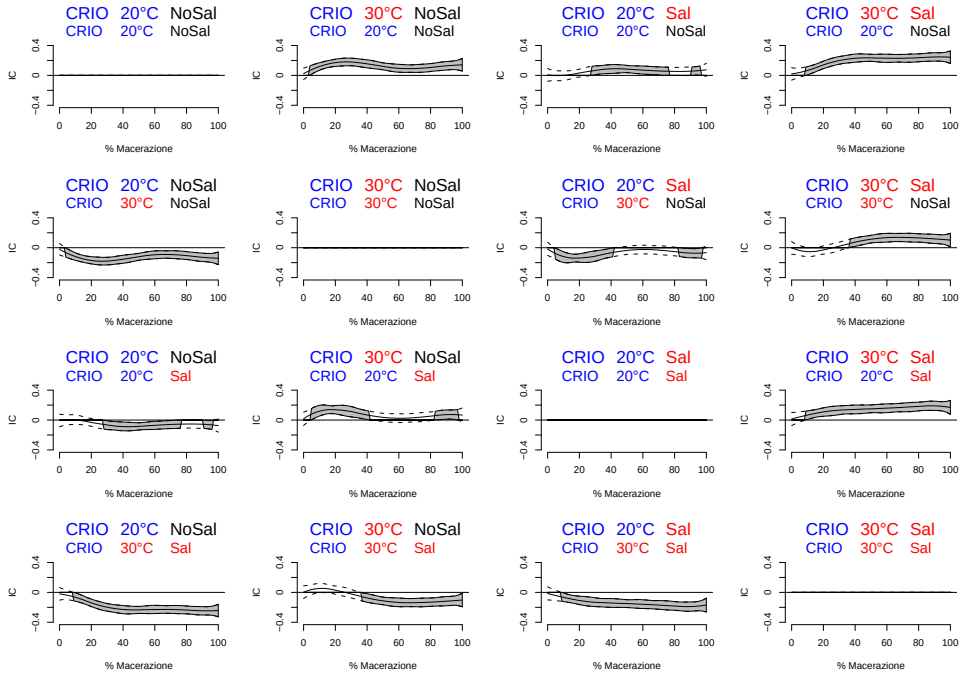

Figura 9: Contrasti per IC (2008)

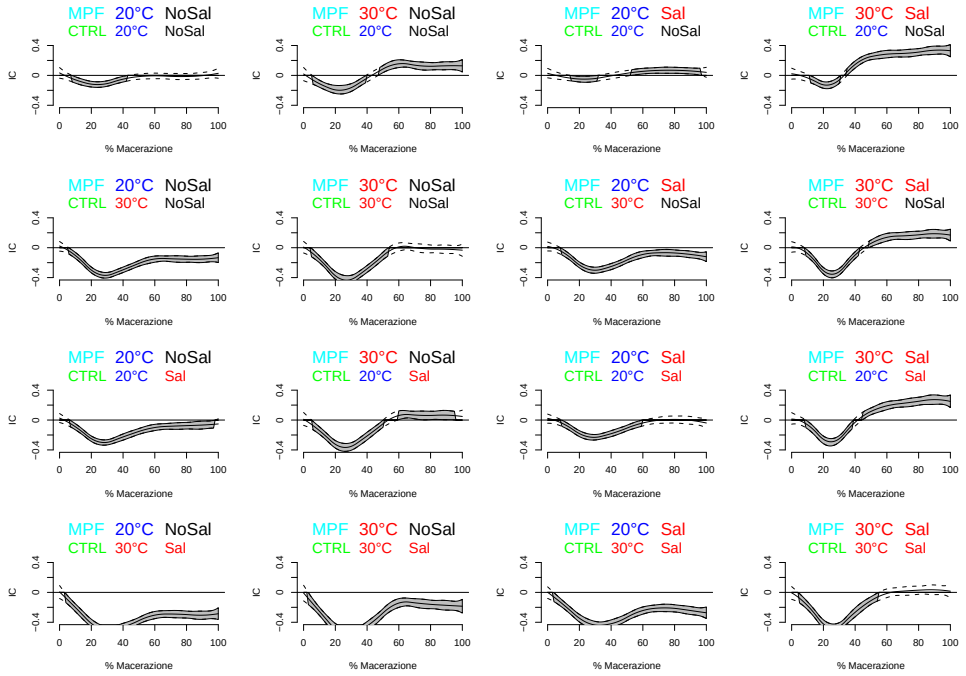

Figura 10: Contrasti per IC (2008)

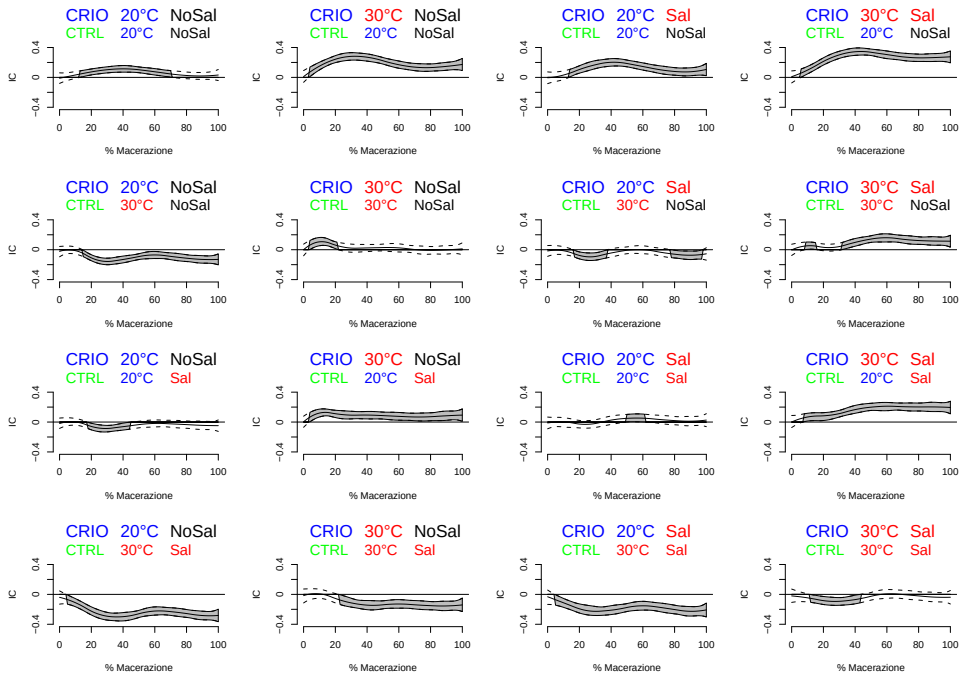

Figura 11: Contrasti per IC (2008)

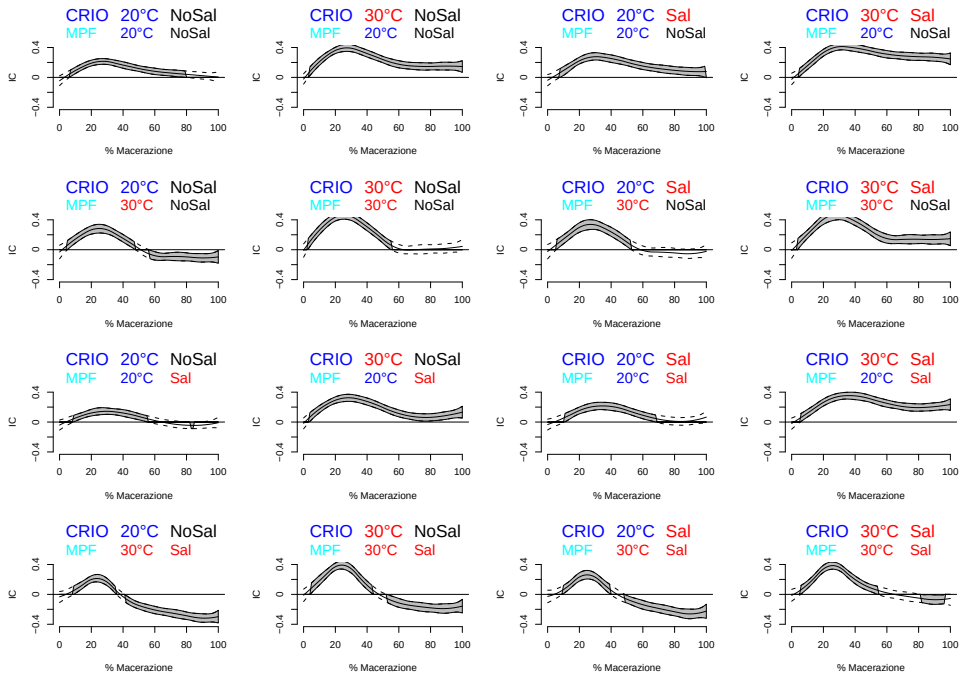

Figura 12: Contrasti per IC (2008)

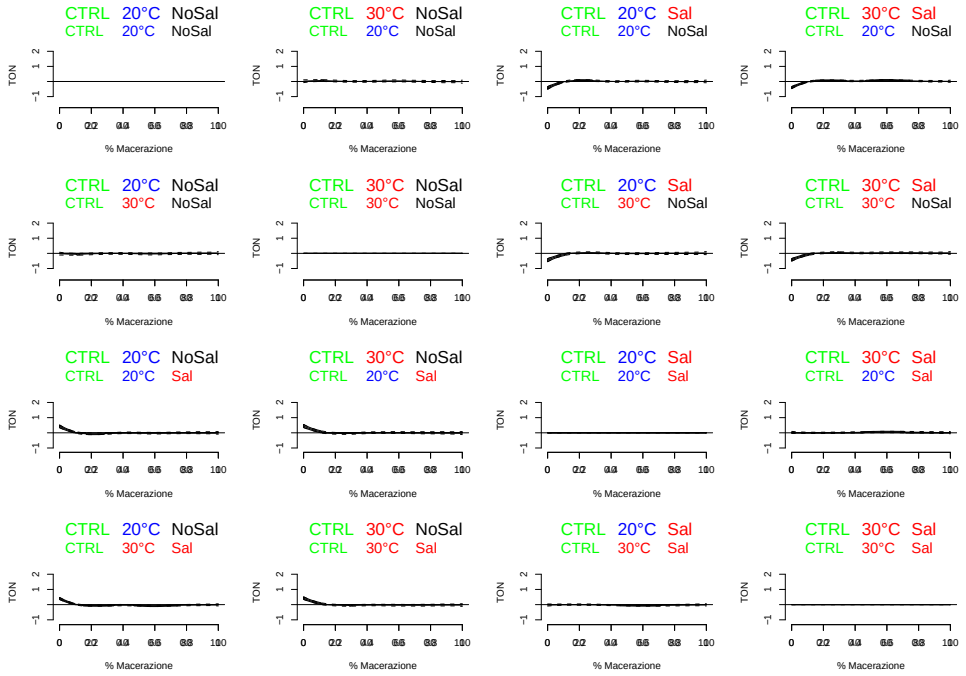

Figura 13: Contrasti per TON (2008)

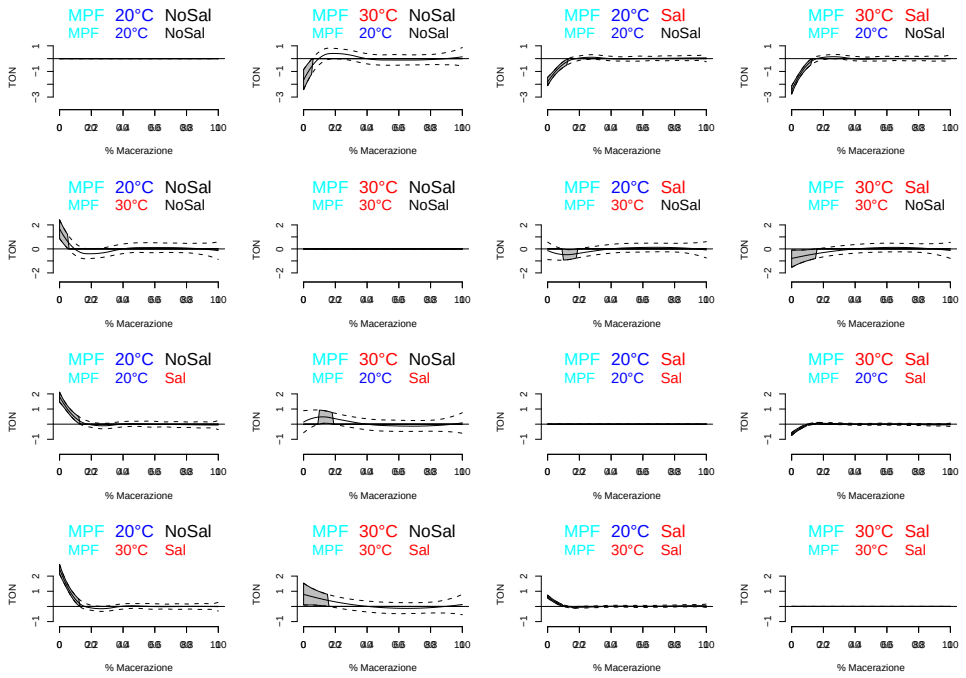

Figura 14: Contrasti per TON (2008)

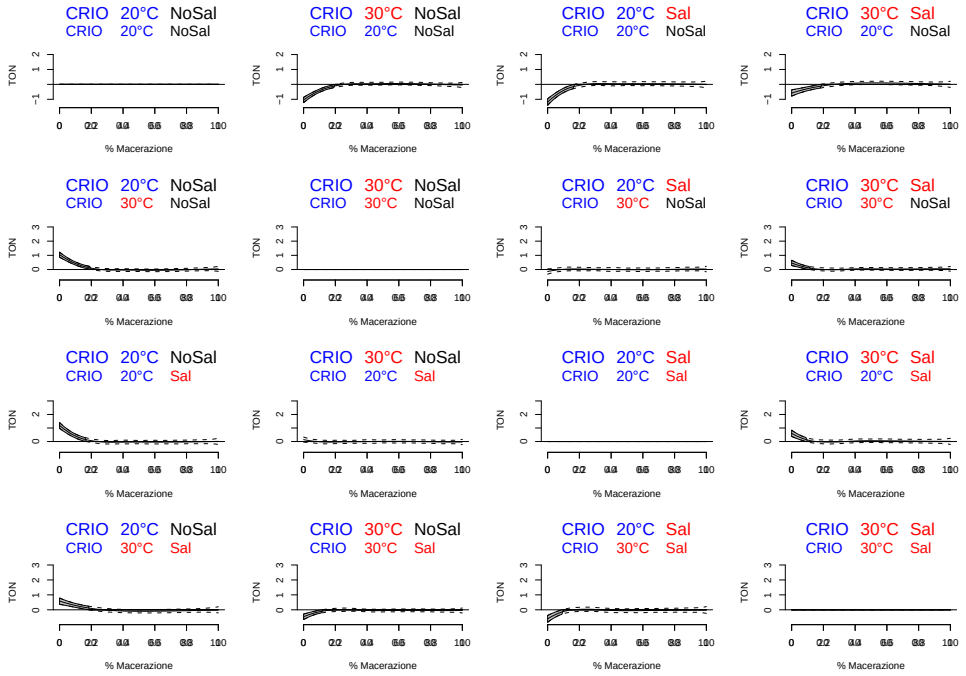

Figura 15: Contrasti per TON (2008)

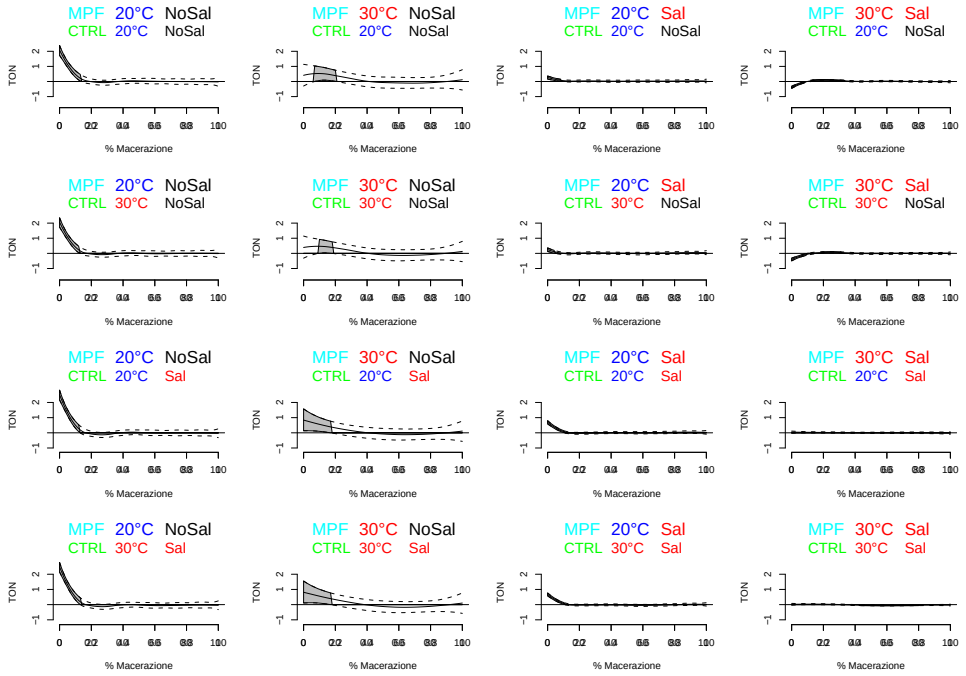

Figura 16: Contrasti per TON (2008)

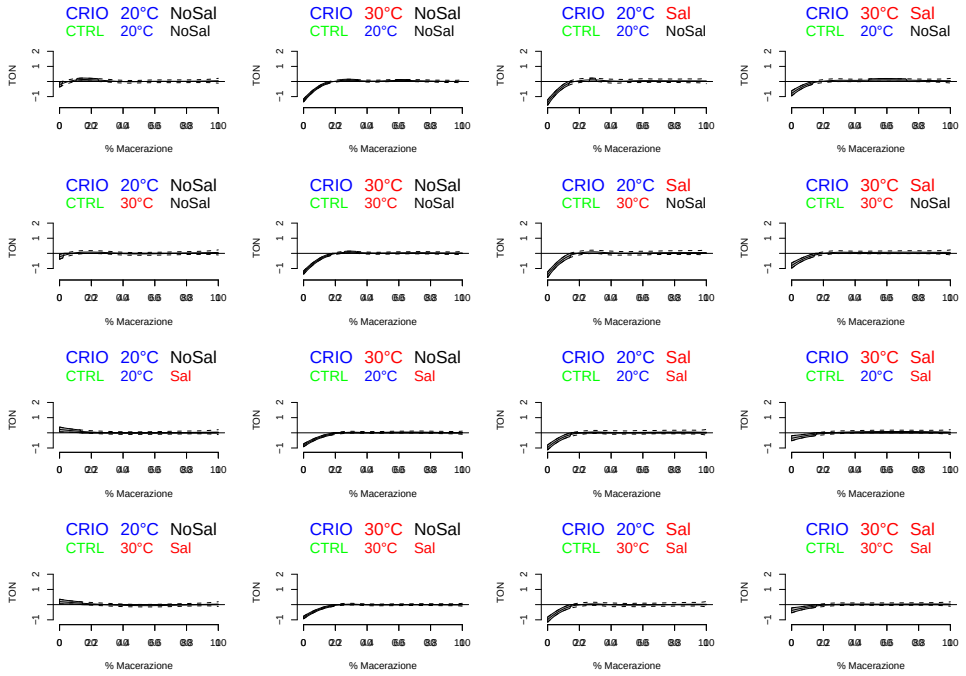

Figura 17: Contrasti per TON (2008)

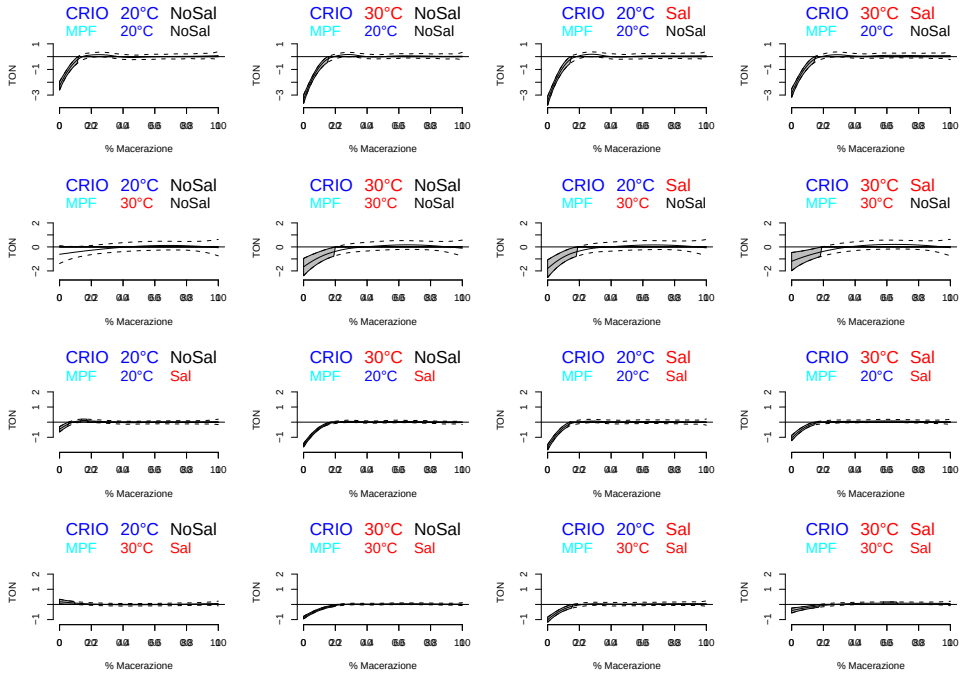

Figura 18: Contrasti per TON (2008)

