

UNIVERSITÀ DEGLI STUDI DI PADOVA  
FACOLTÀ DI INGEGNERIA

DIPARTIMENTO DI PRINCIPI E IMPIANTI DI INGEGNERIA CHIMICA  
“I. Sorgato”



TESI DI LAUREA IN INGEGNERIA CHIMICA  
CLASSE 10 INGEGNERIA INDUSTRIALE  
(DM 509/99)

UTILIZZO DI UN PIANO DI MISURE COMPOSITO CENTRALE PER  
L'OTTIMIZZAZIONE DEL PROCESSO DI PRODUZIONE DI UNA PASTA  
PIGMENTO IN DISPERSIONE ACQUOSA

*Relatore: ing. Paolo Canu*  
*Correlatore: dott. Daniele Foletto*

*Laureando: JACOPO FRAMARIN*

Anno Accademico 2007-2008



# Riassunto

In questa tesi è stata utilizzata una procedura, basata su un approccio con tecniche DOE (Design of experiments), al fine di studiare il comportamento di resa coloristica e viscosità, nella produzione di una pasta pigmento in dispersione acquosa. Sono stati allestiti due piani sperimentali, uno da realizzare in laboratorio e l'altro in impianto pilota, in cui sono stati fatti variare due parametri macchina su tre livelli: velocità dell'agitatore e tempo di permanenza.

Dall'esperienza pregressa è stato possibile utilizzare punti del piano che permettessero il confronto tra superfici di risposta sia su scala laboratorio che su scala industriale. Dopo aver acquisito i dati sperimentali relativi alle risposte scelte come oggetto di indagine per ognuno dei punti sperimentali, ci si è serviti delle informazioni raccolte, per individuare i modelli matematici che descrivessero l'andamento delle risposte stesse sul dominio del piano sperimentale.

Per il controllo dei dati ottenuti si sono sfruttati tre punti aggiuntivi (*check point*) in grado di verificare la bontà dei modelli adottati.

Infine, si sono confrontate le superfici di risposta per verificare l'effettiva applicabilità della metodologia empirica di comparazione attualmente utilizzata, tra le variabili operative di laboratorio e di produzione.

I risultati ottenuti sono stati validati e quindi confrontati mediante l'allestimento di una rete neurale.

Il lavoro descritto in questa Tesi è stato svolto durante un tirocinio di 3 mesi svolto presso la sede di Arzignano (VI) della ditta SAMIA s.a.s. Si ringrazia pertanto l'azienda SAMIA s.a.s per l'opportunità concessa, nonché tutto il personale, e in particolare modo il dott. Daniele Foletto, per la disponibilità mostrata.



# Indice

<b>INTRODUZIONE</b> .....	1
<b>CAPITOLO 1 – Presentazione dell’azienda</b> .....	3
1.1 NOTA INTRODUTTIVA.....	3
1.2 PRESENTAZIONE SAMIA.....	4
1.3 DESCRIZIONE DEL PROCESSO SAMIA s.a.s .....	5
1.3.1 Bagnatura e dispersione (premiscelazione).....	6
1.3.2 Macinazione (raffinazione).....	6
1.3.3 Completamento (Aggiustamento della forza tintoriale).....	6
<b>CAPITOLO 2 – La progettazione degli esperimenti</b> .....	7
2.1 STRATEGIA DELLA SPERIMENTAZIONE.....	7
2.2 INTRODUZIONE AI PIANI FATTORIALI.....	8
2.2.1 Definizioni di basi e principi.....	8
2.3 IL PIANO FATTORIALE A DUE FATTORI.....	9
2.3.1 Analisi statistica del modello a effetti fissi.....	11
2.3.2 Controllo dell’adeguatezza del modello.....	12
2.3.2.1 L’assunzione di normalità.....	15
2.3.2.2 Grafico dei residui in sequenza temporale.....	16
2.3.2.3 Grafico dei residui rispetto ai valori previsti.....	16
2.3.2.4 Grafico dei residui rispetto ad altre variabili.....	17
2.3.3 Scelta della dimensione campionaria.....	17
2.3.4 Un osservazione per cella.....	18
2.4 IL PIANO COMPOSITO CENTRALE.....	18
2.5 ACCOSTAMENTO DI MODELLI PER REGRESSIONE.....	19
2.5.1 Il metodo dei minimi quadrati ordinari.....	20
2.5.2 I parametri di valutazione dei modelli di regressione.....	24
2.5.3 Diagnostica dei modelli di regressione.....	27
2.5.3.1 Residui ridotti.....	27
2.5.3.2 Diagnostica di influenza.....	29
2.5.4 Le trasformazioni della risposta: il metodo di Box Cox.....	31
2.6 RISPOSTE MULTIPLE: LE FUNZIONI DI DESIDERABILITA’ .....	32
<b>CAPITOLO 3 - Parte sperimentale</b> .....	35
3.1 DESCRIZIONE DELL’ APPARECCHIATURA.....	35
3.1.1 Impianto laboratorio....	35
3.1.2 Impianto produzione.....	36
3.2 PROGETTAZIONE DEL PIANO SPERIMENTALE.....	38
3.3 ESECUZIONE DELLE PROVE: RACCOLTA DATI E RISPOSTE.....	39