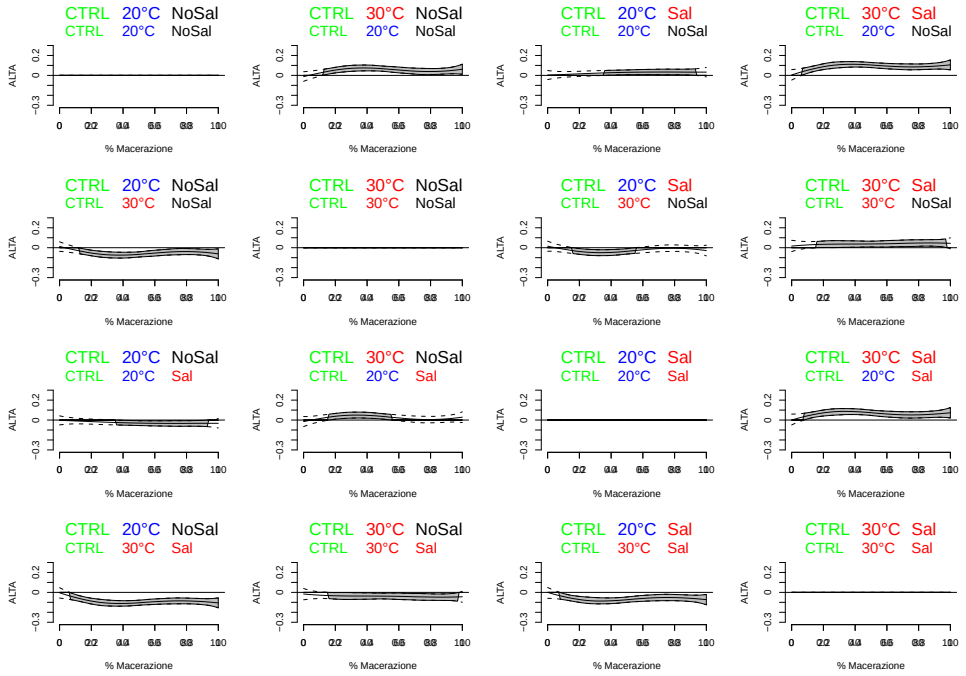

Figura 19: Contrasti per ALTA (2008)

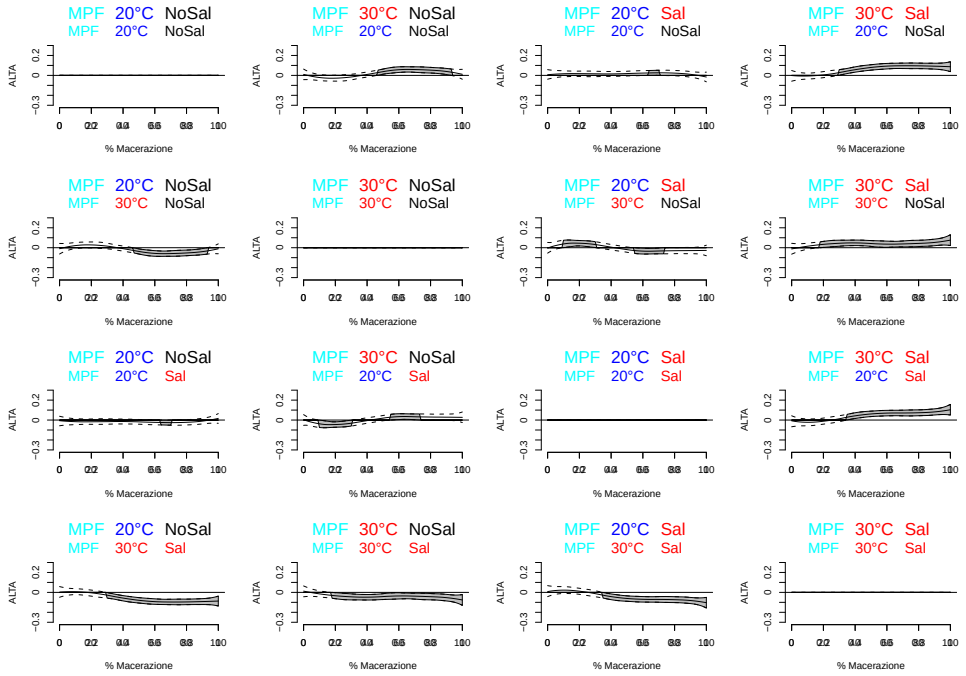

Figura 20: Contrasti per ALTA (2008)

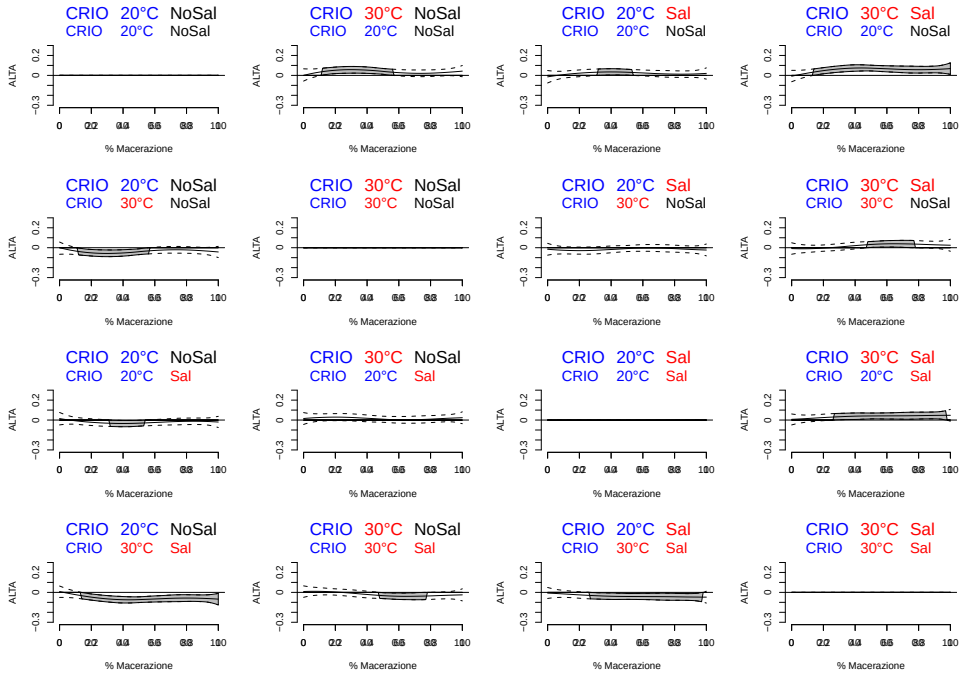

Figura 21: Contrasti per ALTA (2008)

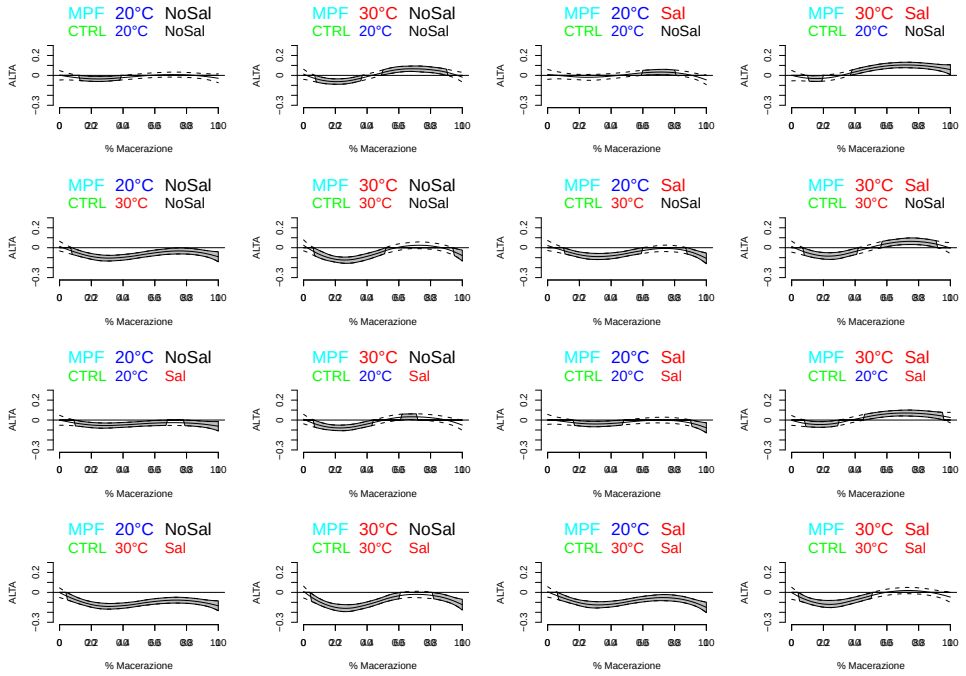

Figura 22: Contrasti per ALTA (2008)

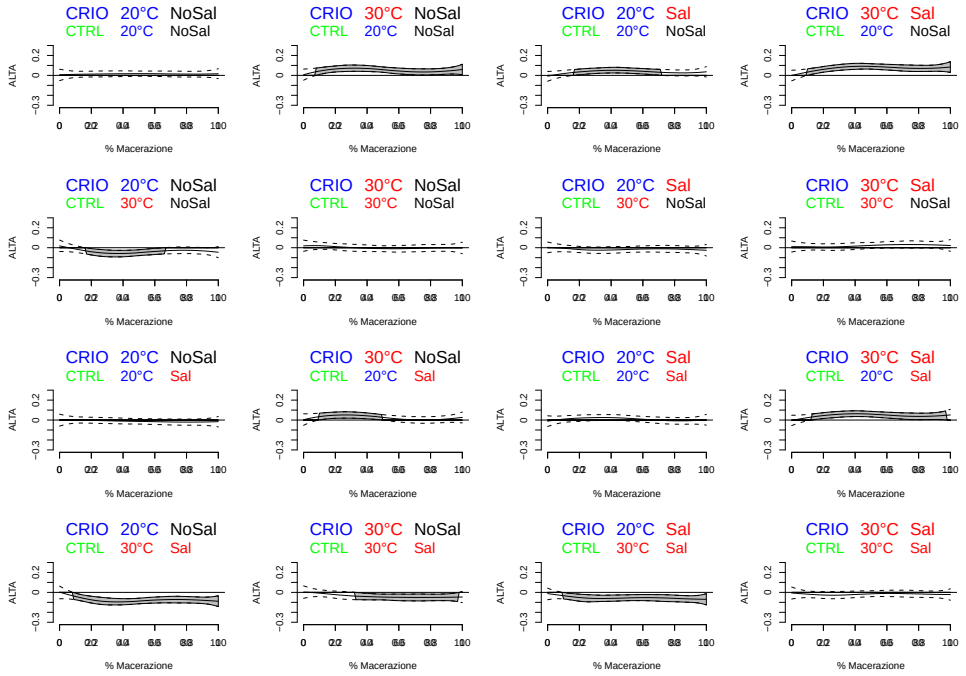

Figura 23: Contrasti per ALTA (2008)

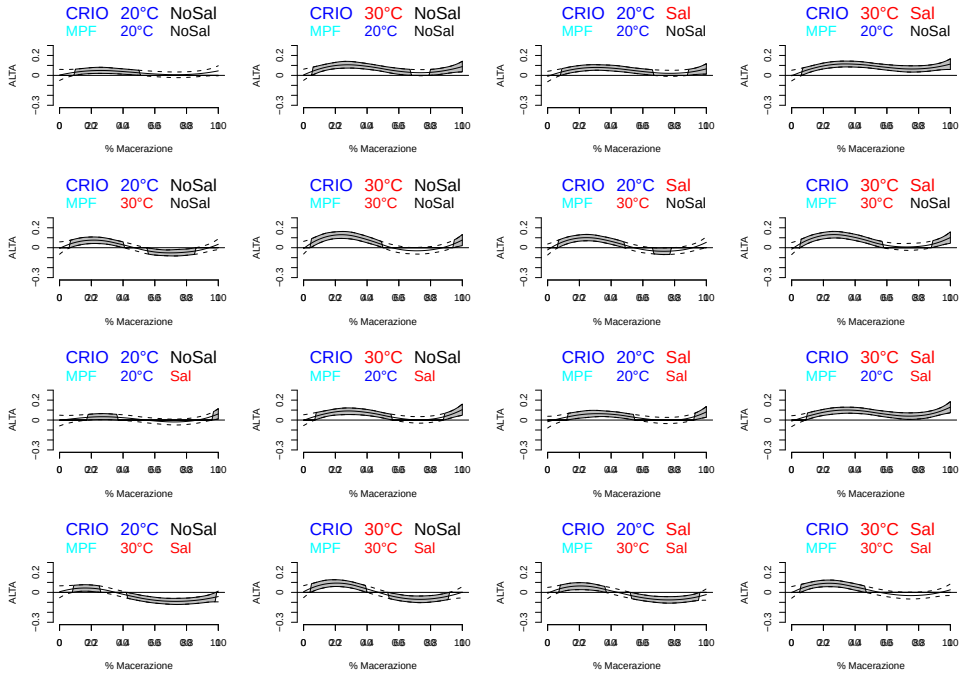

Figura 24: Contrasti per ALTA (2008)

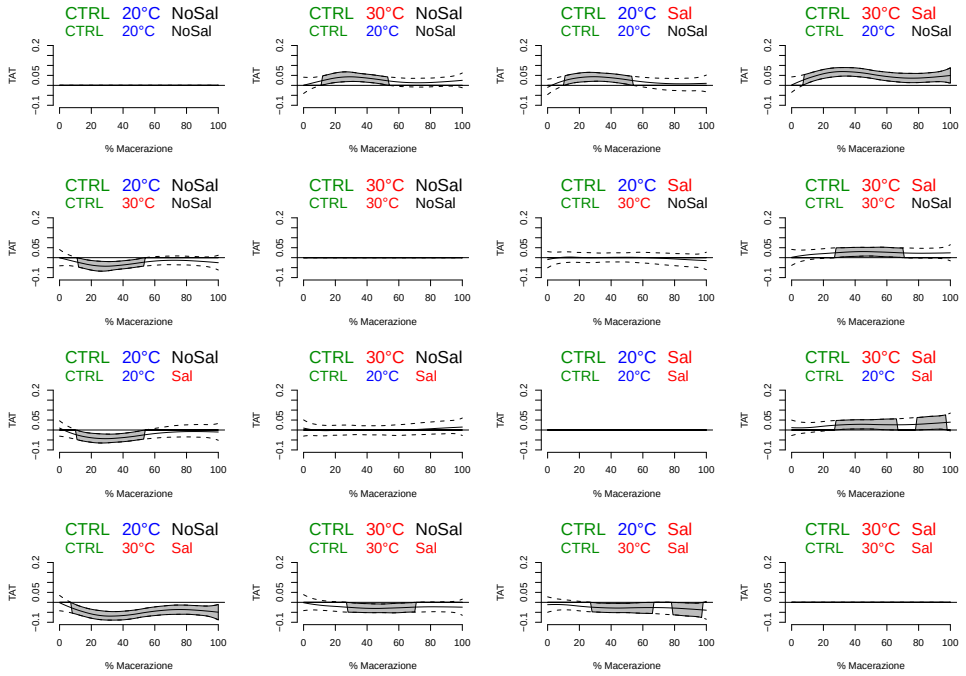

Figura 25: Contrasti per TAT (2008)

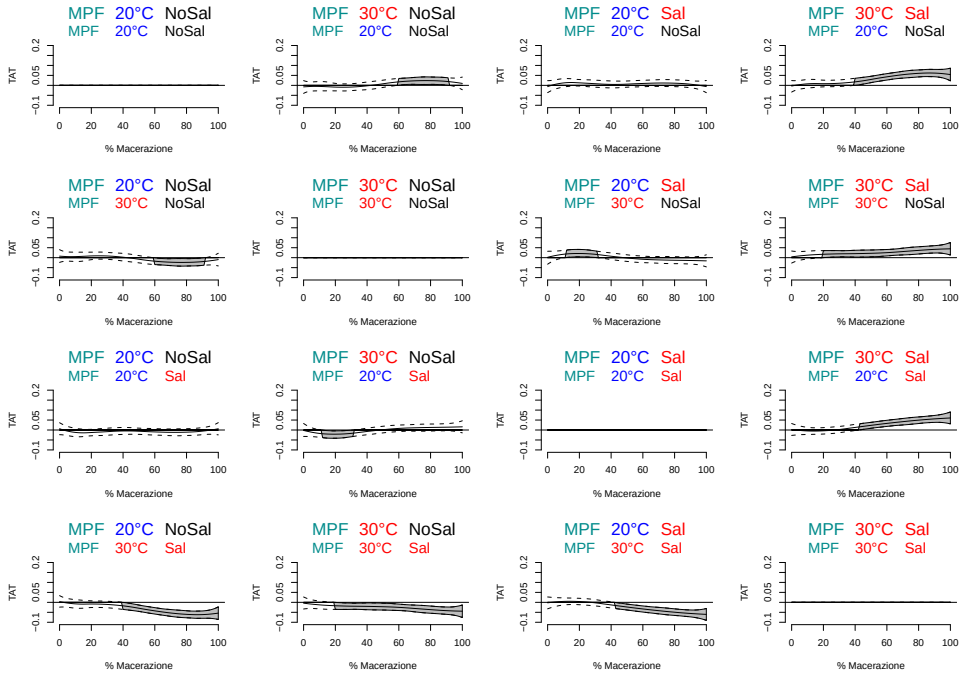

Figura 26: Contrasti per TAT (2008)

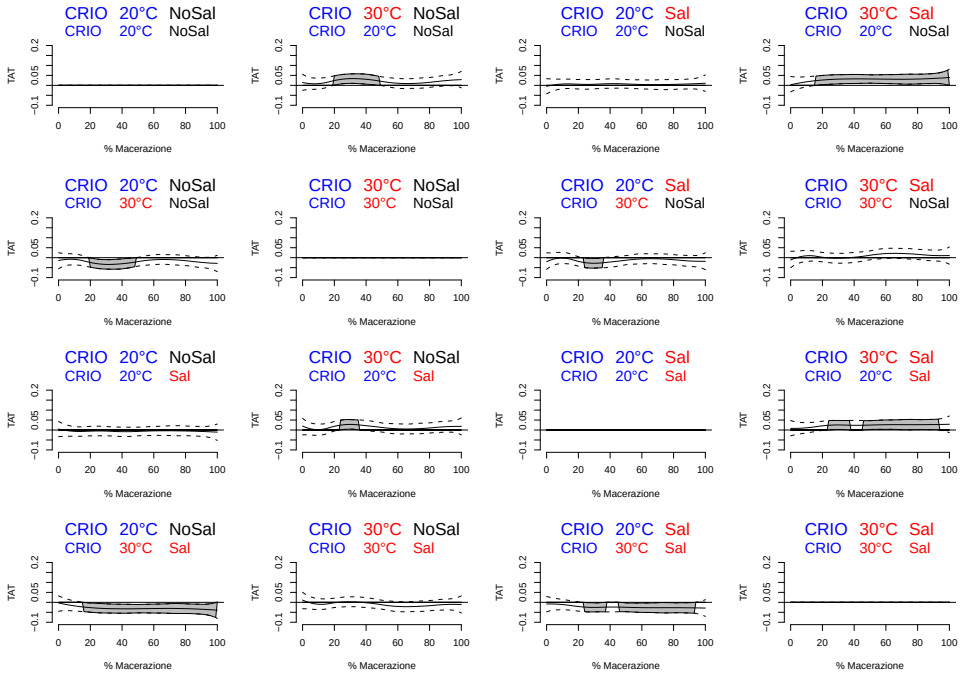

Figura 27: Contrasti per TAT (2008)

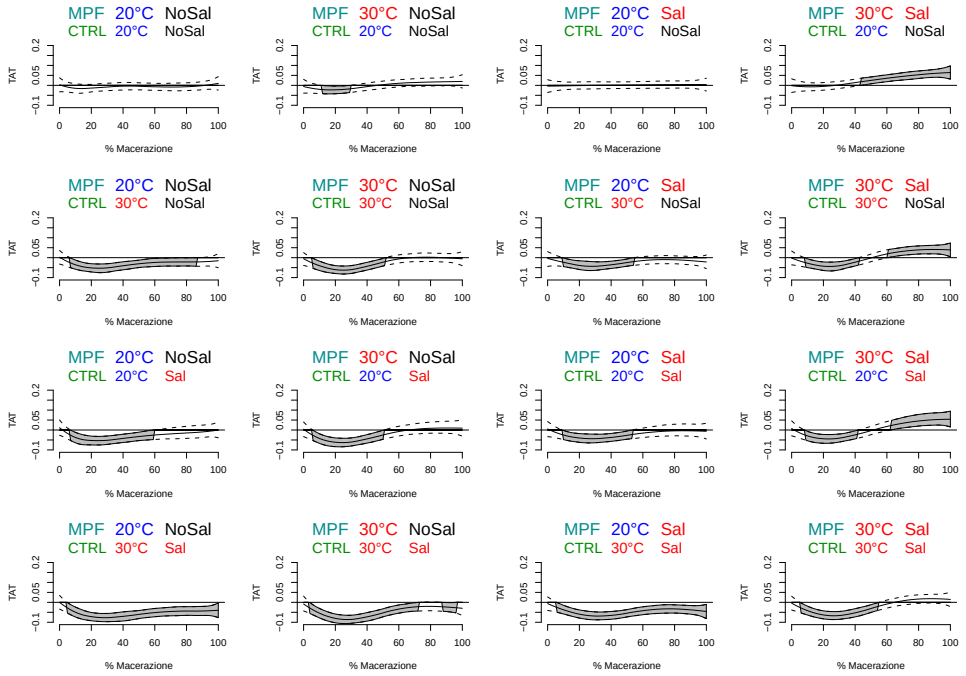

Figura 28: Contrasti per TAT (2008)