3.4	VALUTAZIONE DEL PIANO SPERIMENTALE.....	40
3.5	ANALISI DELLE RISPOSTE.....	44
3.5.1	Resa laboratorio.....	44
3.5.2	Resa produzione.....	53
3.5.3	Viscosità laboratorio.....	57
3.5.4	Viscosità produzione.....	62
3.6	CONFRONTO FRA LE SUPERFICI DI RISPOSTA.....	66
3.7	CHECK-POINT.....	67
3.8	VALUTAZIONE DEI RISULTATI CON RETI NEURALI.....	69
	<b>CONCLUSIONI.....</b>	<b>72</b>
	<b>RIFERIMENTI BIBLIOGRAFICI.....</b>	<b>74</b>

# Introduzione

Con il seguente lavoro si è utilizzato un disegno sperimentale CCD ( central composite design) per esplorare un dominio sperimentale, scelto in base alle conoscenze accumulate nei precedenti studi, che permettesse lo studio di una pasta pigmento in produzione e in laboratorio.

Il lavoro è strutturato in tre capitoli. Il primo capitolo presenta una breve descrizione del processo produttivo SAMIA s.a.s.. Il secondo capitolo tratta i fondamenti teorici della pianificazione sperimentale, la descrizione dei piani fattoriali, in particolare l'uso del piano a due fattori e tre livelli e l'accostamento di modelli. Il terzo capitolo descrive la parte sperimentale di raccolta dati e la relativa analisi.

Per la descrizione degli aspetti teorici si è fatto riferimento ai testi di Montgomery (2005), Todeschini (2003) e Montgomery, Ranger e Hubele (2004); l'elaborazione dei dati per la parte sperimentale è stata condotta utilizzando i software Design-Expert (Vers. 7.0) di Stat-Ease® e NeuroShell 2® di Ward System Group, Inc.® . L'attività è stata svolta nel reparto di produzione e nel laboratorio chimico SAMIA s.a.s., il cui organico è attualmente composto da nove persone: un responsabile del laboratorio facente funzione di responsabile controllo qualità, ricerca e sviluppo, cinque addetti al controllo qualità, due tecnici formulatori e un addetto alla redazione di documenti ad uso interno/esterno e alla corrispondenza ordinaria con clienti e fornitori.





# Capitolo 1

## Presentazione dell'azienda

### 1.1 Nota introduttiva

Il distretto vicentino della concia, situato nella Valle del Chiampo, è oggi il maggiore polo conciario europeo, potendo contare su 721 imprese (dati 2004 CCIAA vicenza), 1500 occupati e un fatturato (2004) di 2859,7 milioni di euro.

La lavorazione del cuoio si può dividere principalmente in due fasi: un primo processo a umido e un successivo processo a secco, assimilabile a un *coating*. Il processo a umido comprende tutte quelle operazioni che, a partire dal materiale grezzo iniziale, provvedono a impartire alla pelle la caratteristica di imputrescibilità e a realizzare una prima nobilitazione del materiale per la commercializzazione, conferendogli morbidezza, pienezza, elasticità e colore. Con il processo a secco, detto rifinitura, si arriva alla fase finale dei procedimenti di produzione del cuoio: lo scopo è di rendere le pelli utilizzabili e idonee all'uso a cui sono destinate. Attraverso trattamenti meccanici e l'applicazione di agenti filmogeni e non filmogeni si ottengono le seguenti caratteristiche, differenti a seconda del tipo di pelle finita:

- tonalità del colore desiderata in versione trasparente, coprente o con effetti;
- aspetto lucido oppure opaco, più o meno brillante;
- tatto di superficie secco, ceroso, untuoso o con mano frenante;
- eliminazione di difetti superficiali, lesioni e ugualizzazione di macchie;
- protezione contro l'azione dello sporco, umidità e prodotti chimici usati dai produttori di manufatti.

Nell'ultimo decennio, la concorrenza di paesi emergenti come Cina, Brasile e India, e l'indiscutibile impatto ambientale causato dalle lavorazioni effettuate hanno provocato un progressivo disimpegno da parte dei conciatori locali nel processo a umido e conseguentemente un potenziamento nell'attività di rifinitura.