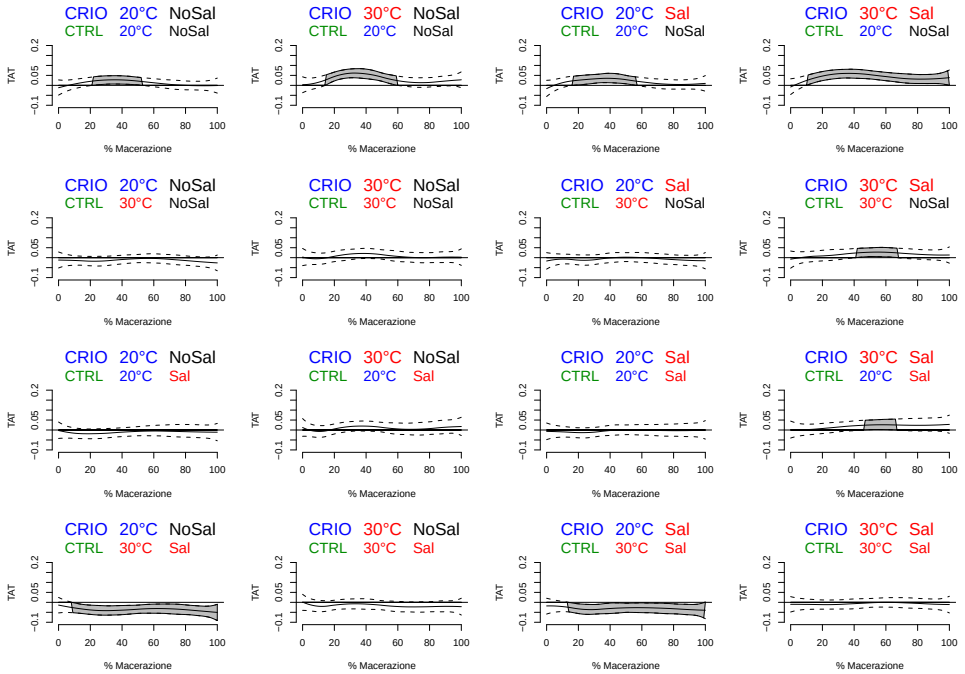

Figura 29: Contrasti per TAT (2008)

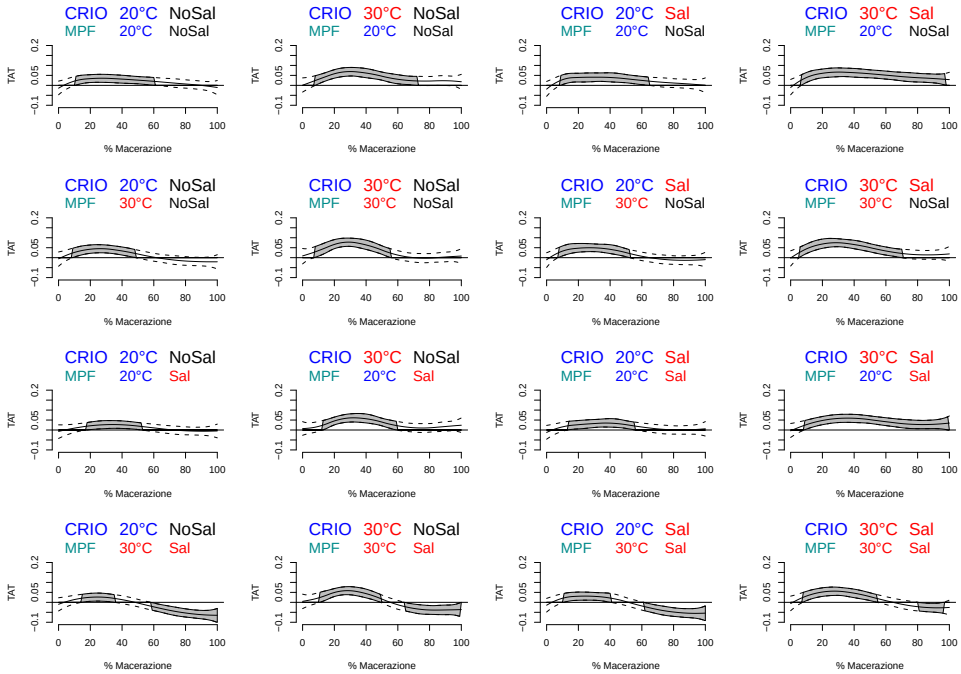

Figura 30: Contrasti per TAT (2008)

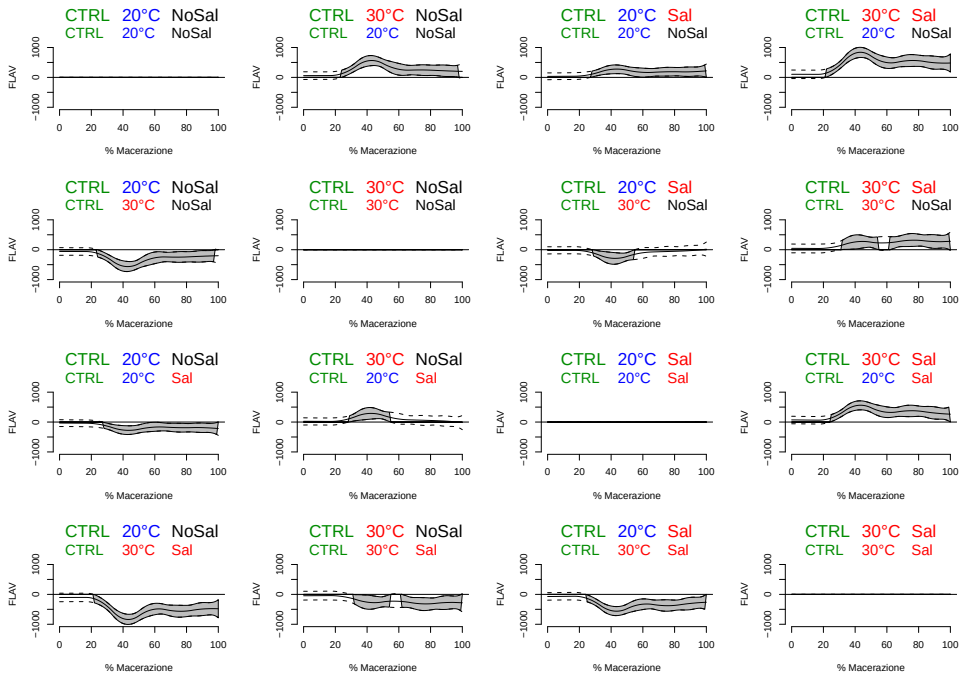

Figura 31: Contrasti per FLAV (2008)

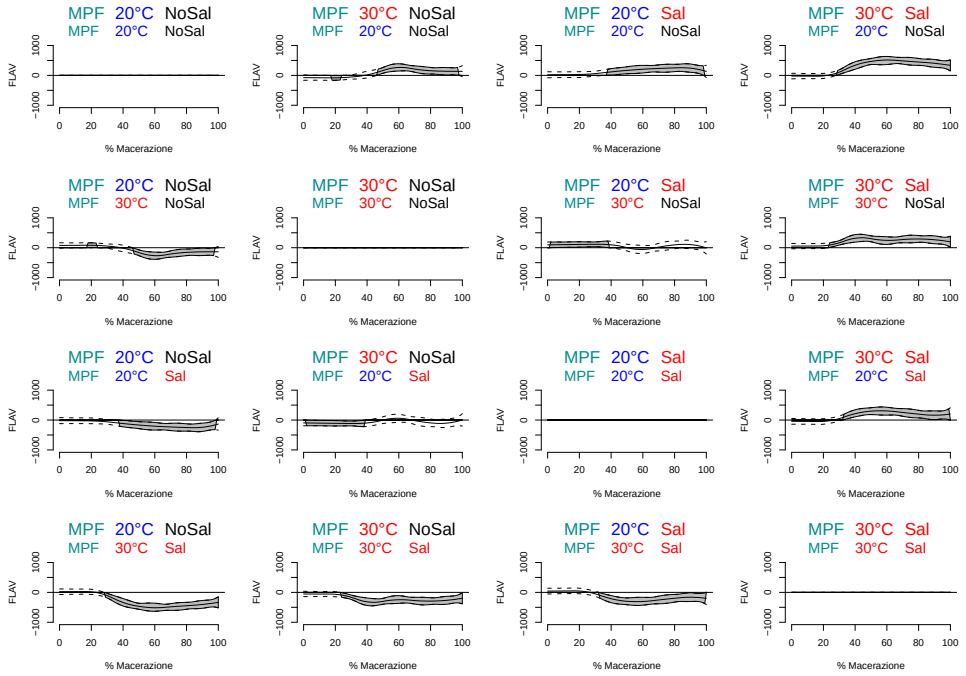

Figura 32: Contrasti per FLAV (2008)

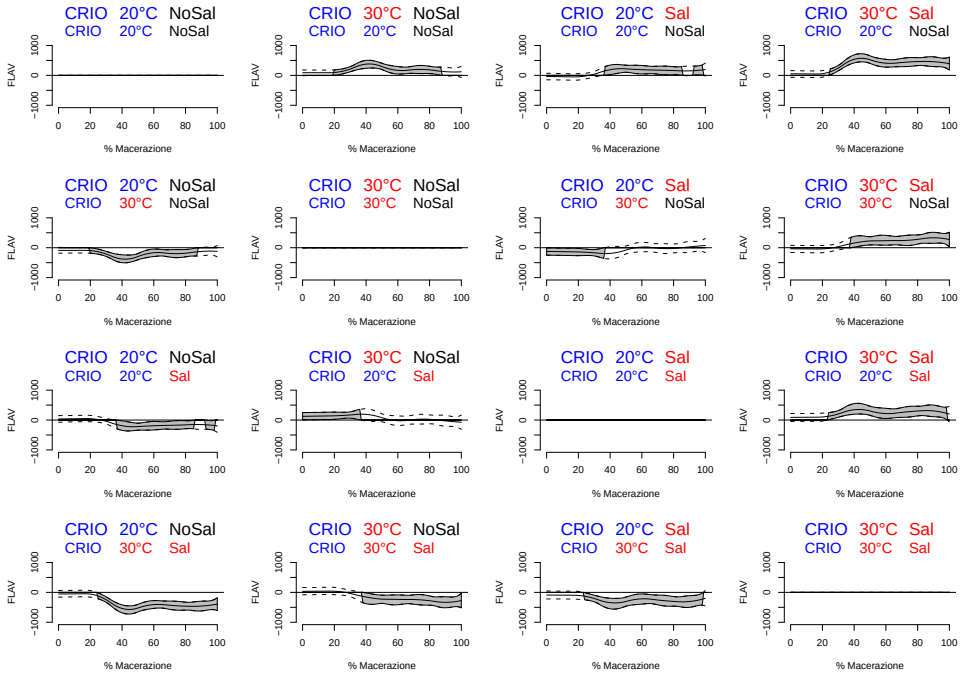

Figura 33: Contrasti per FLAV (2008)

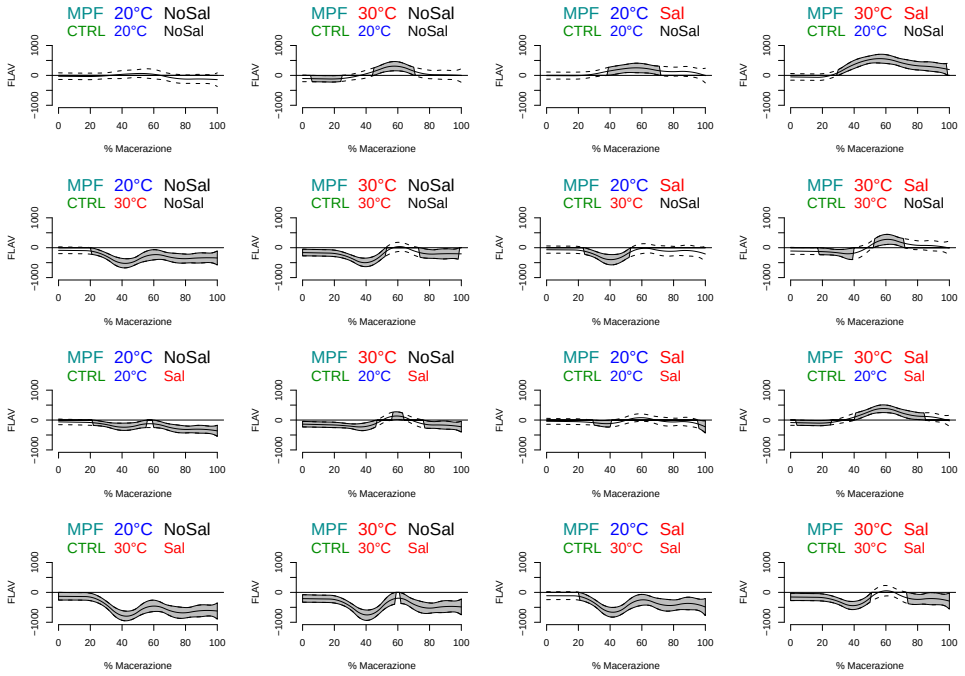

Figura 34: Contrasti per FLAV (2008)

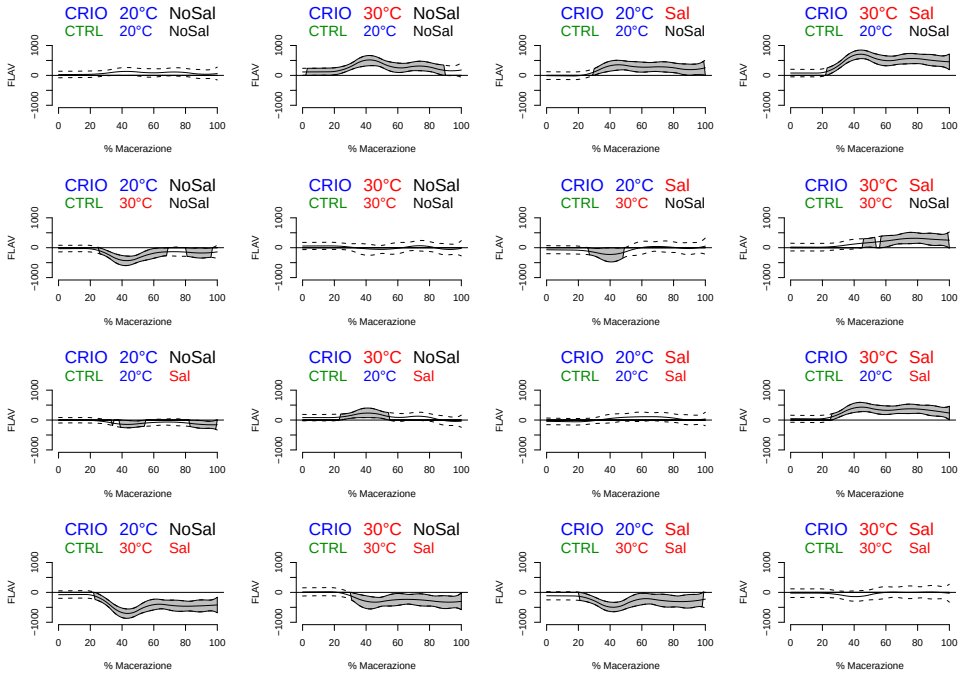

Figura 35: Contrasti per FLAV (2008)

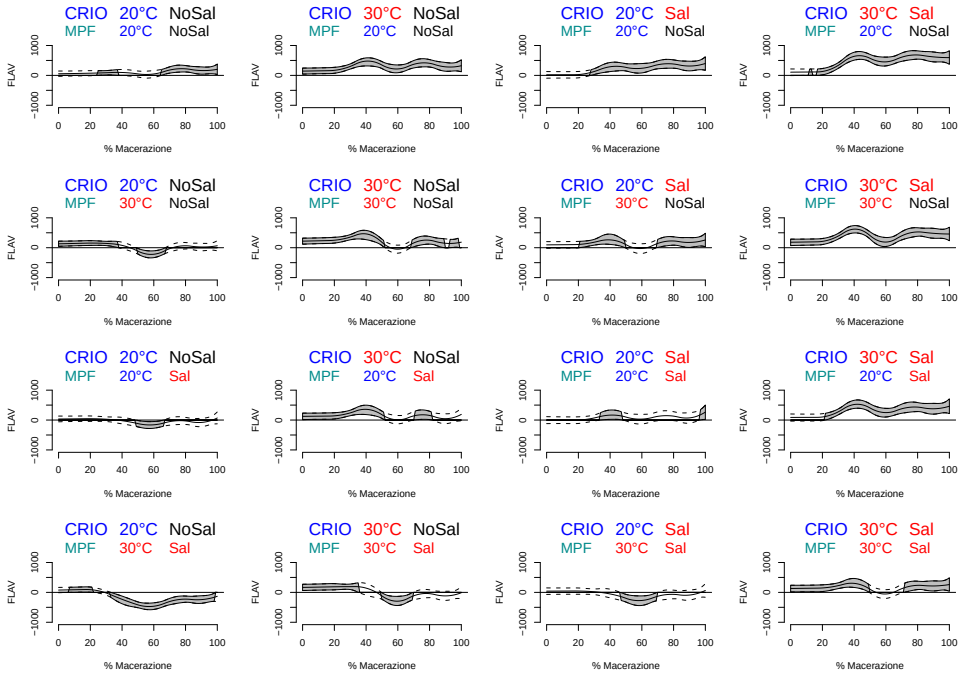

Figura 36: Contrasti per FLAV (2008)

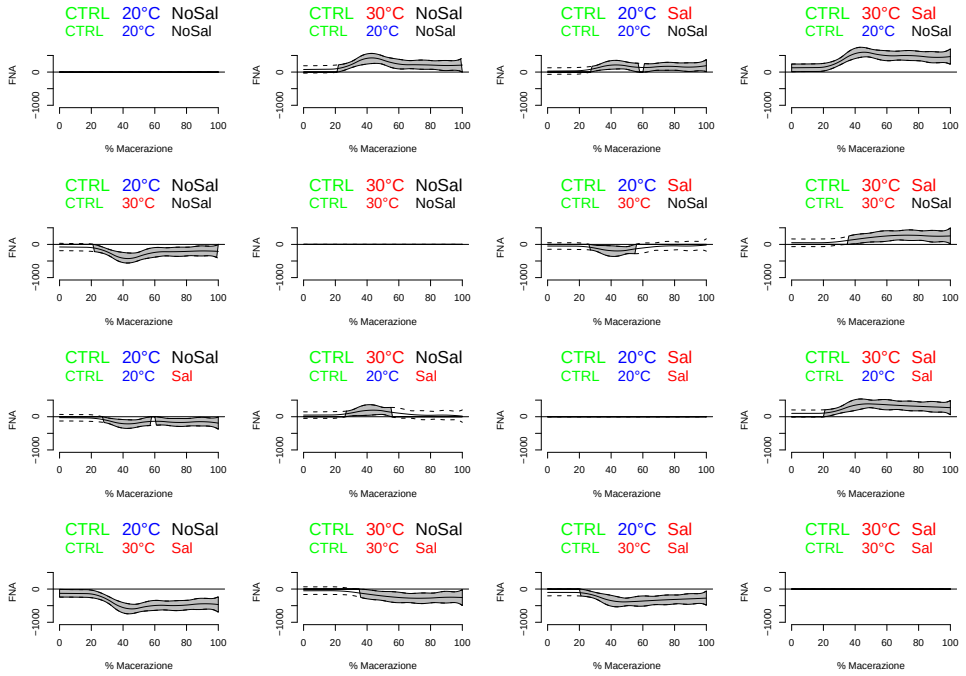

Figura 37: Contrasti per FNA (2008)

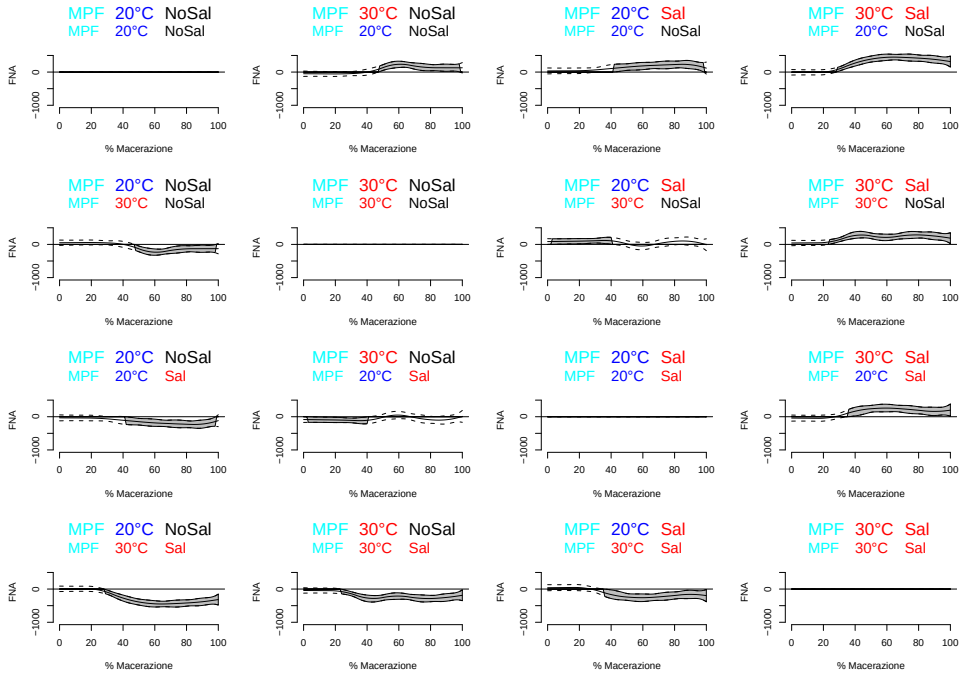

Figura 38: Contrasti per FNA (2008)

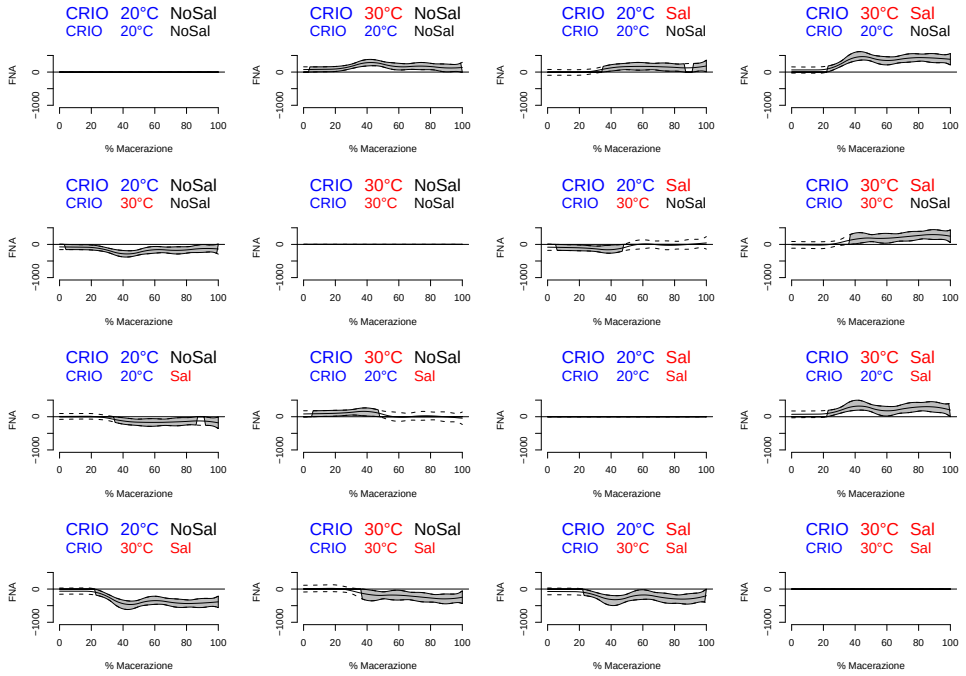

Figura 39: Contrasti per FNA (2008)

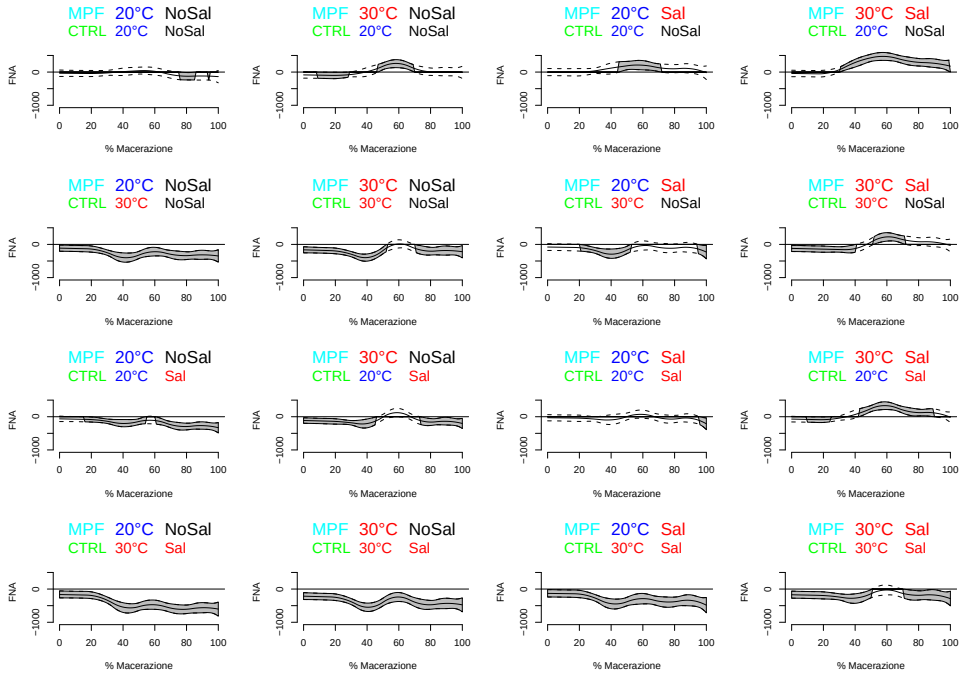

Figura 40: Contrasti per FNA (2008)

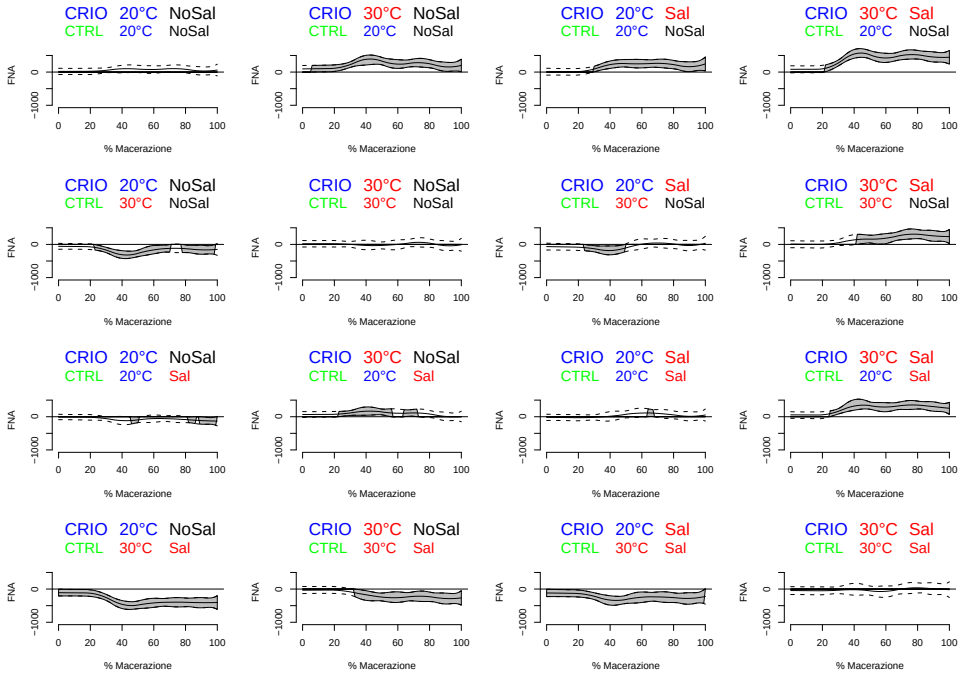

Figura 41: Contrasti per FNA (2008)

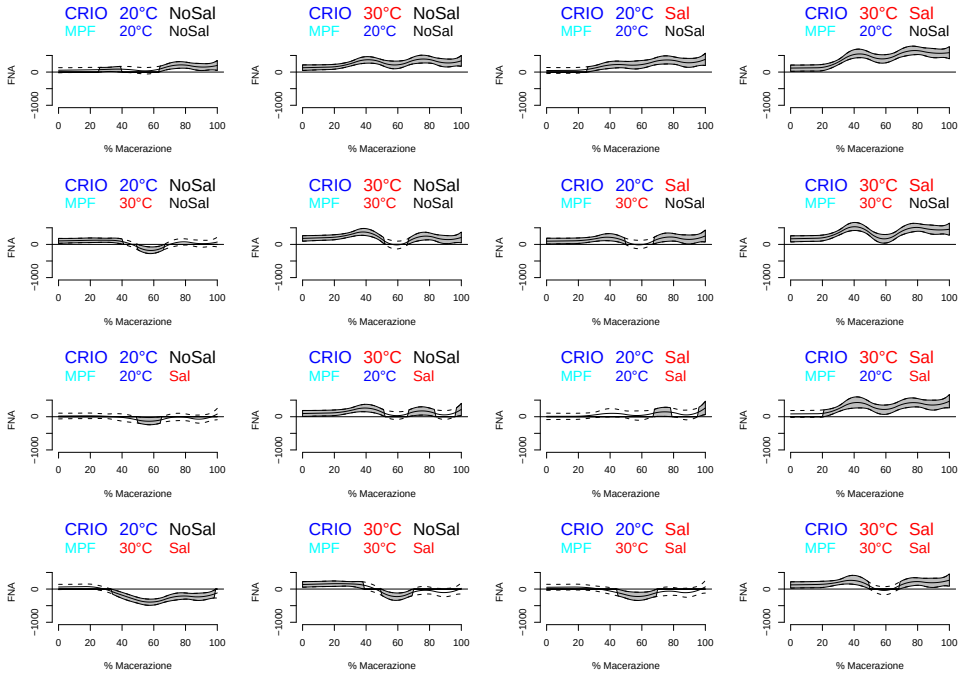

Figura 42: Contrasti per FNA (2008)

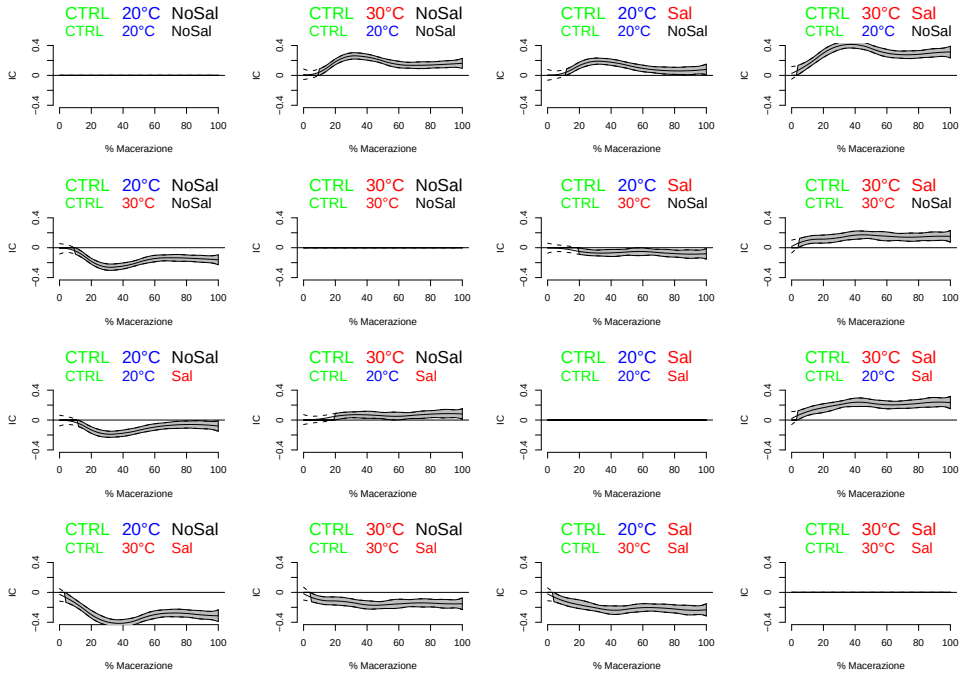

Figura 43: Contrasti per FNA (2008)

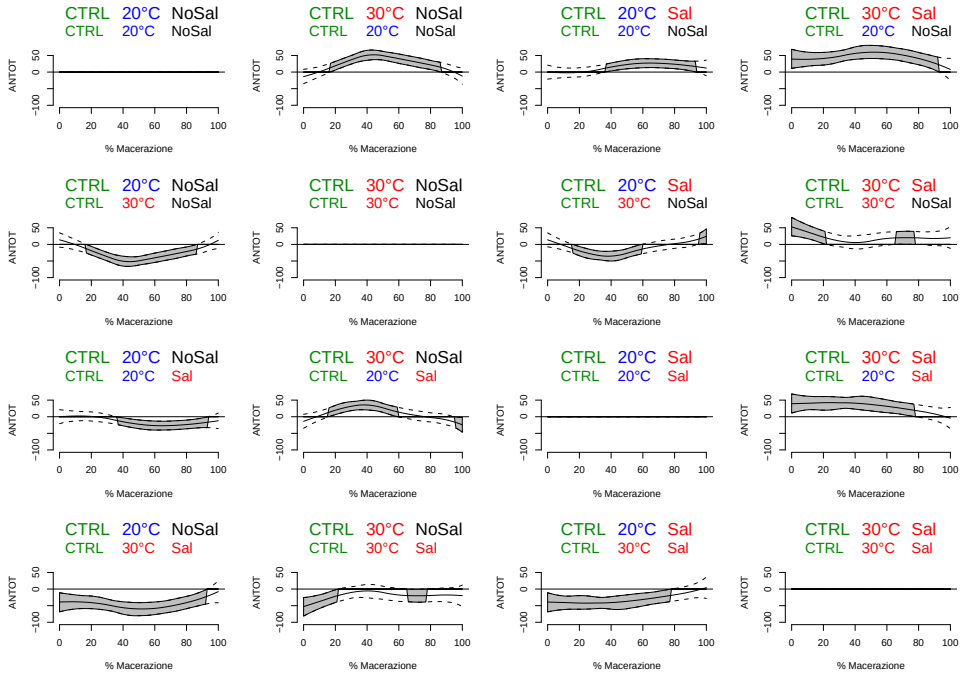

Figura 44: Contrasti per ANTOT (2008)

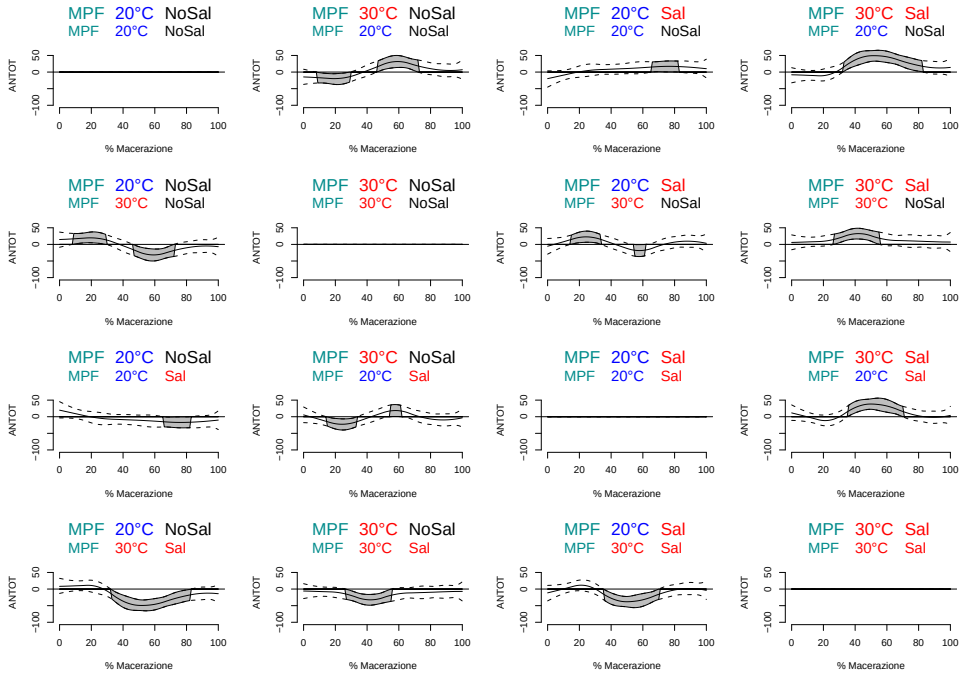

Figura 45: Contrasti per ANTOT (2008)

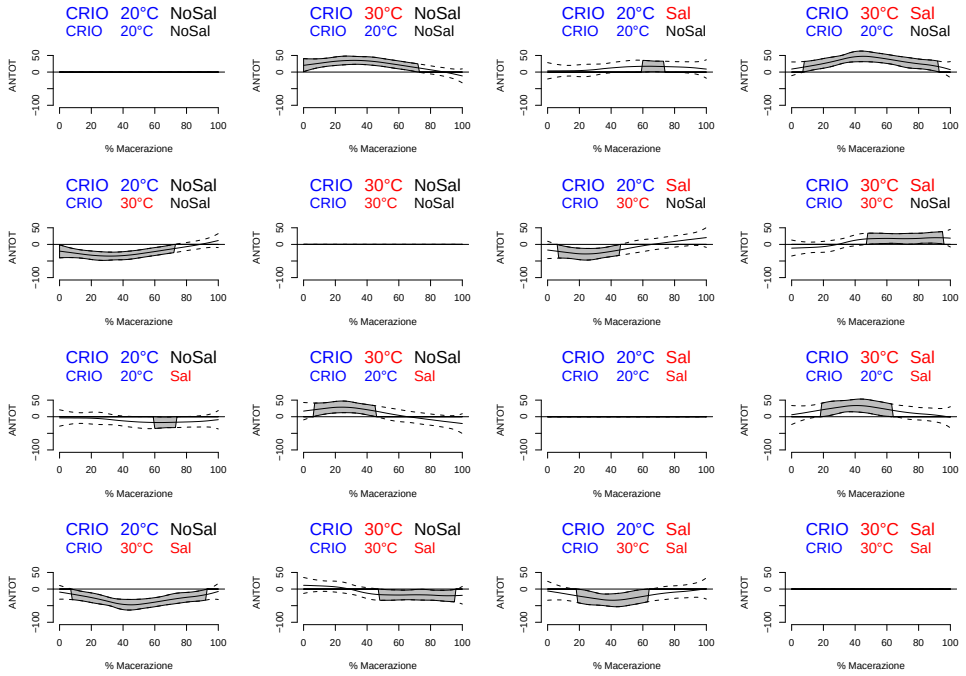

Figura 46: Contrasti per ANTOT (2008)

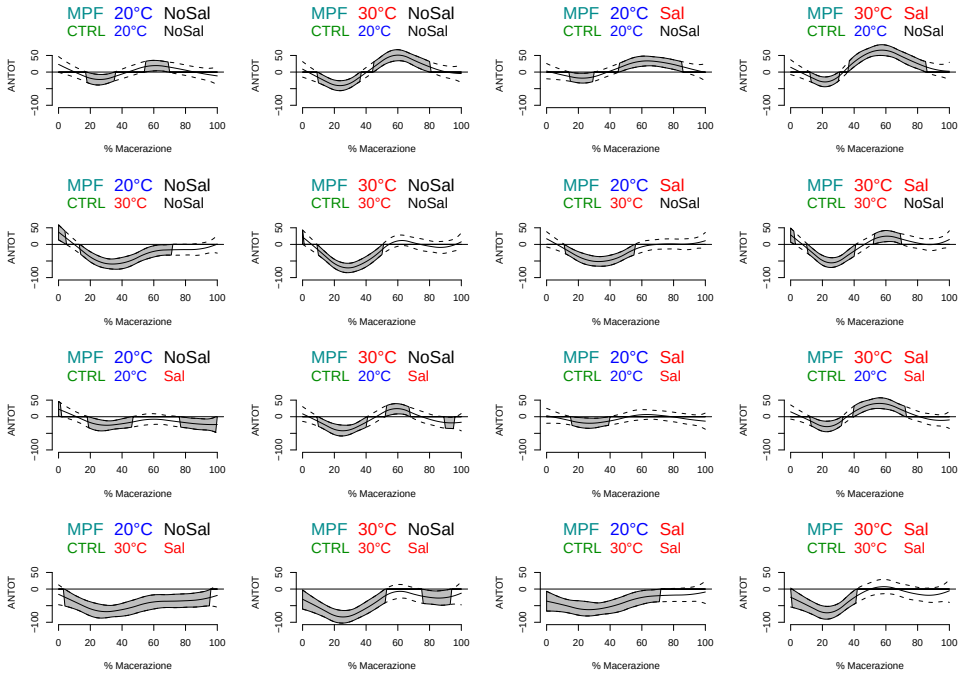

Figura 47: Contrasti per ANTOT (2008)

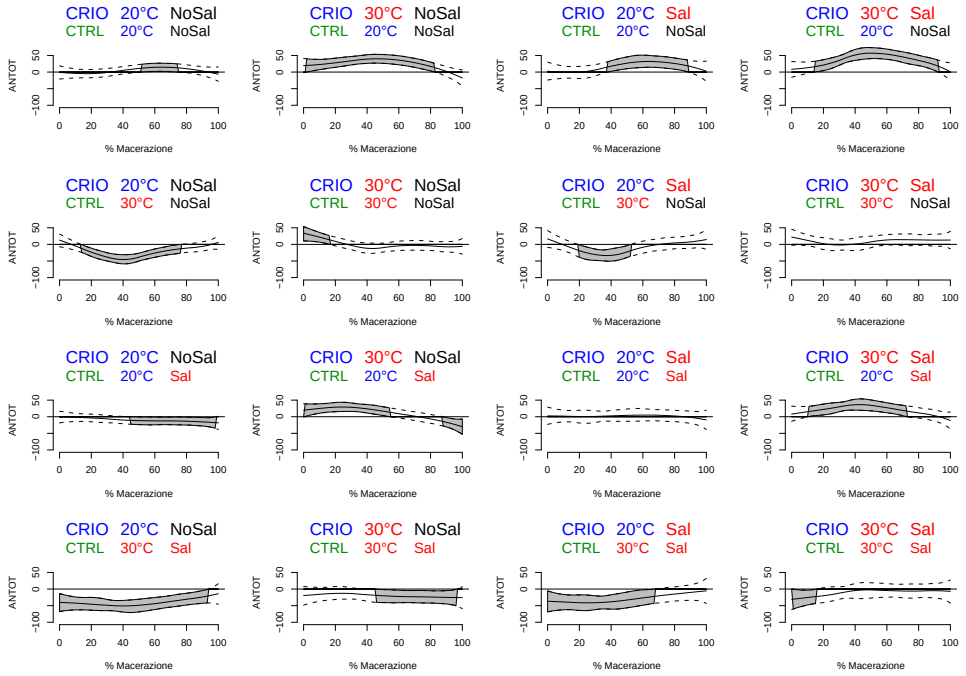

Figura 48: Contrasti per ANTOT (2008)

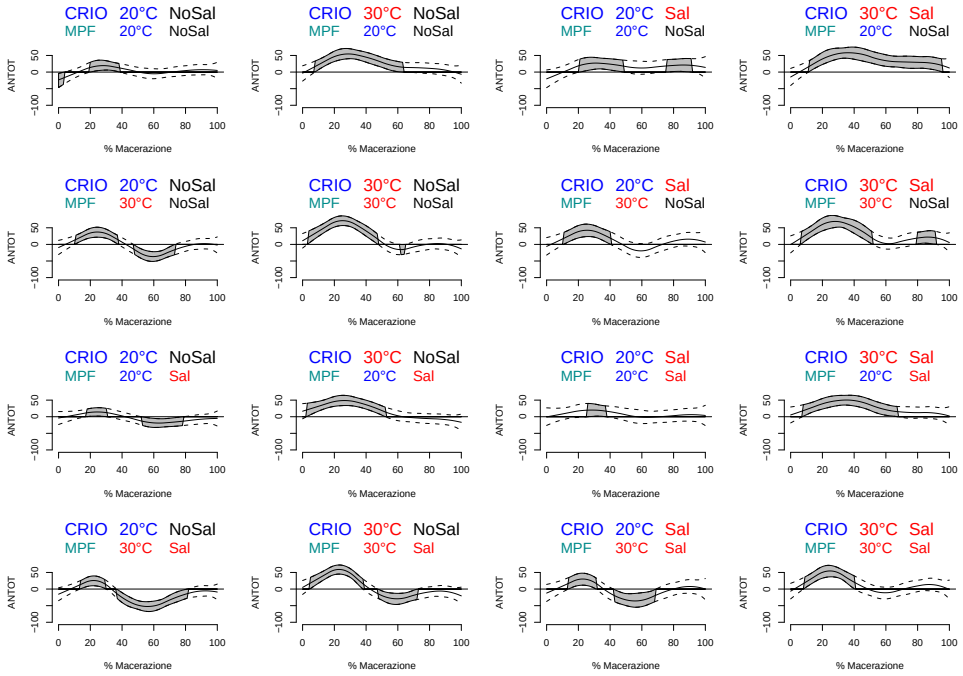

Figura 49: Contrasti per ANTOT (2008)

4 Grafici del 2009

In questa sezione sono riportati i grafici relativi ai risultati del 2009. In primo luogo sono riportate le cinetiche, quindi i contrasti funzionali.