## 1.2 Presentazione di SAMIA s.a.s.

SAMIA s.a.s. è nata ad Arzignano nel 1976 ed è un'azienda leader nella produzione di prodotti chimici per la rifinizione.

Attualmente l'azienda si articola:

- nella Sede di Arzignano (VI), oltre 14500 m<sup>2</sup> coperti, in cui sono concentrate le attività amministrative, commerciali, tecnico-produttive e di laboratorio;
- nel deposito di S. Croce sull'Arno (PI), attraverso il quale si sviluppa la distribuzione dei prodotti standard e l'assistenza dei clienti in loco;
- in una vasta rete commerciale di agenzie e depositi che permette a SAMIA s.a.s. di essere presente in tutto il mondo, in modo particolarmente capillare in Estremo Oriente (Cina, Thailandia, Korea del Sud, Giappone) e in Sud America (Brasile, Argentina, Uruguay, Perù).

Nel 1996 l'azienda ha conseguito la certificazione di qualità secondo la norma UNI EN ISO 9002:1994, dal 2002 ha un sistema integrato (qualità, ambiente, sicurezza) conforme alle Norme UNI EN ISO 9001: 2000; UNI EN ISO 14001 e alla specifica OHSAS 18001.

La gamma di prodotti (oltre seicento) comprende paste base pigmentate, coloranti, agenti filmogeni e ausiliari (modificatori di tatto, penetranti, reticolanti, etc.); ormai da due decenni il punto di forza di SAMIA s.a.s. è la produzione di dispersioni di pigmento ad uso conciario per cui è il maggiore produttore italiano.

L'attività produttiva è discontinua, a batch, suddivisa in reparti secondo la tipologia di prodotto, e consiste in processi comprendenti operazioni di miscelazione, emulsione, dispersione e raffinazione sia in fase acqua che in fase solvente.

La capacità produttiva, grazie a una serie di moderni impianti, tecnologicamente avanzati, è superiore alle 10000 tonn/anno di prodotto finito, e ciò le permette di competere in tutto il mondo con concorrenti di dimensioni ben maggiori quali le numerose multinazionali presenti nel settore.

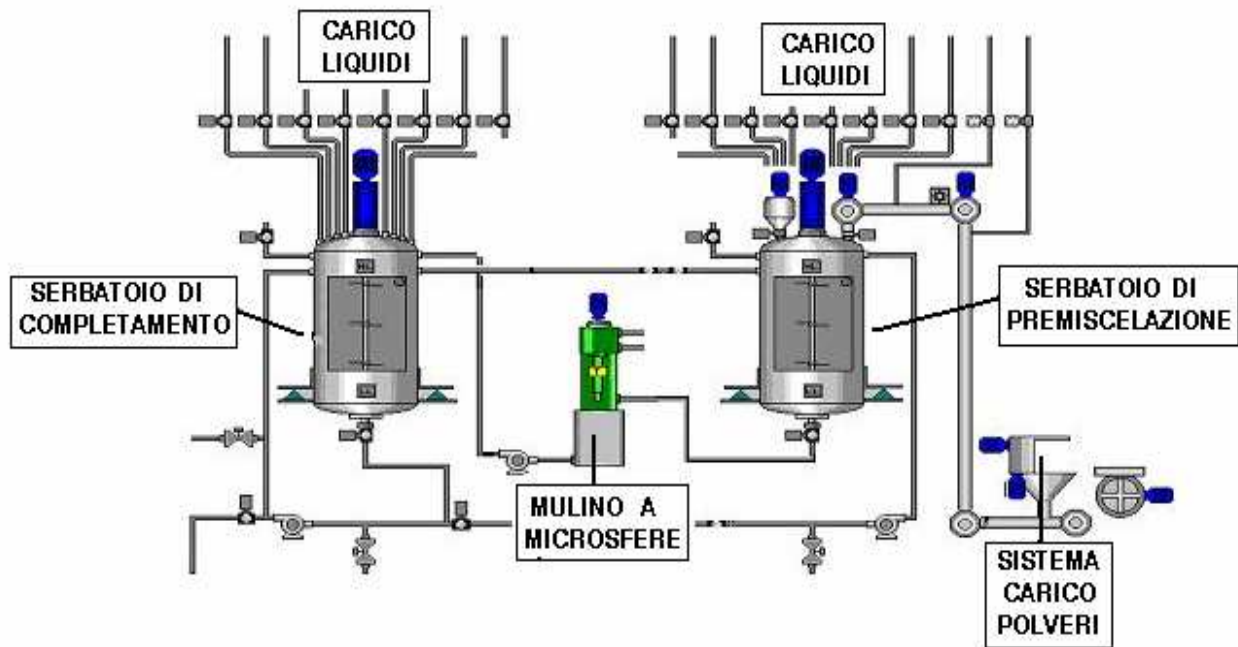
Il mercato in cui si trova a operare richiede la profusione di ingenti risorse nell'attività di assistenza post-vendita; pertanto l'azienda è dotata di un laboratorio interno attrezzato per le prove applicative e garantisce con personale qualificato un costante supporto tecnico presso il cliente. Tale assistenza comprende anche un reparto di campionatura colori in quanto il core-business, come già anticipato, è la produzione di dispersioni di pigmento. Infine il laboratorio chimico garantisce il controllo qualità su materie prime, semilavorati, prodotti finiti, oltrechè, attraverso la decennale esperienza dei suoi formulatori, lo sviluppo di prodotto e di processo.

### 1.3 Descrizione del processo SAMIA s.a.s.

L'attività principale di SAMIA s.a.s. è la produzione di dispersioni di pigmento da utilizzarsi nel settore conciario, in particolare dispersioni di pigmento in fase acquosa.

Le produzioni prevalenti sono relative a batch da oltre cinque tonnellate e vengono effettuate in impianti costituiti da (vedi figura 1.1):

- sistemi per il carico delle materie prime (sia solide che liquide)
- miscelatori equipaggiati con agitatori
- macchine raffinatrici
- sistema di supervisione, controllo e gestione



**Figura 1.1** Schema di impianto per la produzione di dispersioni di pigmento

Il processo produttivo è schematizzabile in tre fasi:

- Bagnatura e dispersione dei pigmenti (premiscelazione)
- Macinazione (raffinazione)
- Completamento (aggiustamento della forza tintoriale)

### *1.3.1 Bagnatura e dispersione (premiscelazione)*

Nella bagnatura l'aria presente sulla superficie del pigmento è sostituita con una fase liquida ovvero l'interfaccia, da solido/gas (pigmento/aria), diviene solido/liquido (pigmento/soluzione acquosa). Nella dispersione gli agglomerati di pigmento vengono rotti e subiscono una prima riduzione di dimensione attraverso forze di taglio generate con opportuni agitatori.

Per la bagnatura e la dispersione l'azione meccanica di agitazione viene supportata dall'utilizzo di specifici additivi chimici: gli additivi bagnanti accelerano la bagnatura mentre gli additivi disperdenti vengono impiegati sia per rendere possibile o agevolare la raffinazione, sia per stabilizzare il prodotto in corso di produzione, nel magazzinaggio e nell'applicazione.

Spesso lo stesso additivo riunisce azione bagnante e disperdente.

### *1.3.2 Macinazione (raffinazione)*

La macinazione permette l'ulteriore riduzione nelle dimensioni degli agglomerati di pigmento: la raffinazione varia da pigmento a pigmento e si ottiene tramite passaggi in mulini a microsferiche del premiscelato. In questa fase si raggiungono le caratteristiche chimico-fisiche che vengono richieste al prodotto finito, quali la lucentezza e l'intensità del colore. La viscosità e la reologia in genere sono fondamentali per utilizzare la macchina al massimo del suo potere raffinanante e ciò, unitamente al fatto che la macinazione è la fase più lenta dell'intero processo, comporta un costante monitoraggio nonché un preventivo lavoro di aggiustamento nella composizione del premiscelato nell'ottica di ridurre i tempi di lavorazione.

### *1.3.3 Completamento (aggiustamento della forza tintoriale)*

Una volta terminata la fase di raffinazione, la pasta concentrata viene "completata" ovvero vengono controllati e regolati i parametri chimico-fisici per garantire il rispetto del capitolato di ciascun prodotto.

Il parametro più importante, in quanto indispensabile al cliente per l'agevole utilizzo della dispersione di pigmento, è senza dubbio la forza tintoriale, la quale viene aggiustata nell'intervallo di accettabilità riportato nelle specifiche di vendita mediante aggiunta dei cosiddetti prodotti di completamento (nella maggior parte dei casi acqua e/o leganti e prodotti addensanti).

# Capitolo 2

## La progettazione degli esperimenti

### 2.1 Strategia della sperimentazione

Nell'ingegneria la sperimentazione gioca un ruolo importante nella progettazione di nuovi prodotti, nello sviluppo di processi produttivi e nel miglioramento del processo. L'obiettivo può essere sovente quello di sviluppare un processo robusto, vale a dire affetto il meno possibile da fonti di variabilità esterne. E' possibile comunemente rappresentare il processo come una combinazione di macchine, metodi, personale ed altre risorse che trasformano un certo input (spesso una materia prima), in un output caratterizzato da una o più risposte osservabili. Alcune delle variabili di processo  $x_1, x_2, \dots, x_p$  sono controllabili, mentre altre variabili  $z_1, z_2, \dots, z_q$ , non sono controllabili (sebbene esse possano esserlo a scopo di prova). Gli obiettivi dello sperimentatore possono includere:

- Determinare quali variabili hanno maggiore influenza sulla risposta  $y$
- Determinare quali valori assegnare alle variabili influenti  $x$ , in modo che la risposta  $y$  risulti quasi sempre prossima al valore nominale desiderato
- Determinare quali valori assegnare alle variabili influenti  $x$ , in modo che la variabilità nella risposta  $y$  sia piccola
- Determinare quali valori assegnare alle variabili influenti  $x$ , in modo che l'effetto delle variabili non controllabili  $z_1, z_2, \dots, z_q$  sulla risposta  $y$  sia minimizzato.