I grafici dei contrasti vanno letti come segue: la linea superiore del titolo di ogni grafico indica il trattamento sotto esame, mentre la riga sottostante indica il trattamento di riferimento. L'area grigia indica i tempi in cui i due trattamenti sono diversi con un intervallo di credibilità del 95 %.

I parametri TON e IC non sono stati analizzati perchè non ritenuti meritevoli di approfondimento sulla base di quanto valutato per la stagione 2008. I grafici delle rilevazioni giornaliere sono stati comunque riportati.

4.1 Cinetiche del 2009

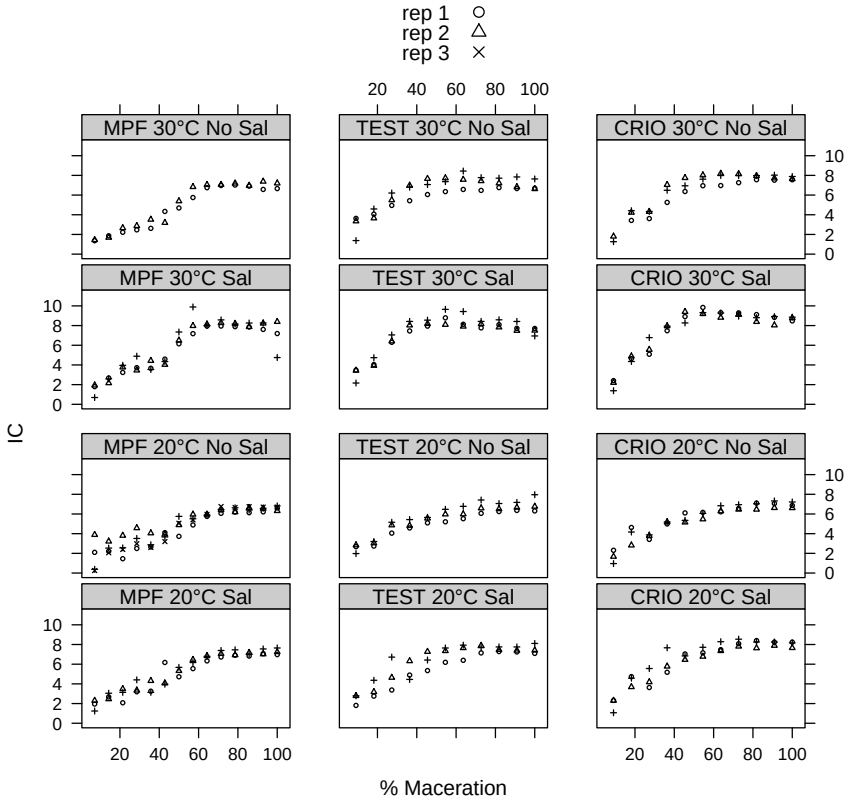

Figura 50: Cinetica di IC (2009)

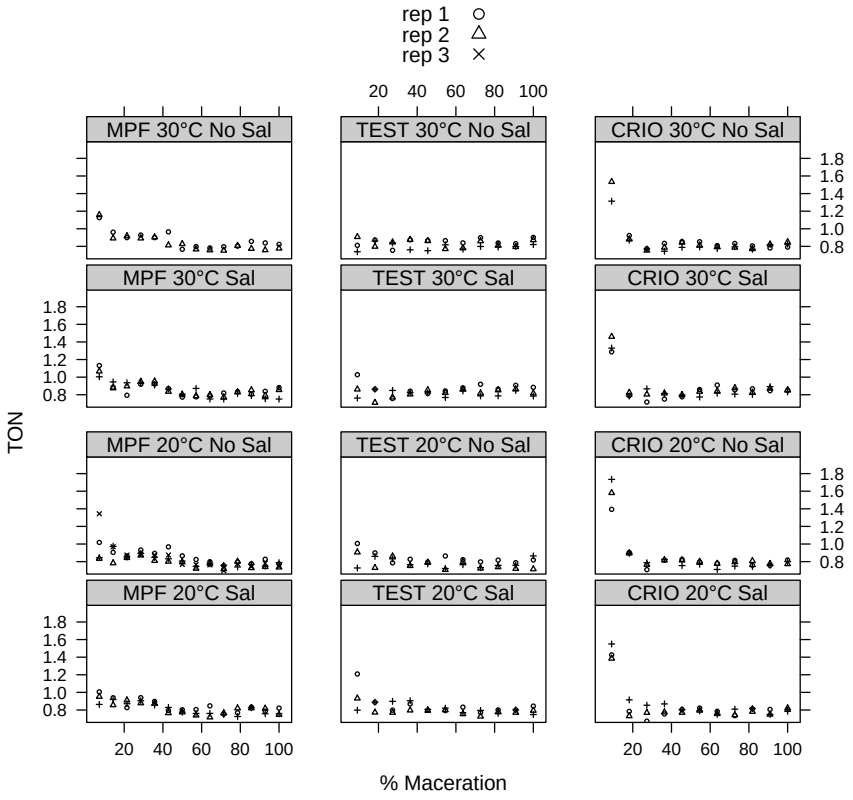

Figura 51: Cinetica di TON (2009)

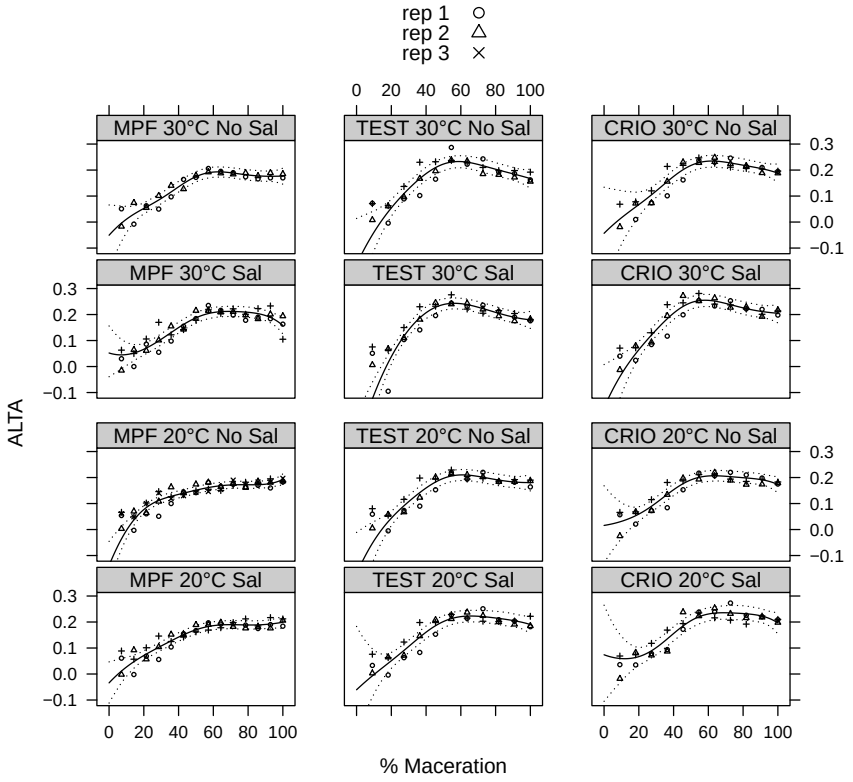

Figura 52: Cinetica di ALTA (2009)

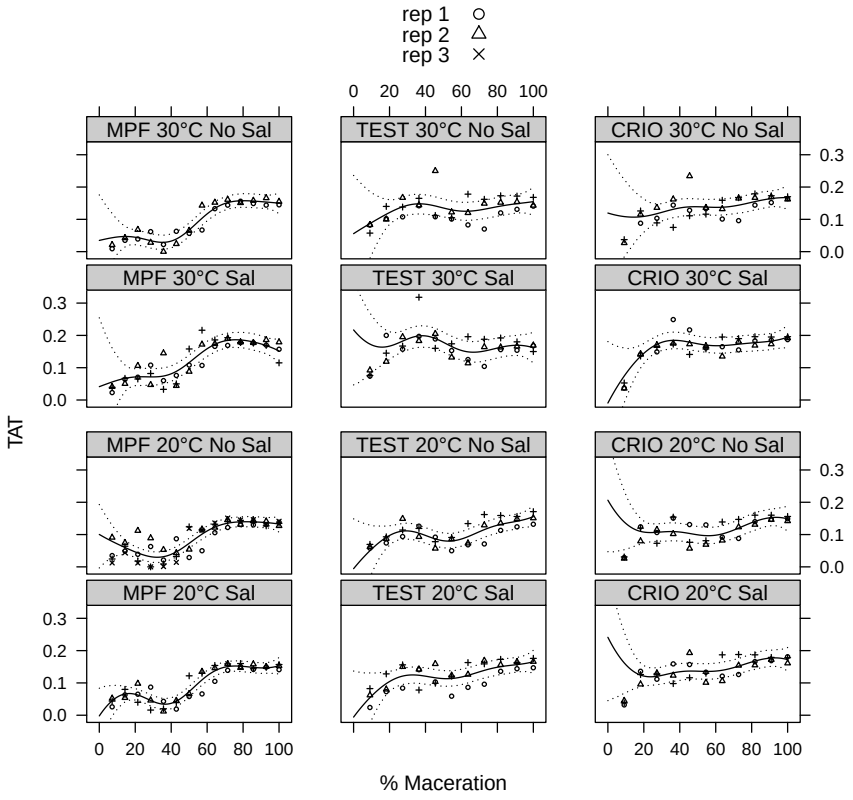

Figura 53: Cinetica di TAT (2009)

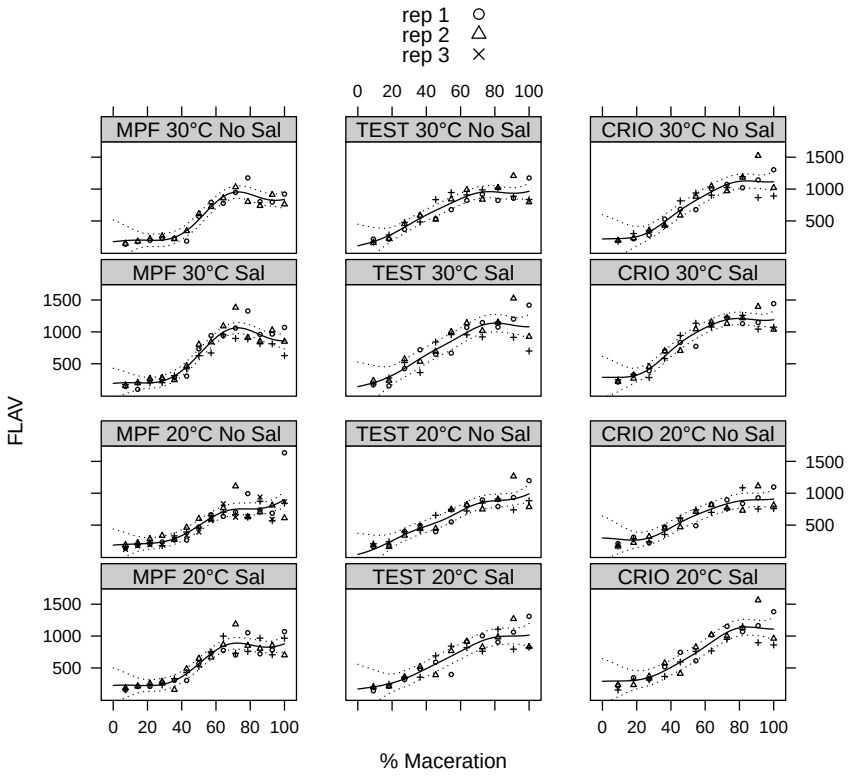

Figura 54: Cinetica di FLAV (2009)

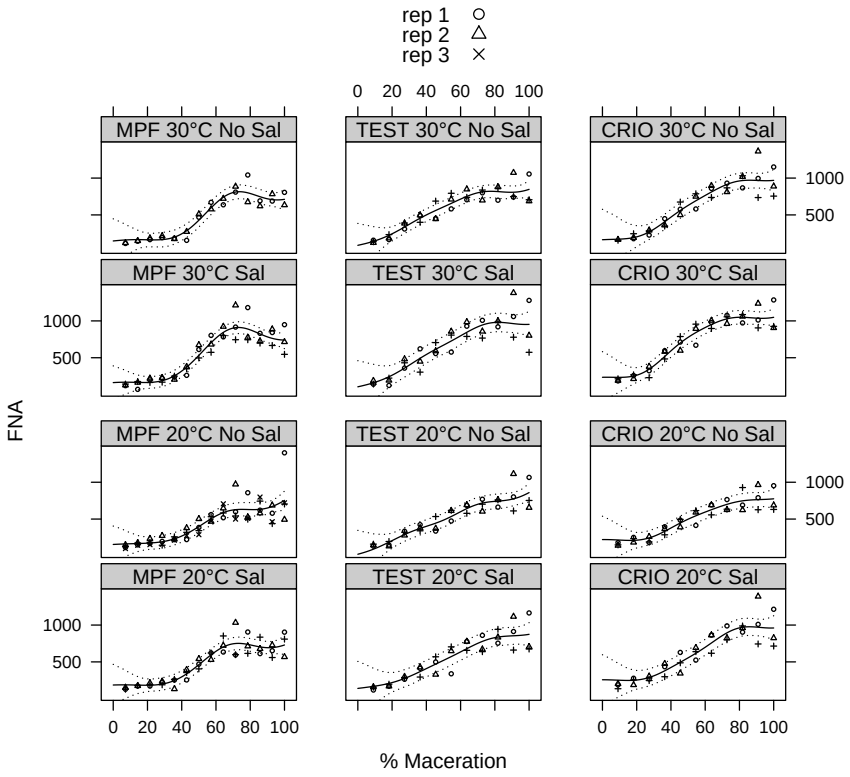

Figura 55: Cinetica di FNA (2009)

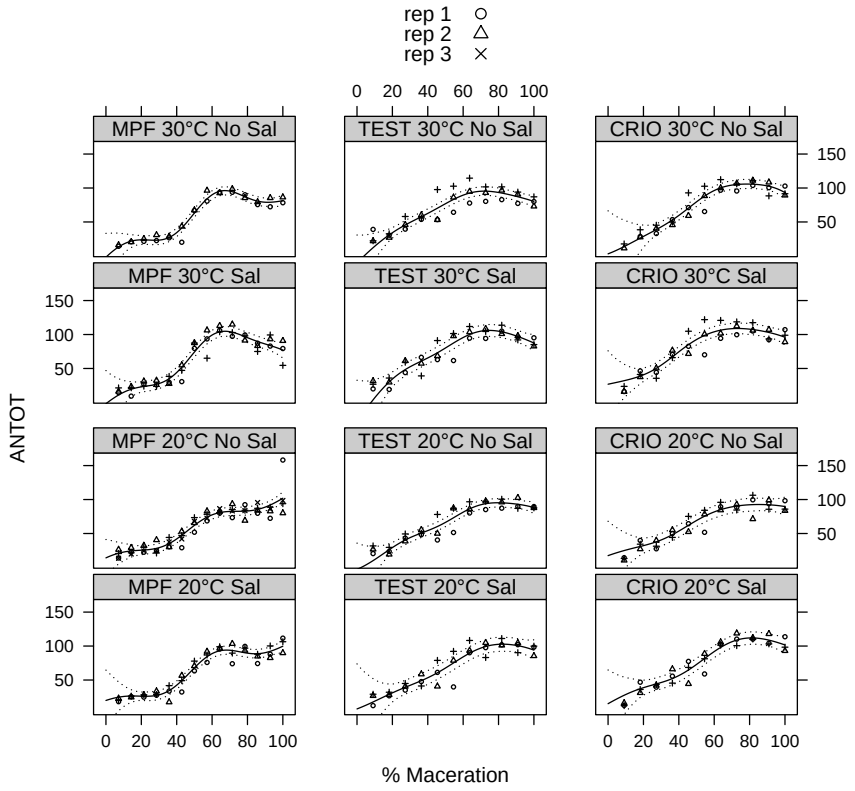

Figura 56: Cinetica di ANTOT (2009)