Gli esperimenti coinvolgono spesso diversi fattori (o variabili); uno degli obiettivi di chi conduce l'esperimento (detto lo **sperimentatore**) è spesso quello di determinare l'influenza che questi fattori hanno sulla risposta del sistema.

La **strategia sperimentale** è un approccio generale per pianificare e condurre un esperimento.

A disposizione dello sperimentatore ci sono diverse strategie sperimentali:

- L'**approccio a tentativi** si esegue quando gli sperimentatori conoscono molto bene il sistema che stanno studiando e per il quale si ha una notevole esperienza pratica. Consiste nel condurre le prove scegliendo una combinazione arbitraria di fattori e facendo variare uno o più livelli dei fattori, non a caso, ma guidati dalla conoscenze tecniche e teoriche del sistema.
- L'**approccio un fattore alla volta** si impiega, dopo aver scelto un insieme di livelli base per ciascun fattore, variando in successione i livelli di ogni fattore mantenendo gli altri fattori costanti al loro livello base. Il maggiore svantaggio della strategia un-fattore-alla-volta, è

che non riesce a tenere conto di possibili **interazioni** tra i fattori. Un'interazione consiste nel fatto che un fattore non causa lo stesso effetto sulla risposta, al variare dei livelli di un altro fattore. Le interazioni tra fattori sono spesso presenti ed in quei casi la strategia un-fattore-alla-volta darà risultati scadenti.

- L'**approccio fattoriale** segue una strategia secondo cui sono provate tutte le possibili combinazioni di fattori e livelli e per questo esso è considerato il modo corretto per condurre esperimenti con più fattori.

La pianificazione statistica degli esperimenti è necessaria se si vogliono ricavare conclusioni fondate su solide basi scientifiche.

## 2.2 Introduzione ai piani fattoriali

### 2.2.1 Definizioni di base e principi

Col termine **piano fattoriale** ci si riferisce ad esperimenti completi o repliche degli stessi, in cui sono provate tutte le possibili combinazioni di fattori e livelli. Per esempio, se ci sono  $a$  livelli del fattore  $A$  e  $b$  livelli del fattore  $B$ , ogni **replicazione** contiene tutte le  $a \cdot b$  combinazioni dei trattamenti, dove con il termine replicazione si intende la ripetizione dell'esperimento di base, cioè in questo caso del piano. Spesso si dice che i fattori inseriti in un piano fattoriale sono **incrociati**.

L'effetto di un fattore, definito come la variazione nella risposta prodotta da una variazione nel livello del fattore, viene chiamato **effetto principale**, poiché, di solito, ci si riferisce ai fattori primari d'interesse nell'esperimento. Per esempio, si consideri un semplice esperimento fattoriale a due fattori con entrambi i fattori del piano a due livelli. Si indichino questi livelli rispettivamente come "basso" e "alto". L'effetto principale del fattore  $A$ , in questo piano a due livelli, può essere pensato come la differenza tra la risposta media al livello basso di  $A$  e la risposta media al livello alto di  $A$ .

In alcuni esperimenti si può notare che la differenza nella risposta tra i livelli di un fattore non è la stessa per tutti i livelli degli altri fattori; quando ciò avviene esiste interazione tra i fattori. Il concetto in questione può essere illustrato graficamente, riportando le risposte in funzione di un fattore ( $A$ ), per entrambi i livelli dell'altro fattore ( $B$ ): se le linee ottenute sono approssimativamente parallele, ciò sta ad indicare l'assenza di interazione tra i fattori  $A$  e  $B$ . In caso contrario, se è evidente invece che le linee non sono parallele, esiste interazione tra i fattori. Non ci si deve tuttavia limitare ad un'analisi dei dati di questo tipo, poiché le interpretazioni sono soggettive e in taluni casi la presentazione grafica e le conseguenti deduzioni possono essere fuorvianti.