4.2 Contrasti del 2009

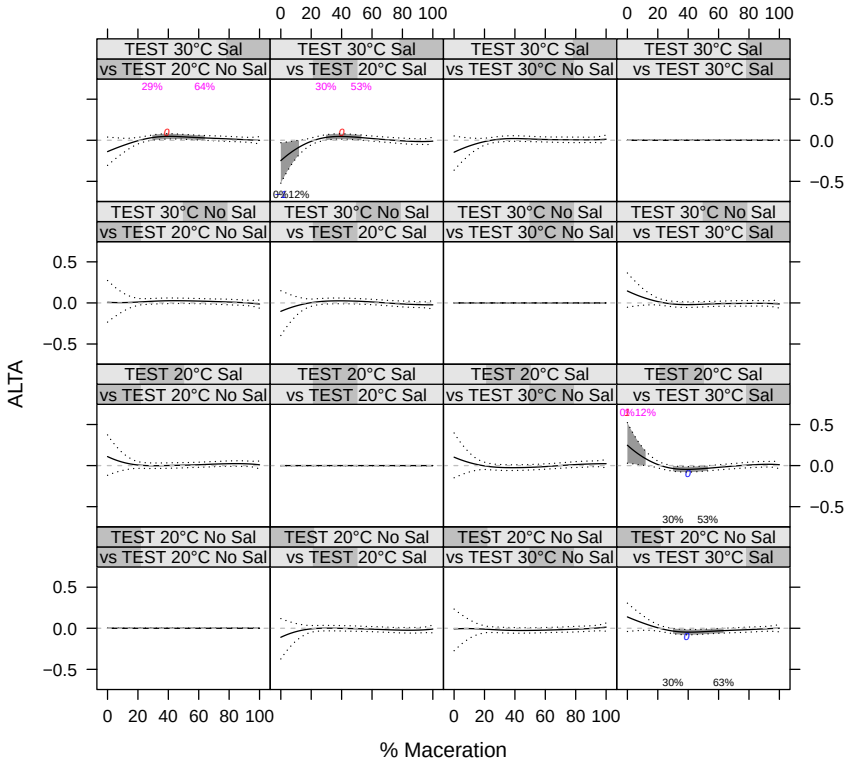

Figura 57: Contrasti per ALTA (2009): test contro test

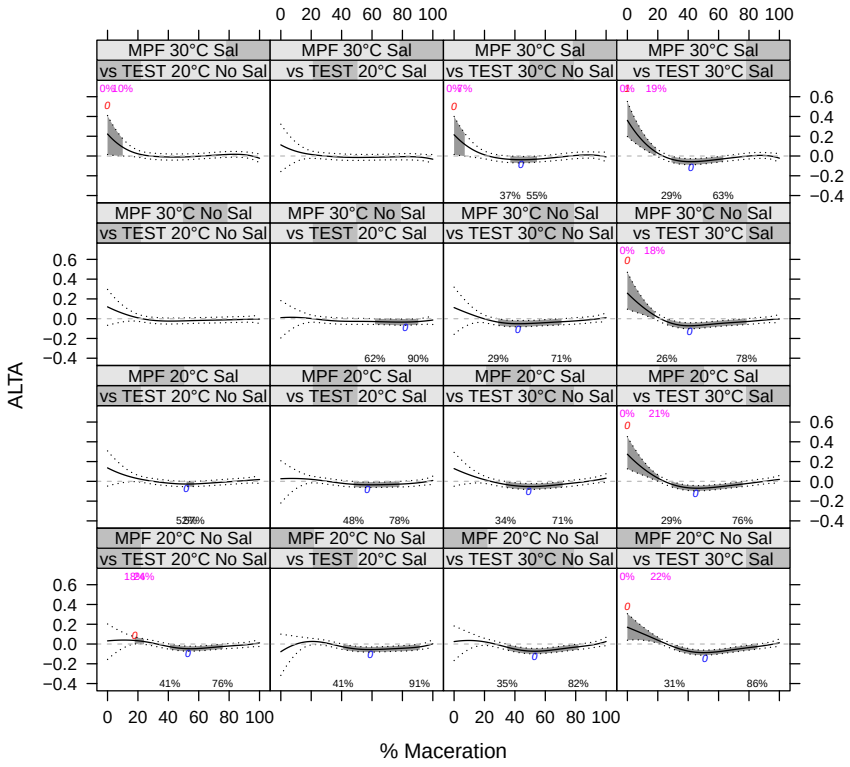

Figura 58: Contrasti per ALTA (2009): mpf contro test

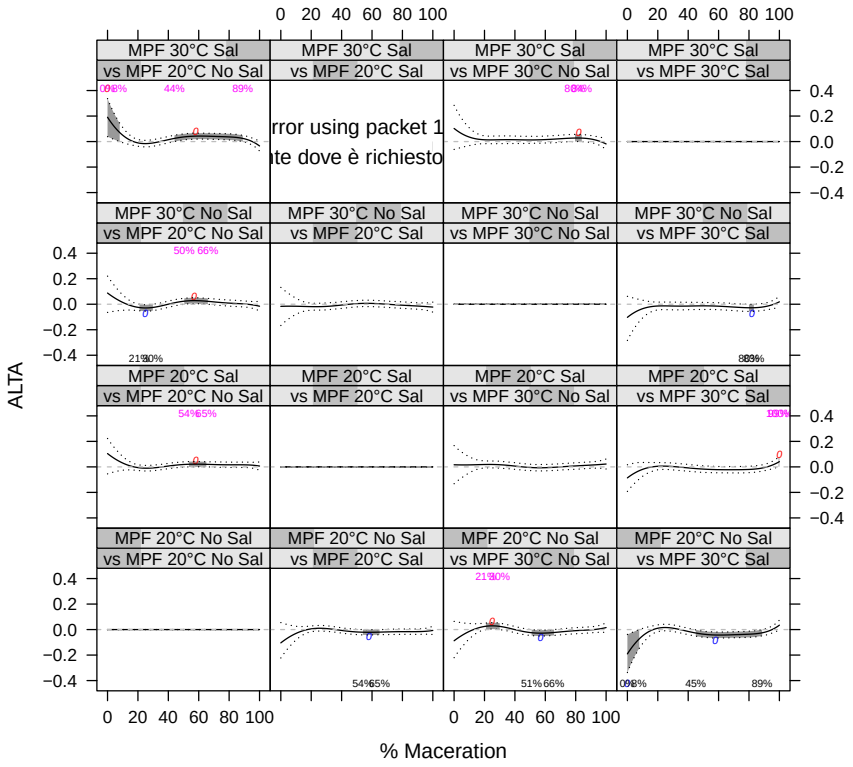

Figura 59: Contrasti per ALTA (2009): mpf contro mpf

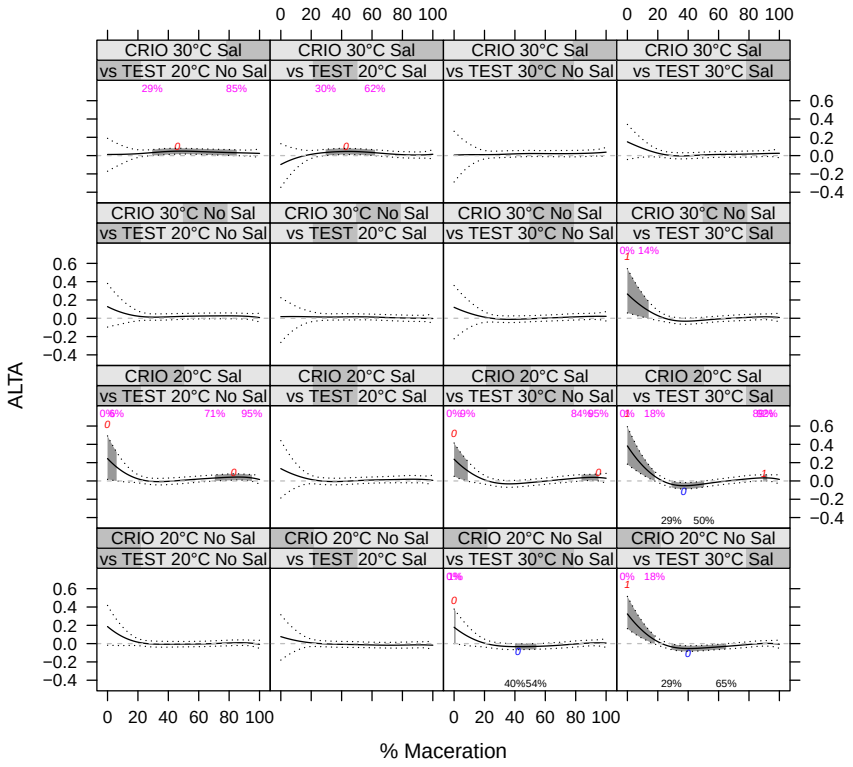

Figura 60: Contrasti per ALTA (2009): crio contro test

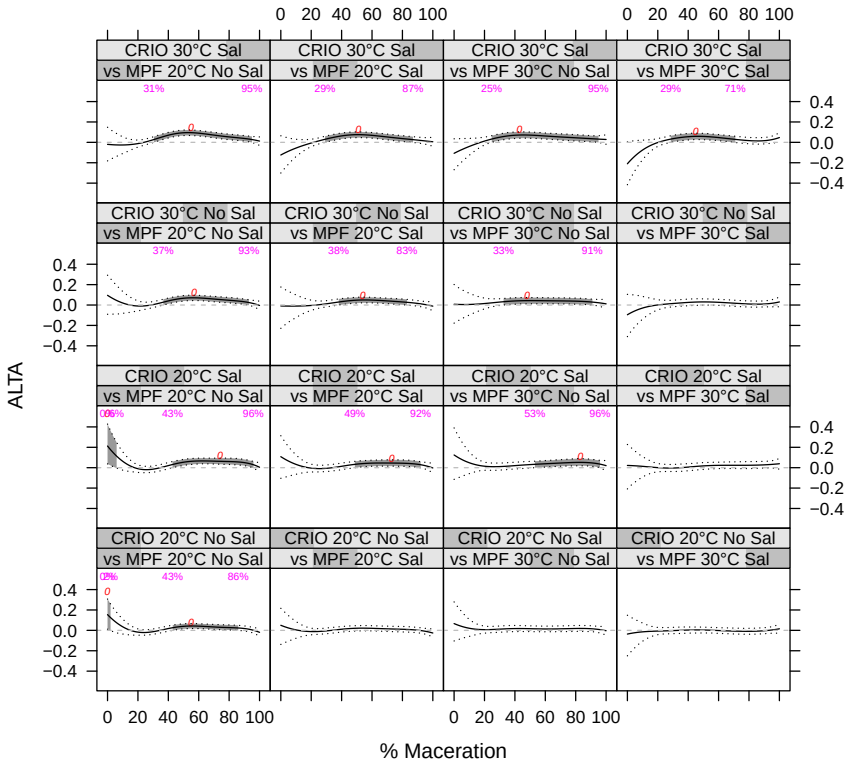

Figura 61: Contrasti per ALTA (2009): crio contro mpf

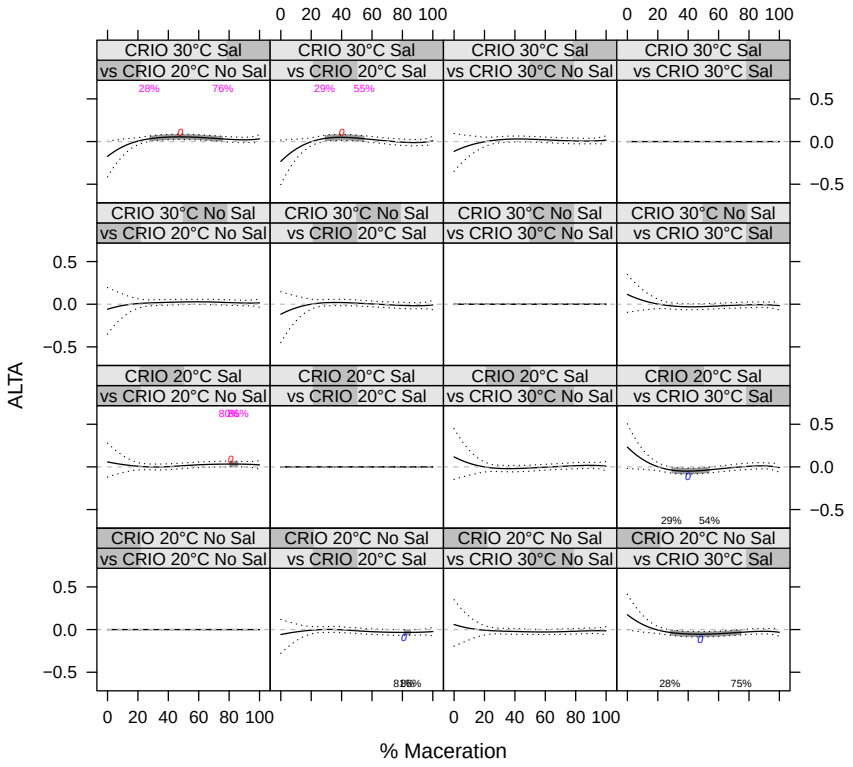

Figura 62: Contrasti per ALTA (2009): crio contro crio

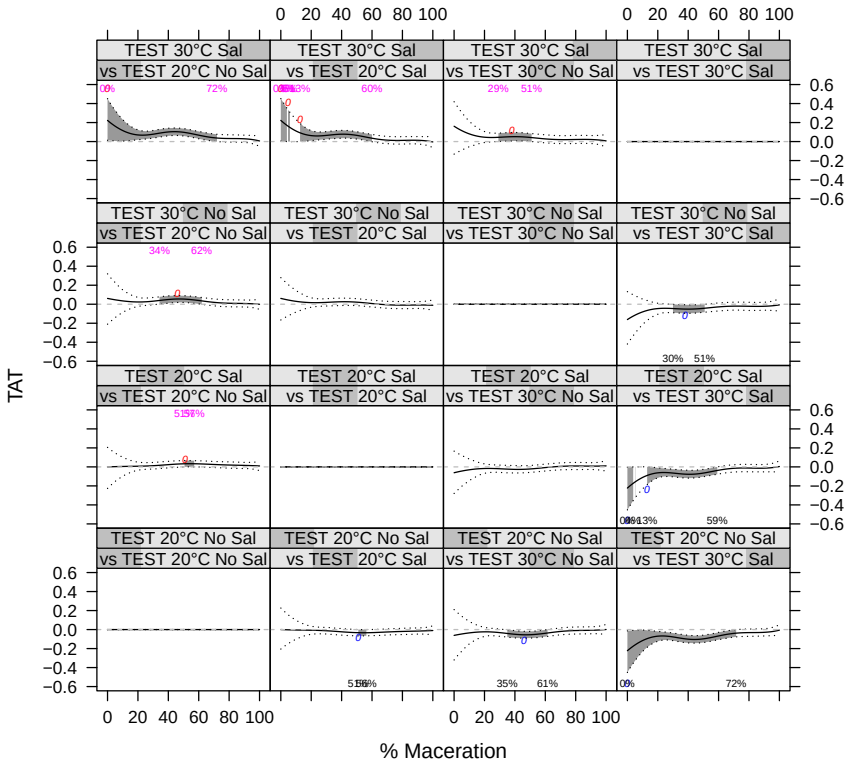

Figura 63: Contrasti per TAT (2009): test contro test

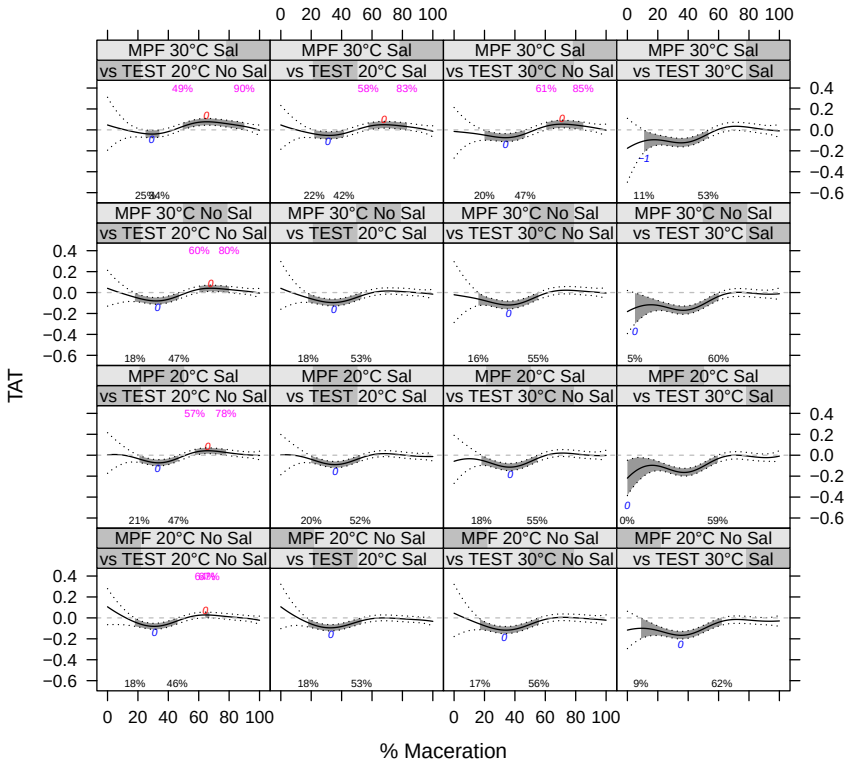

Figura 64: Contrasti per TAT (2009): mpf contro test

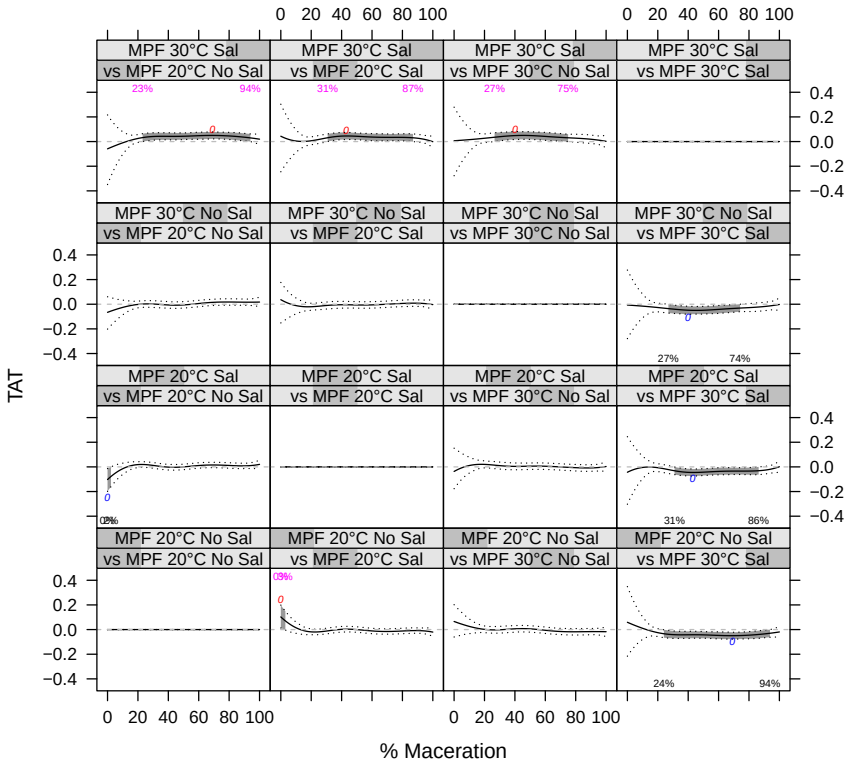

Figura 65: Contrasti per TAT (2009): mpf contro mpf

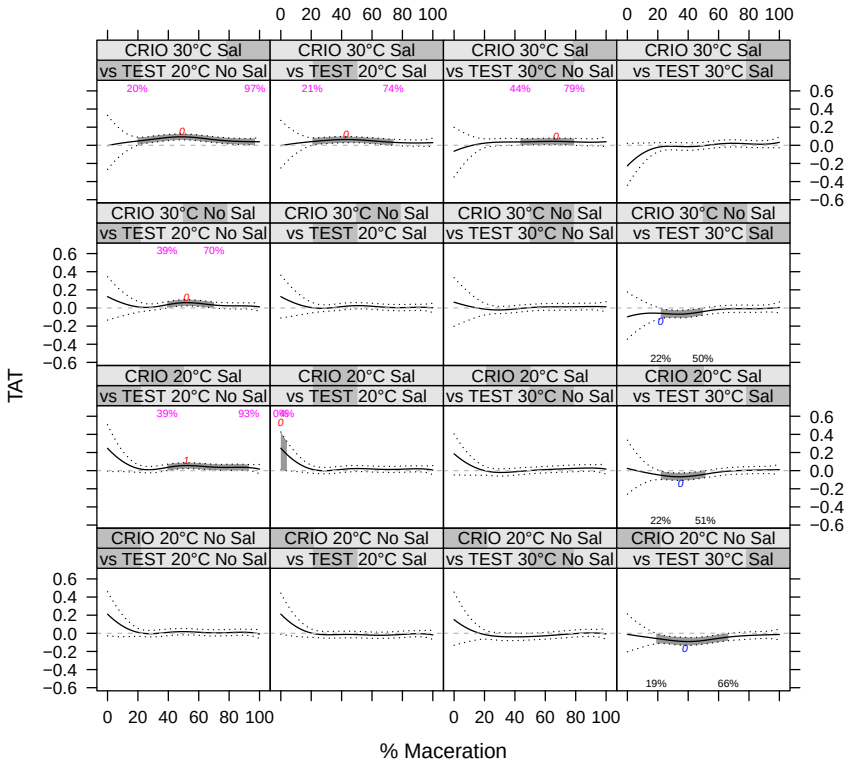

Figura 66: Contrasti per TAT (2009): crio contro test

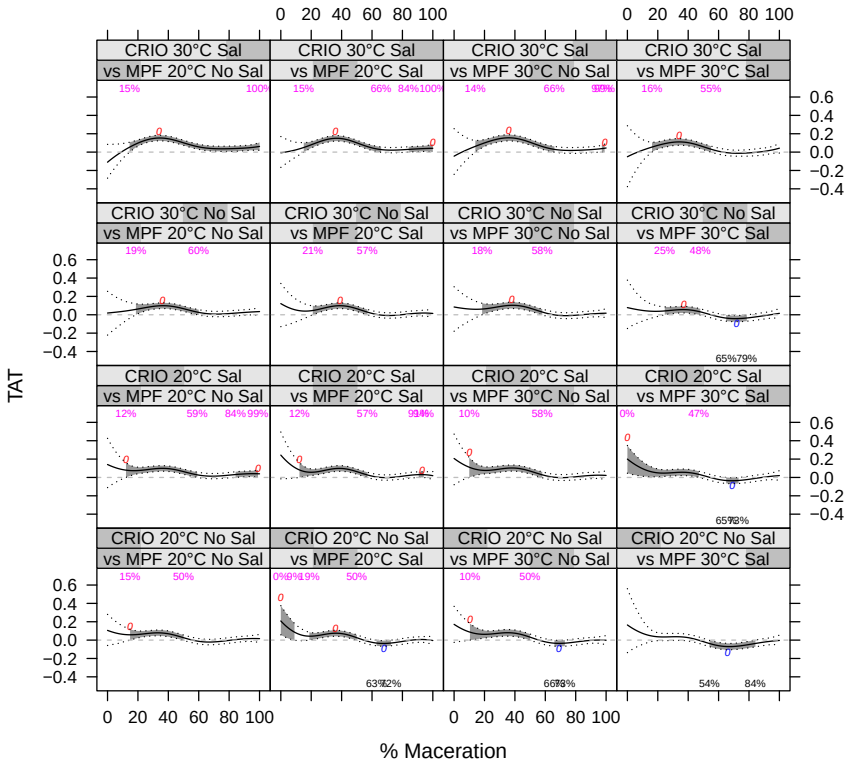

Figura 67: Contrasti per TAT (2009): crio contro mpf

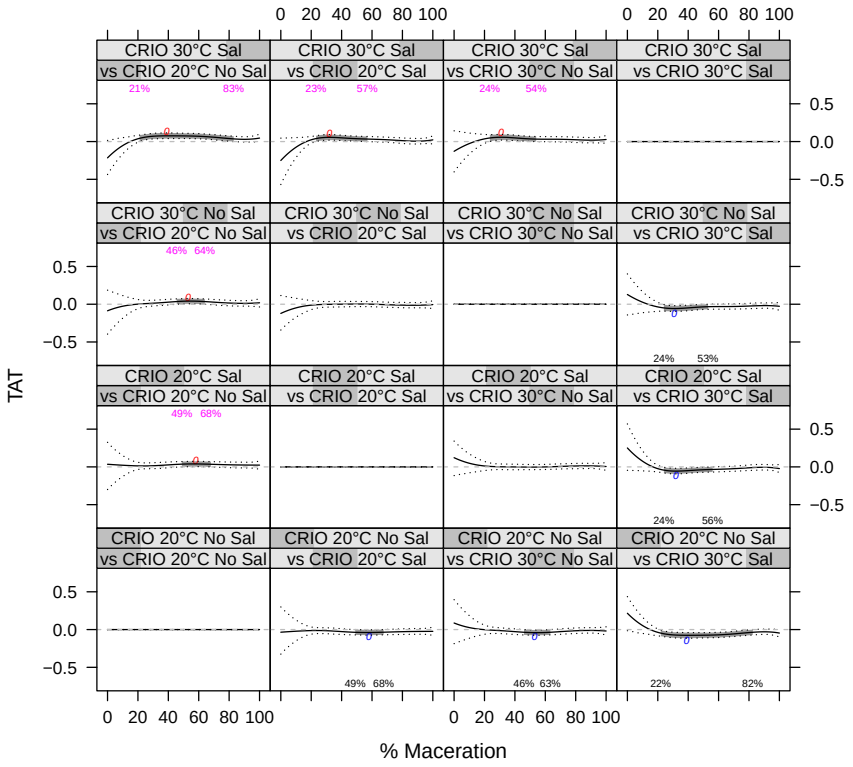

Figura 68: Contrasti per TAT (2009): crio contro crio

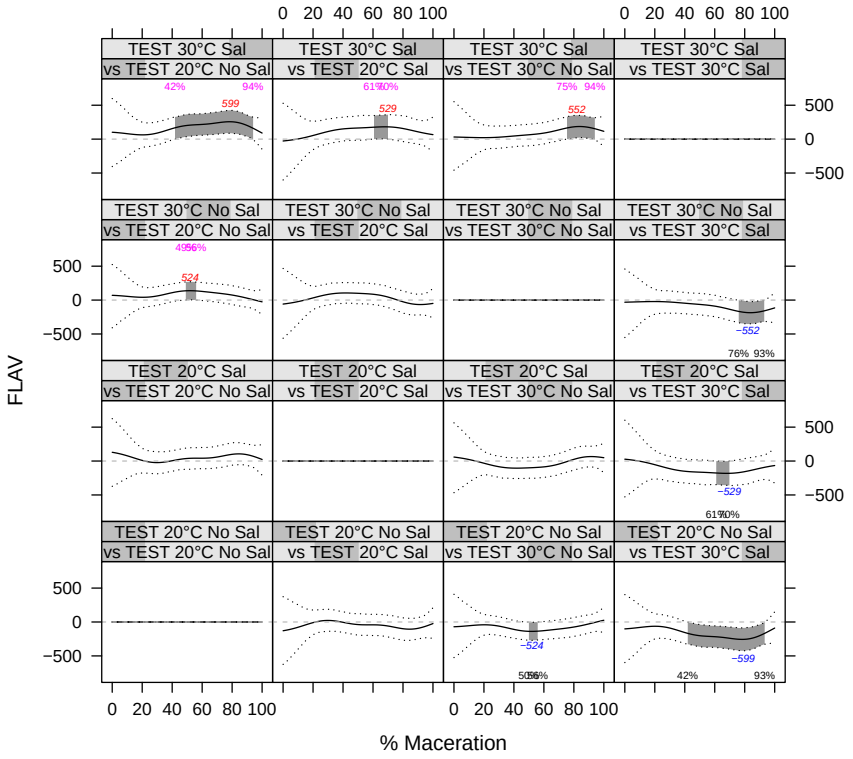

Figura 69: Contrasti per FLAV (2009): test contro test

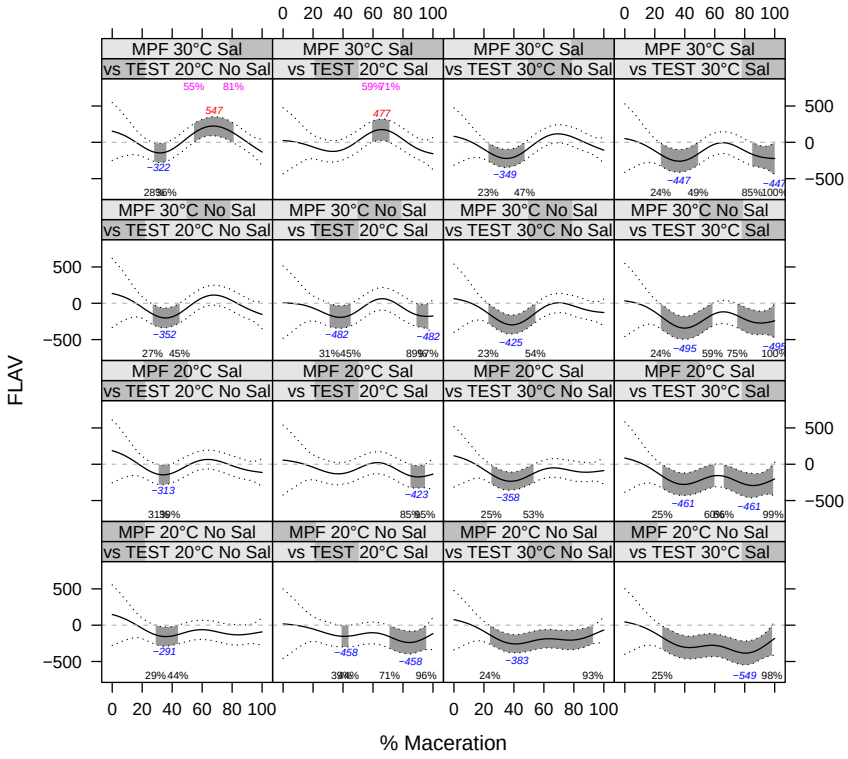

Figura 70: Contrasti per FLAV (2009): mpf contro test

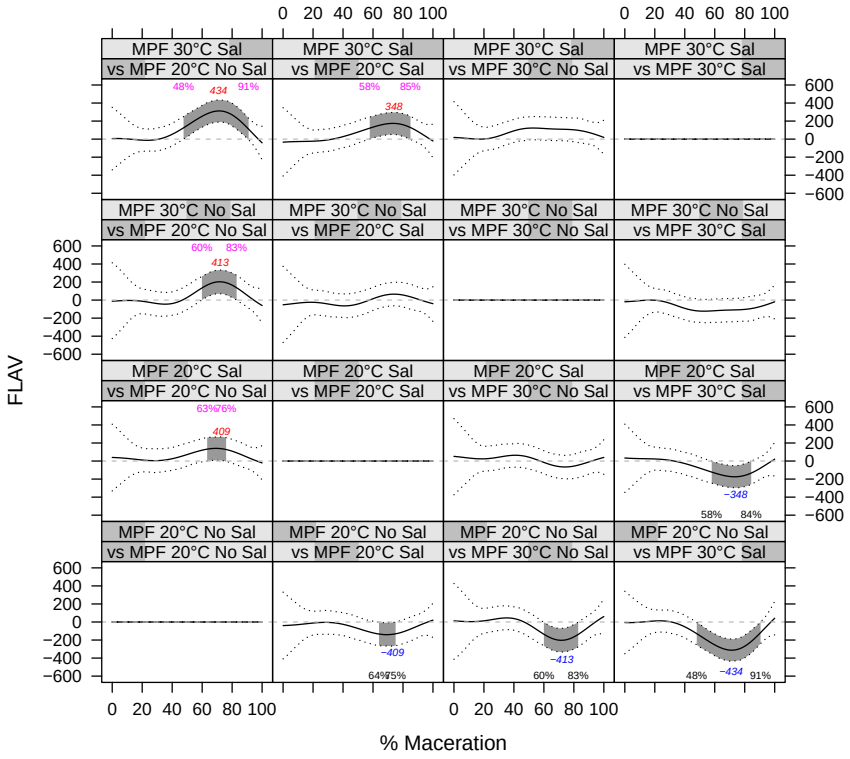

Figura 71: Contrasti per FLAV (2009): mpf contro mpf

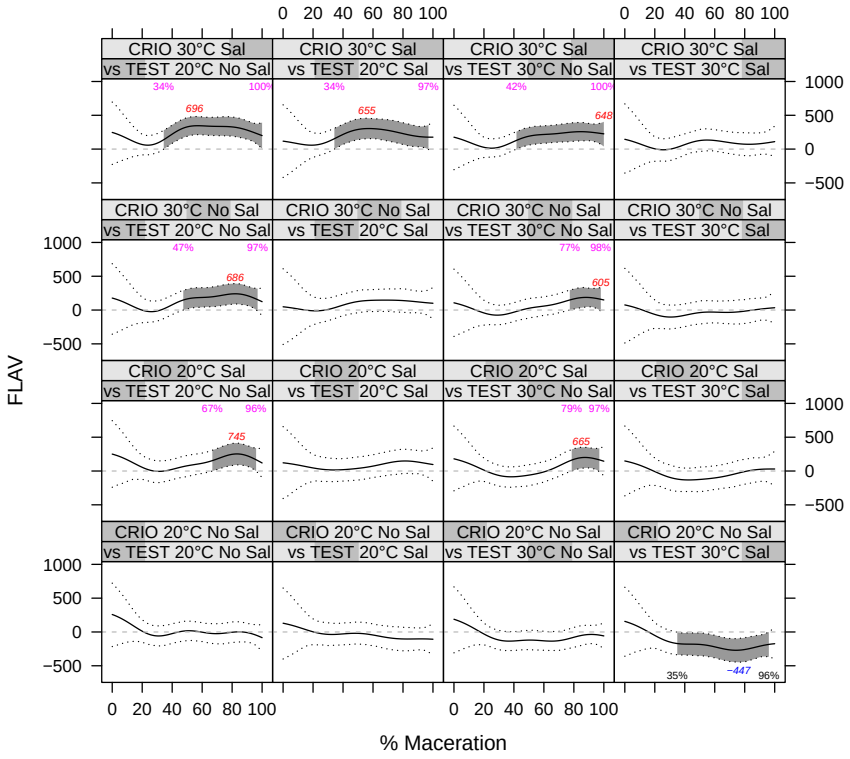

Figura 72: Contrasti per FLAV (2009): crio contro test

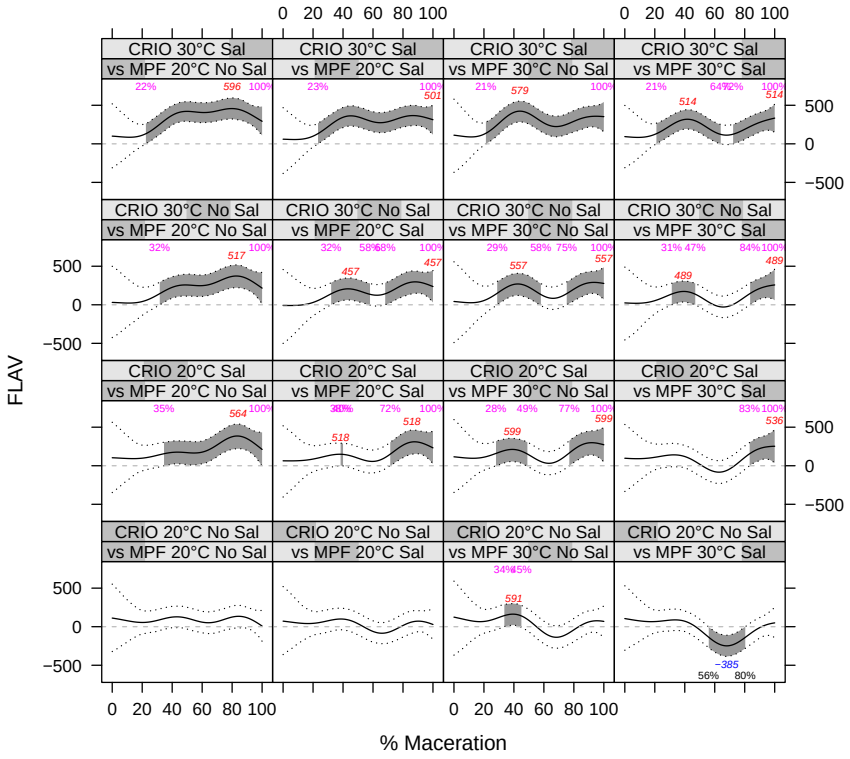

Figura 73: Contrasti per FLAV (2009): crio contro mpf

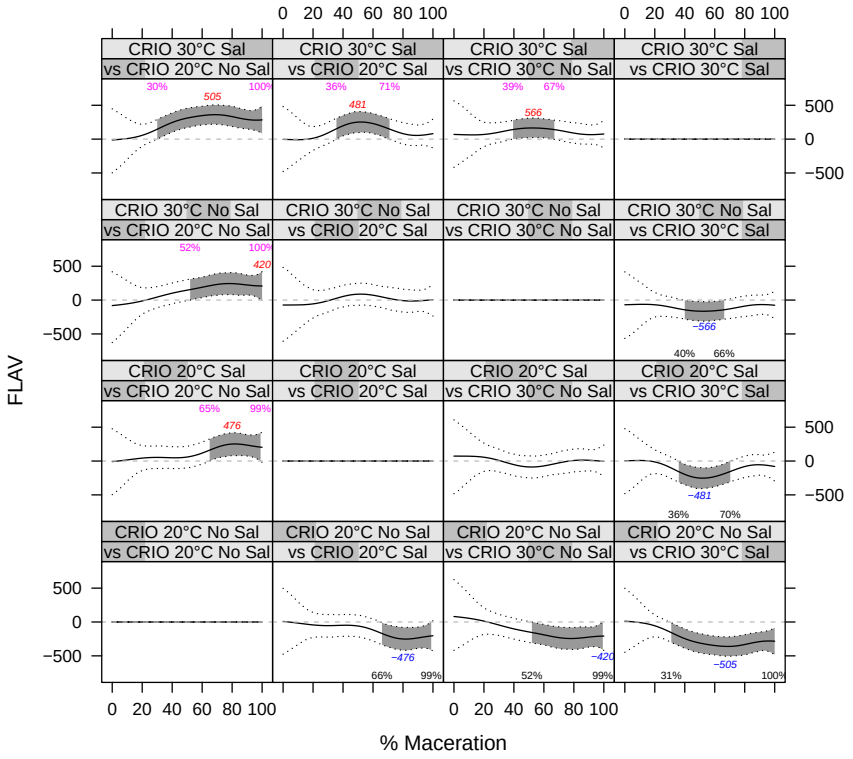

Figura 74: Contrasti per FLAV (2009): crio contro crio

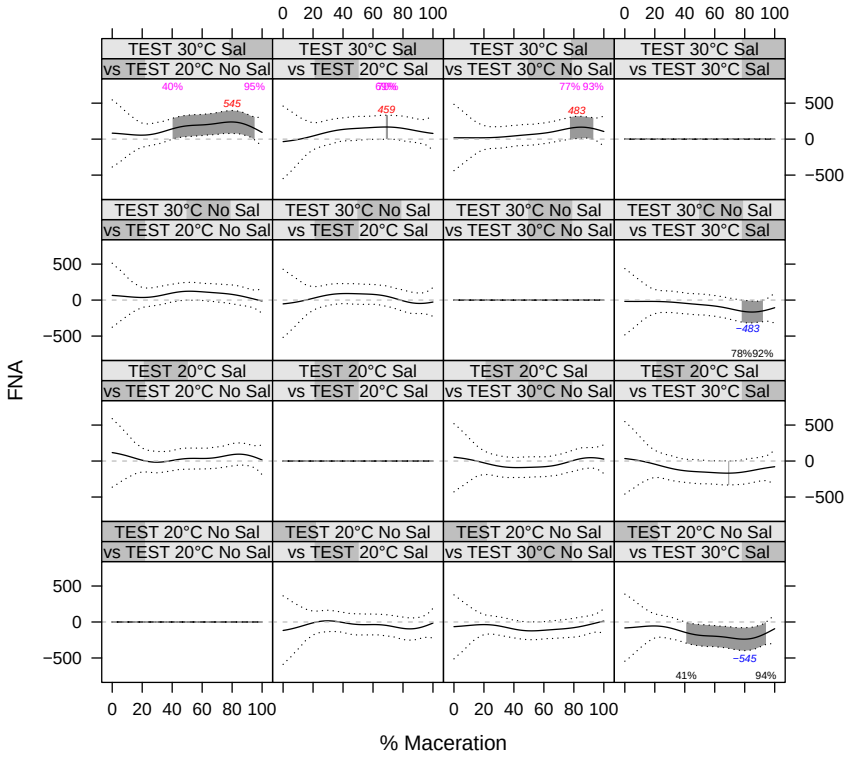

Figura 75: Contrasti per FNA (2009): test contro test

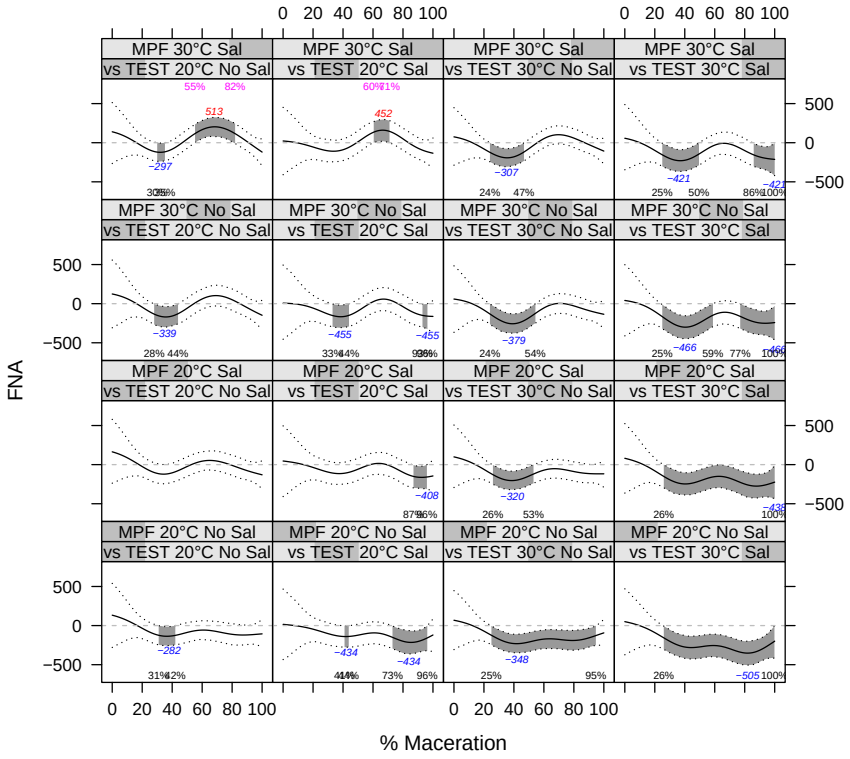

Figura 76: Contrasti per FNA (2009): mpf contro test

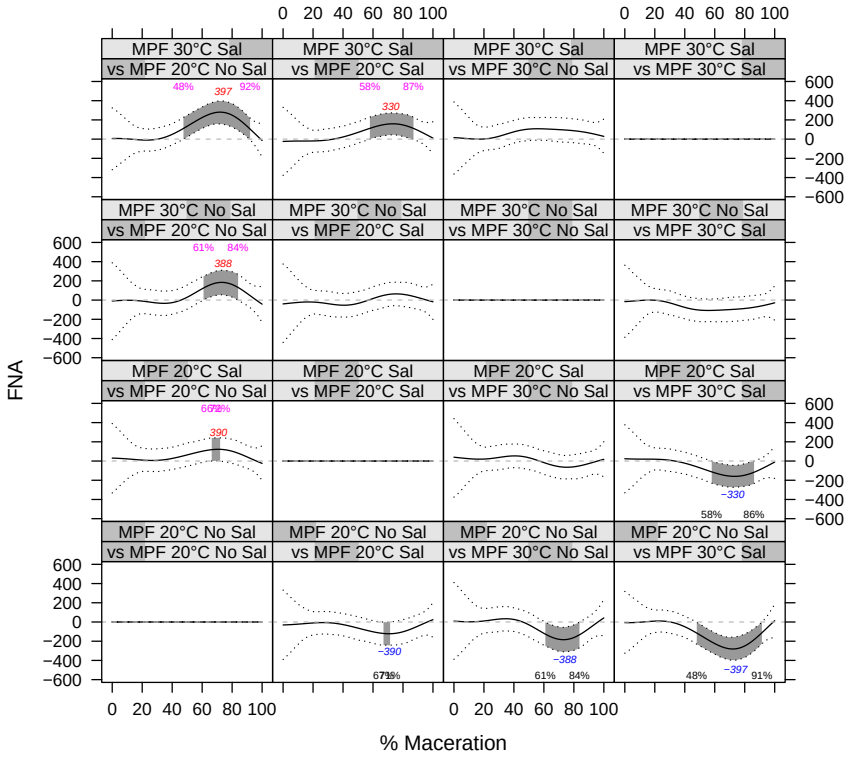

Figura 77: Contrasti per FNA (2009): mpf contro mpf

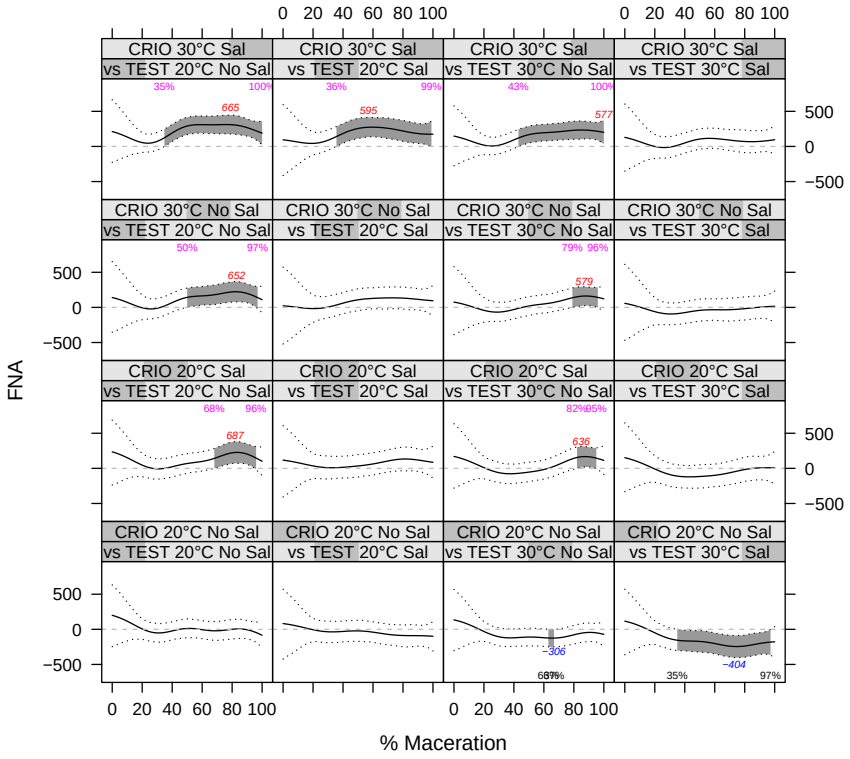

Figura 78: Contrasti per FNA (2009): crio contro test

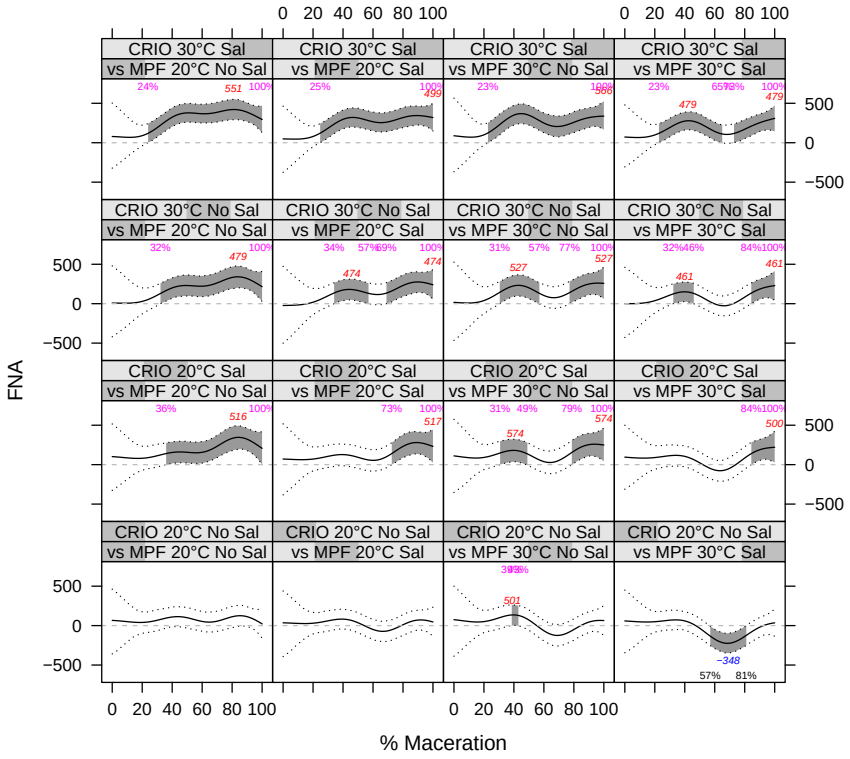

Figura 79: Contrasti per FNA (2009): crio contro mpf

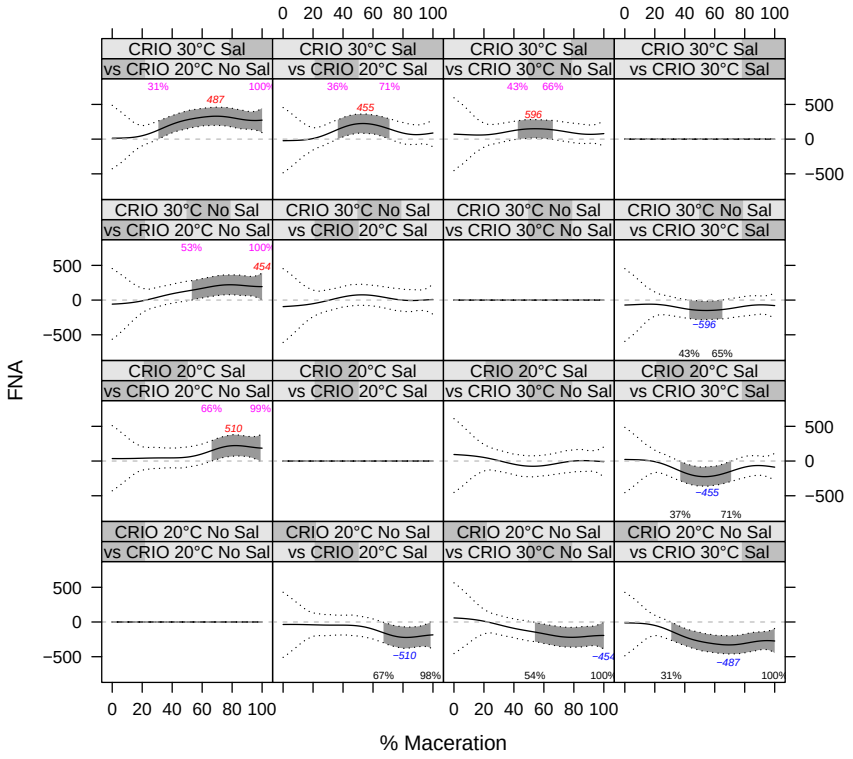

Figura 80: Contrasti per FNA (2009): crio contro crio

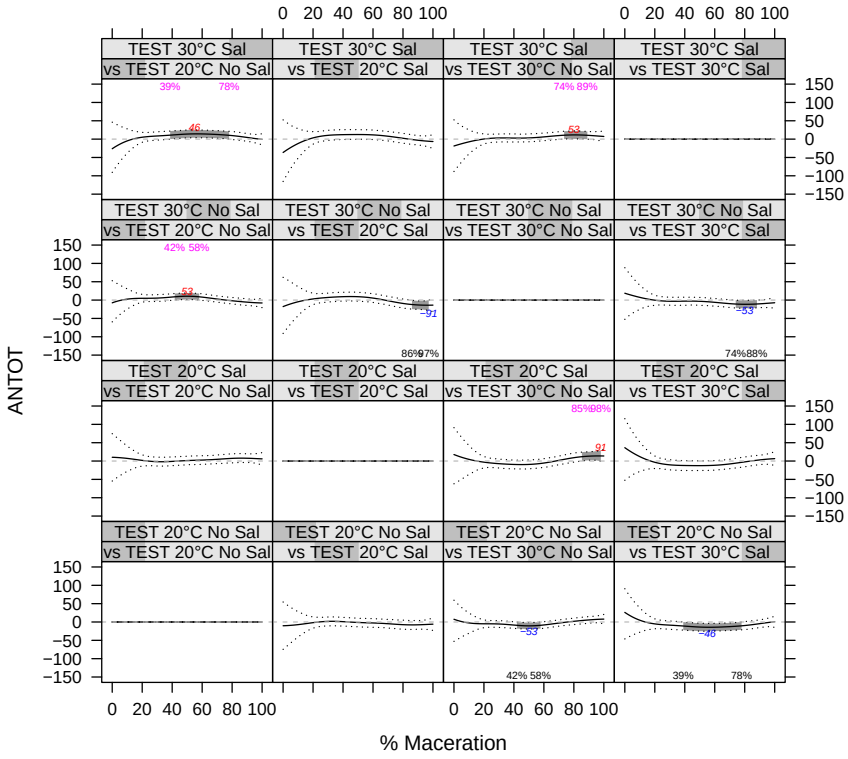

Figura 81: Contrasti per ANTOT (2009): test contro test

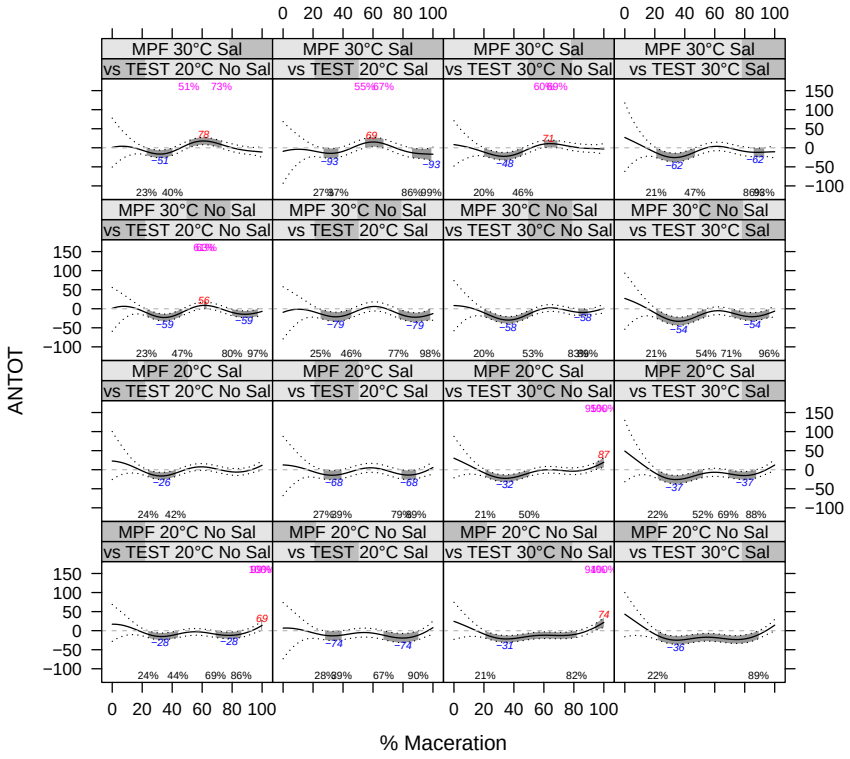

Figura 82: Contrasti per ANTOT (2009): mpf contro test

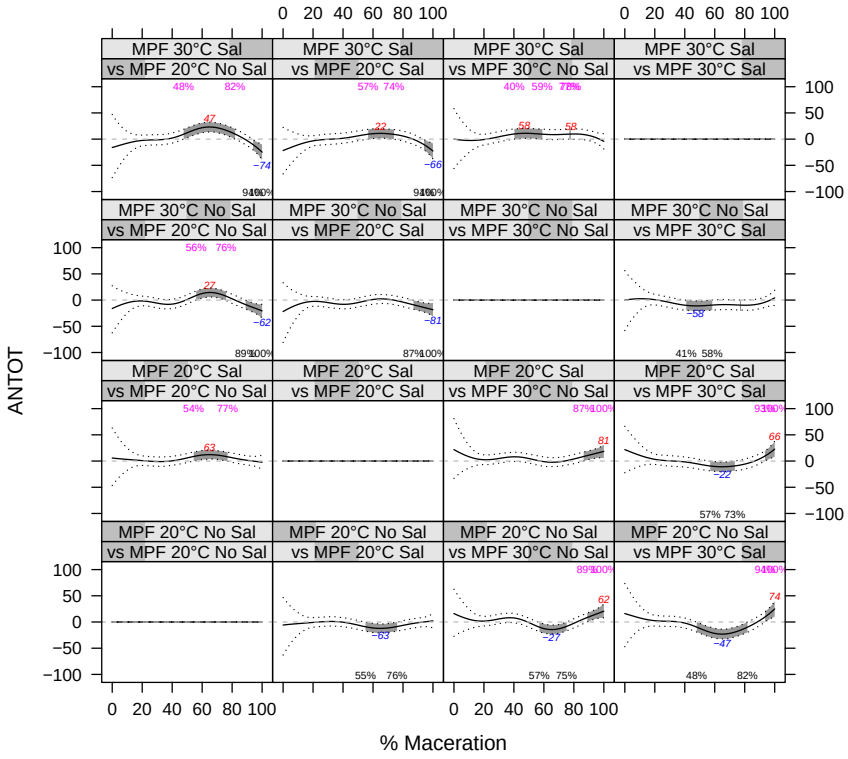

Figura 83: Contrasti per ANTOT (2009): mpf contro mpf

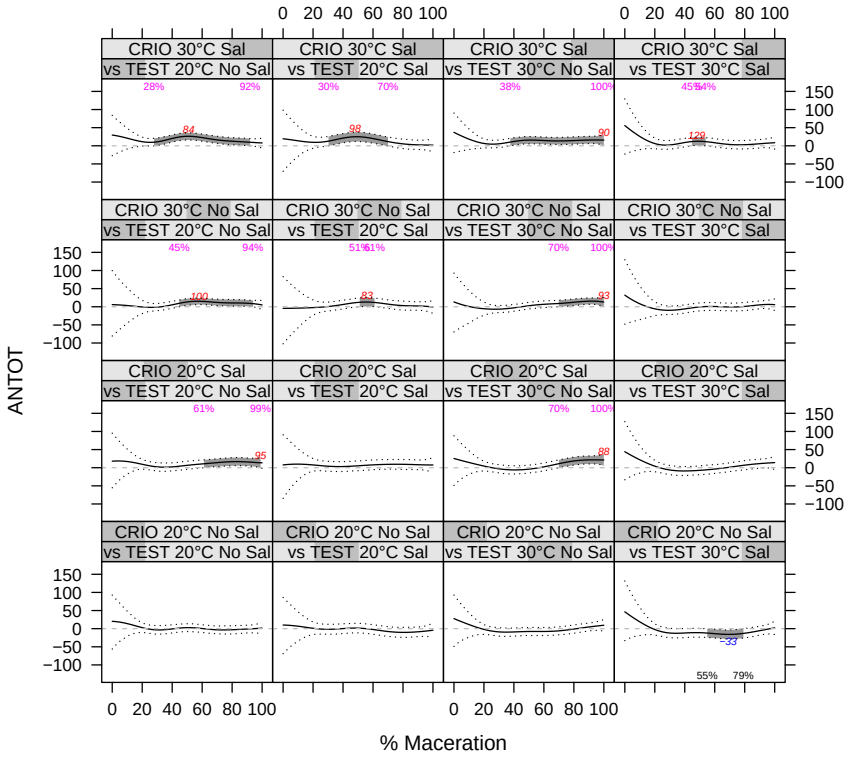

Figura 84: Contrasti per ANTOT (2009): crio contro test

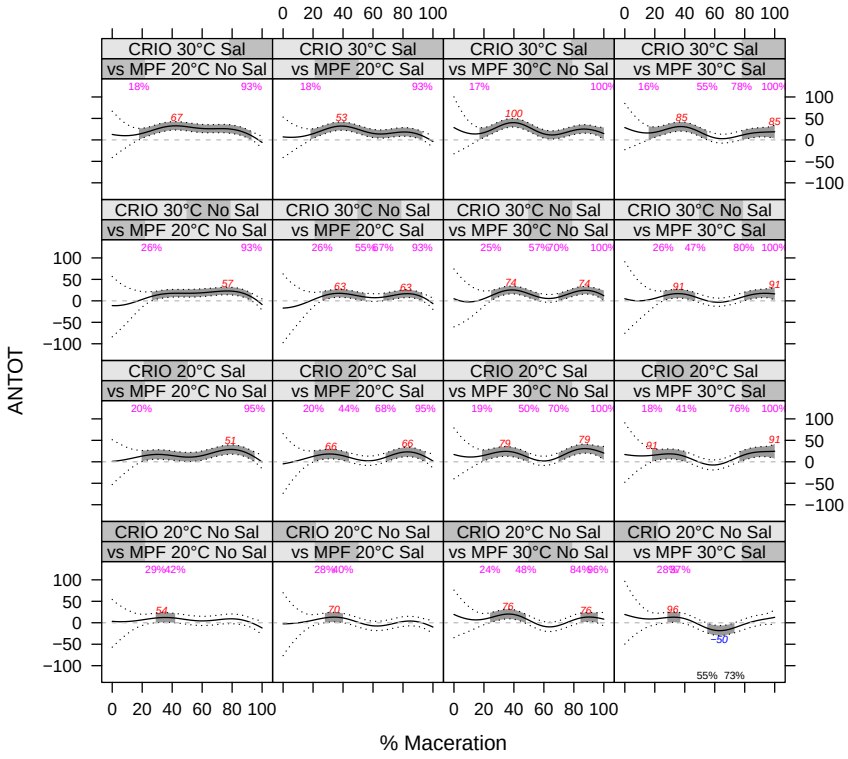

Figura 85: Contrasti per ANTOT (2009): crio contro mpf

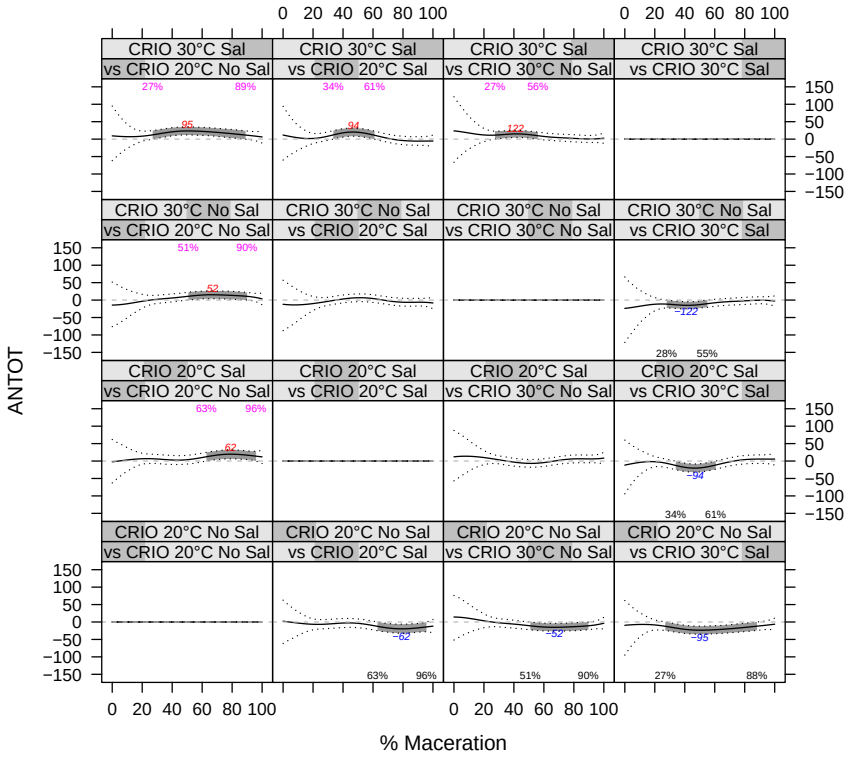

Figura 86: Contrasti per ANTOT (2009): crio contro crio

5 Prospettive

I modelli per le cinetiche della linea B.4 hanno svelato interessanti caratteristiche attribuibili ai trattamenti applicati. La maggior difficoltà tecnica è legata all'elevato numero di parametri richiesti per ottenere modelli adeguati in relazione alle poche osservazioni per vasca, tra le 10 e le 20 totali.

I modelli risultanti per il 2008 e 2009 non sono direttamente relati ed è tuttora in corso un approfondimento tecnico in laboratorio per valutare eventuali effetti di break strutturale imputabili alla diluizione oppure ad altre condizioni di misura.

In attesa del completamento delle analisi con i dati del 2010 e dell'estensione dei modelli sviluppati, è comunque possibile usare i modelli presentati per anno ed interpretare operativamente i risultati entro anno.

I risultati contenuti in questa relazione sono oggetto di corrente discussione critica congiuntamente con gli specialisti di Tuscania.

6 Bibliografia

Alcuni riferimenti bibliografici sono elencati come segue:

Barnet V., 1982, Comparative Statistical Inference, Wiley, New York.

Bengtsson,H., The R.oo package - Object-Oriented Programming with References Using Standard R Code. In Kurt Hornik, Friedrich Leisch and Achim Zeileis, editors, Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20-22, Vienna, Austria.
http://www.ci.tuwien.ac.at/Conferences/DSC-2003/

Berger J.O., 1985, Statistical Decision Theory and Bayesian Analysis, Springer-Verlag, Berlin.

Bernardo J.M., Smith A.F.M., 1994, Bayesian Theory, Wiley, New York.

Besag J., Green P., Higdon D., Mengersen K., 1995, Bayesian computation and stochastic systems, Statistical Science, 10(1):3-66.

Box G.E.P., Cox D.R., 1964, The analysis of transformations (with discussion), Journal of the Royal Statistical Society, B26:211-252.

Box G.E.P., Tiao G.C., 1973, Bayesian Inference in Statistical Analysis, Addison-Wesley, Reading.

Brockwell P.J., Davies R.A.,1998, Time Series: theory and Methods, Springer Verlag, New York.

Brumback, B.A., Ruppert, D. and Wand, M.P.,1999, Comment on paper by Shively, Kohn and Wood. Journal of the American Statistical Association, 94, 794–797.

Casella G., George E.I., 1992, Explaining the Gibbs sampler, The American Statistician, 46(3):167-174.

Casella G., Berger R.L., 1990, Statistical Inference, Duxbury, Wadsworth,Belmont.

Chib S., Greenberg E., 1995, Understanding the Metropolis-Hastings algorithm, The American Statistician, 49(4):327-335.

Cowell R.G., Dawid A.P., Lauritzen S.L., Spiegelhalter D.J., 1999, Probabilistic Networks and Expert Systems, Springer Verlag.

Cowles M.K., Carlin B.P., 1996, Markov Chain Monte Carlo convergence diagnostics: a comparative review, Journal of the American Statistical Association, 91(434):883-904.

De Groot M.H., 1970, Optimal statistical decisions, McGraw Hill.

Gelman A., Rubin D.B., 1992a, Inference from iterative simulation using multiple sequences, Statistical Sciences, 7(4):457-511.