C'è un altro modo per illustrare il concetto di interazione. Si supponga che entrambi i fattori del piano siano **quantitativi**, cioè che i loro livelli possano essere messi in corrispondenza con i punti su una determinata scala numerica, come ad esempio accade per temperatura, pressione o tempo. Allora si può rappresentare il **modello di regressione** dell'esperimento fattoriale a due fattori come:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon, \quad (2.1)$$

dove  $y$  è la risposta,  $\beta$  sono parametri i cui valori devono essere determinati,  $x_1$  è la variabile che rappresenta il fattore  $A$ ,  $x_2$  è la variabile che rappresenta il fattore  $B$  ed  $\varepsilon$  è un termine di errore casuale. Le variabili  $x_1$  e  $x_2$  sono definite su una **scala codificata** da -1 a +1 (i livelli basso e alto di  $A$  e  $B$ ) e  $x_1 x_2$  rappresenta l'interazione tra  $x_1$  e  $x_2$ .

Le stime dei parametri in questo modello di regressione risultano legate alle stime degli effetti dei fattori. Dalla rappresentazione grafica di questo modello si ottiene un diagramma tridimensionale chiamato **grafico della superficie di risposta**, in cui si ritrova la superficie dei valori di  $y$  generata dalle varie combinazioni di  $x_1$  e  $x_2$ . La presenza di interazione sostanziale altera il piano della superficie di risposta, cioè inserisce una forma di curvatura.

In genere, quando un'interazione è grande, i corrispondenti effetti principali hanno un significato pratico limitato. Un'interazione significativa spesso maschera la significatività degli effetti principali: può accadere che il fattore  $A$  abbia un effetto, ma che esso dipenda dal livello del fattore  $B$ ; in altri termini, la conoscenza dell'interazione  $AB$  può essere più utile della conoscenza dell'effetto principale. In presenza di interazioni significative, lo sperimentatore deve di solito esaminare i livelli di un fattore, per esempio  $A$ , in relazione coi livelli degli altri fattori, per trarre conclusioni sull'effetto principale di  $A$ .

In sintesi si può dunque concludere che i piani fattoriali hanno numerosi vantaggi: essi sono più efficienti degli esperimenti ad un fattore alla volta e sono inoltre indispensabili in presenza di interazioni, per evitare conclusioni fuorvianti; infine i piani fattoriali consentono di stimare gli effetti d'ogni fattore a differenti livelli degli altri, consentendo conclusioni valide in un'ampia gamma di condizioni sperimentali.

## 2.3 Il piano fattoriale a due fattori

I più semplici tipi di piani fattoriali prevedono solo due fattori o insiemi di trattamenti. Sia  $y_{ijk}$  la risposta osservata quando il fattore  $A$  è all' $i$ -esimo livello ( $i = 1, 2, \dots, a$ ) ed il fattore  $B$  è al  $j$ -esimo livello ( $j = 1, 2, \dots, b$ ) per la  $k$ -esima replicazione ( $k = 1, 2, \dots, n$ ). In generale un esperimento a due fattori apparirà come in Tabella 1.1. L'ordine in cui le  $a \cdot b \cdot n$  osservazioni sono rilevate è casuale; si tratta di un **piano completamente casualizzato**. Con il termine casualizzazione s'intende che, sia l'ordine in cui le singole prove dell'esperimento sono condotte, sia l'allocazione del materiale sperimentale, sono stabiliti in maniera casuale.

Casualizzando, infatti, si mediano gli effetti di fattori estranei che possono essere presenti e si fa in modo che gli errori siano variabili casuali indipendenti.

**Tabella 2.1** *Disposizione generale di un piano fattoriale a due fattori*

		Fattore B			
		1	2	...	b
Fattore A	1	$y_{111}, y_{112},$ $\dots, y_{11n}$	$y_{121}, y_{122},$ $\dots, y_{12n}$		$y_{1b1}, y_{1b2},$ $\dots, y_{1bn}$
	2	$y_{211}, y_{212},$ $\dots, y_{21n}$	$y_{221}, y_{222},$ $\dots, y_{22n}$		$y_{2b1}, y_{2b2},$ $\dots, y_{2bn}$
	a	$y_{a11}, y_{a12},$ $\dots, y_{a1n}$	$y_{a21}, y_{a22},$ $\dots, y_{a2n}$		$y_{ab1}, y_{ab2},$ $\dots, y_{abn}$

Le osservazioni in un esperimento fattoriale possono essere descritte da un modello, che si può scrivere in diversi modi. Un possibile modello è il **modello degli effetti**:

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \quad \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, b \\ k = 1, 2, \dots, n \end{cases} \quad (2.2)$$

dove  $\mu$  è l'effetto medio generale,  $\tau_i$  è l'effetto dell' $i$ -esimo livello del fattore di riga A,  $\beta_j$  è l'effetto del  $j$ -esimo livello del fattore di colonna B,  $(\tau\beta)_{ij}$  è l'effetto dell'interazione tra  $\tau_i$  e  $\beta_j$  e  $\varepsilon_{ijk}$  è una componente di errore casuale. Entrambi i fattori sono assunti come **fissi**, cioè specificatamente scelti dallo sperimentatore (non elementi di un campione casuale), e gli effetti dei trattamenti sono definiti come scarti dalla media generale; quindi  $\sum_{i=1}^a \tau_i = 0$  e  $\sum_{j=1}^b \beta_j = 0$ .

Analogamente, gli effetti di interazione sono fissi e sono definiti in modo che  $\sum_{i=1}^a (\tau\beta)_{ij} = \sum_{j=1}^b (\tau\beta)_{ij} = 0$ . In questo caso in cui si considerano effetti fissi, si vogliono verificare le ipotesi di uguaglianza tra le medie di popolazione dei trattamenti e le conclusioni cui si perviene saranno valide solo per i livelli dei fattori considerati nell'analisi. Le stesse conclusioni non possono essere estese a trattamenti simili che non sono stati esplicitamente considerati. Poiché ci sono  $n$  replicazioni dell'esperimento, ci sono  $a \cdot b \cdot n$  osservazioni totali.

Un altro possibile modello per l'esperimento fattoriale è il **modello delle medie**:



$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \quad \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, b \\ k = 1, 2, \dots, n \end{cases} \quad (2.3)$$

dove la media dell' $ij$ -esima cella è:

$$\mu_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij}.$$

Si potrebbe usare anche, come in § 2.2.1, un **modello di regressione**, particolarmente utile quando uno o più fattori sono quantitativi.

Nel fattoriale a due fattori, i due fattori (o trattamenti) di riga e di colonna,  $A$  e  $B$ , sono d'eguale interesse. Specificatamente, l'interesse è rivolto a **valutare ipotesi** sull'eguaglianza di effetti di trattamenti di riga, del tipo:

$$\begin{aligned} H_0 &: \tau_1 = \tau_2 \dots = \tau_a = 0 \\ H_1 &: \text{almeno un } \tau_i \neq 0 \end{aligned} \quad (2.4)$$

e l'eguaglianza di effetti di trattamento di colonna, del tipo:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 \dots = \beta_b = 0 \\ H_1 &: \text{almeno un } \beta_j \neq 0. \end{aligned} \quad (2.5)$$

Vi è anche interesse a determinare se i trattamenti di riga e di colonna *interagiscono*. Quindi si desidera valutare anche:

$$\begin{aligned} H_0 &: (\tau\beta)_{ij} = 0 \text{ per tutti } i, j \\ H_1 &: \text{almeno un } (\tau\beta)_{ij} \neq 0. \end{aligned} \quad (2.6)$$

Queste ipotesi vengono verificate usando l'**analisi della varianza a due fattori**.

### 2.3.1 Analisi statistica del modello a effetti fissi

Si indichi con  $y_{i..}$  il totale di tutte le osservazioni effettuate col livello  $i$ -esimo del fattore  $A$ , con  $y_{.j}$  il totale di tutte le osservazioni col  $j$ -esimo livello del fattore  $B$ ,  $y_{ij}$  il totale di tutte le osservazioni nella cella  $ij$ -esima e  $y_{...}$  il totale di tutte le osservazioni. Si definiscano  $\bar{y}_{i..}$ ,  $\bar{y}_{.j}$ ,  $\bar{y}_{ij}$  e  $\bar{y}_{...}$ , come le corrispondenti medie di riga, di colonna, di cella e generale. Espresso matematicamente:

$$Y_{i..} = \sum_{j=1}^b \sum_{k=1}^n Y_{ijk} \quad \bar{Y}_{i..} = \frac{Y_{i..}}{bn} \quad i = 1, 2, \dots, a$$

$$Y_{.j} = \sum_{i=1}^a \sum_{k=1}^n Y_{ijk} \quad \bar{Y}_{.j} = \frac{Y_{.j}}{an} \quad j = 1, 2, \dots, b \quad (2.7)$$

$$Y_{ij} = \sum_{k=1}^n Y_{ijk} \quad \bar{Y}_{ij} = \frac{Y_{ij}}{n} \quad i = 1, 2, \dots, a; \quad j = 1, 2, \dots, b$$

$$Y_{...} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n Y_{ijk} \quad \bar{Y}_{...} = \frac{Y_{...}}{abn}$$