Gelman A., Rubin D.B., 1992b, A single series from the Gibbs sampler provide a false sense of security, in Bernardo J.M., Berger J.O., Dawid A.P., Smith F.M. (eds.), Bayesian Statistics 4, Oxford University Press.

Gelman A., Carlin J.B., Stern H.S., Rubin D.B., 1995, Bayesian Data Analysis, Chapman and Hall, New York.

Gelman A., 1996, Inference and monitoring convergence, in Gilks W.R., Richardson S., (eds.), Markov Chain Monte Carlo in Practice, Chapman and Hall, New York.

Geyer C.J., 1992, Practical Markov chain Monte Carlo, Statistical Science, 7(4):473-511.

Gilks W.R., Richardson S., Spiegelhalter D.J., 1996, Markov Chain Monte Carlo in Practice, Chapman and Hall, New York.

Hammersley J.M., Handscomb D.C., 1964, Monte Carlo Methods, Chapman and Hall, New York.

Hastings W.K., 1970, Monte Carlo sampling methods using Markov chains and their applications, Biometrika, 57(1):97-109.

Lindley D., 1987, The probability approach to the treatment of uncertainty in artificial intelligence and expert systems, Statistical Science, 2(1):3-44.

Lindley D., 2000, The phylosophy of statistics, The Statistician, 49:293-337.

Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller M.N., Teller E., 1953, Equations of state calculations by fast computing machines, J. Chem. Phys., 21:1087-1092.

Mood A.M., Graybill F.A., Boes D.C., 1974, Introduction to the Theory of Statistics, McGraw-Hill, New York.

O'Hagan A., 1994, Bayesian Inference, Kendall's Advanced Theory of Statistics, Edward Arnold, London.

Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T., 1990, Numerical Recipes in C, Cambridge University Press, New York.

R Development Core Team, 2006, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ripley B.D., 1987, Stochastic Simulation, Wiley, New York.

Rosini E., 1988, Introduzione all'agroclimatologia, volume 1,2,3. ERSA-SMR, Bologna.

Smith A.F., Gelfand A.E., 1992, Bayesian statistics without tears: a sampling-resampling perspective, The American Statistician, 46(2):84-88.

Sneyers, R., 1990, On the statistical analysis of series of observations.WMO, Technical Note No. 143, Geneva, Switzerland.

Stefanini, F.M., 2006, Sintesi probabilistica di serie agroclimatologiche: le temperature, Technical Report.

Tanner M.A., 1993, Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, Springer-Verlag, New York.

Thomas D.C., Gauderman W.J., 1996, Gibbs sampling methods in genetics, in Gilks W.R., Richardson S., (eds.), Markov Chain Monte Carlo in Practice, Chapman and Hall, New York.

Thompson E.A., 1994, Monte Carlo likelihood in genetic mapping, Statistical Science, 9(3):355-366.

Tierney L., 1994, Markov Chains for exploring posterior distributions, Annals of Statistics, 22:1701-1762.

Venables W.N., Ripley B.D., 1994, Modern Applied Statistics with S-Plus, Springer-Verlag, New York.

Zhao, Y., Staudenmayer, J., Coull, B.A. and Wand, M.P. (2006). General Design Bayesian Generalized Linear Mixed Models. Statistical Science, (2006), 21, 35-51.