La somma dei quadrati totale corretta è l'indice statistico utilizzato come misura della variabilità totale dei dati. Intuitivamente ciò è ragionevole poiché, se la si dividesse per il numero totale dei gradi di libertà appropriato, si otterrebbe la **varianza campionaria** delle  $Y$ . La varianza campionaria è l'indice comunemente adottato per misurare la variabilità. Essa può essere scritta come:

$$\begin{aligned} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{...})^2 &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n [(\bar{Y}_{i...} - \bar{Y}_{...}) + (\bar{Y}_{.j} - \bar{Y}_{...}) + (\bar{Y}_{ij} - \bar{Y}_{i...} - \bar{Y}_{.j} + \bar{Y}_{...}) + (Y_{ijk} - \bar{Y}_{ij})]^2 = \\ &= bn \sum_{i=1}^a (\bar{Y}_{i...} - \bar{Y}_{...})^2 + an \sum_{j=1}^b (\bar{Y}_{.j} - \bar{Y}_{...})^2 + n \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{ij} - \bar{Y}_{i...} - \bar{Y}_{.j} + \bar{Y}_{...})^2 + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{ij})^2, \end{aligned} \quad (2.8)$$

poiché i sei prodotti incrociati del lato destro hanno valore zero. Si noti che la somma dei quadrati totale è stata scomposta in una somma di quadrati  $SS_A$  dovuta a "righe" o fattore  $A$ , una somma dei quadrati  $SS_B$  dovuta a "colonne" o fattore  $B$ , una somma dei quadrati  $SS_{AB}$  dovuta all'interazione tra  $A$  e  $B$  ed una somma dei quadrati dovuta all'errore  $SS_E$ . Dall'espressione dell'ultima componente della (1.7), si deduce che ci devono essere almeno due replicazioni per ottenere una somma di quadrati dell'errore. La (1.7) può essere scritta in simboli come:

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E. \quad (2.9)$$

Essendoci  $abn$  osservazioni in totale,  $SS_T$  ha pertanto  $abn - 1$  gradi di libertà, dove i gradi di libertà corrispondono al numero di elementi indipendenti presenti in quella somma di quadrati. Allo stesso modo poiché ci sono  $a$  livelli del fattore  $A$  e  $b$  livelli del fattore  $B$ , gli effetti principali di  $A$  e  $B$  hanno  $a - 1$  e  $b - 1$  gradi di libertà. I gradi di libertà dell'interazione sono semplicemente i gradi di libertà delle celle  $(ab - 1)$  meno il numero di gradi di libertà dei due effetti principali  $A$  e  $B$ ; cioè  $ab - 1 - (a - 1) - (b - 1) = (a - 1)(b - 1)$ . All'interno di ciascuna delle  $a \cdot b$  caselle ci sono  $n - 1$  gradi di libertà tra le  $n$  replicazioni; quindi ci sono  $ab(n - 1)$  gradi di libertà per l'errore. Si noti che i gradi di libertà sul lato destro della (1.8) assommano al numero totale di gradi di libertà.

Ciascuna somma di quadrati, divisa per i propri gradi di libertà, è un **quadrato medio** (MS). Esaminando i **valori attesi**, o valori a lungo andare, di tali quadrati medi, si osserva che:

$$E(MS_A) = E\left(\frac{SS_A}{a-1}\right) = \sigma^2 + \frac{bn \sum_{i=1}^a \tau_i^2}{a-1}$$

$$E(MS_B) = E\left(\frac{SS_B}{b-1}\right) = \sigma^2 + \frac{an \sum_{j=1}^b \beta_j^2}{b-1}$$

$$E(MS_{AB}) = E\left(\frac{SS_{AB}}{(a-1)(b-1)}\right) = \sigma^2 + \frac{n \sum_{i=1}^a \sum_{j=1}^b (\tau\beta)_{ij}^2}{(a-1)(b-1)}$$

$$E(MS_E) = E\left(\frac{SS_E}{ab(n-1)}\right) = \sigma^2.$$
(2.10)

Pertanto  $MS_E$  è uno **stimatore** congiunto della varianza campionaria, cioè una funzione delle osservazioni campionarie, che non contiene parametri incogniti, corrispondente a quel parametro. Si noti che, se le ipotesi nulle di assenza di effetti di trattamento di riga, di colonna e di interazione sono vere, allora i valori attesi dei quadrati medi  $MS_A$ ,  $MS_B$ ,  $MS_{AB}$  e  $MS_E$  sono tutte stime della varianza della popolazione  $\sigma^2$ ; tuttavia se vi sono differenze per esempio tra gli effetti di trattamento di riga, allora il valore atteso di  $MS_A$  sarà maggiore del valore atteso di  $MS_E$ . Analogamente, se sono presenti effetti del trattamento di colonna, o d'interazione, allora i valori attesi dei corrispondenti quadrati medi saranno maggiori del valore atteso di  $MS_E$ . Pertanto, per valutare la significatività di entrambi i fattori e della loro interazione, è sufficiente dividere il corrispondente quadrato medio per il quadrato medio dell'errore. Valori elevati di tale rapporto stanno ad indicare che i dati non confermano l'ipotesi nulla.

Se si assume che il modello (1.1) sia adeguato e che i termini di errore  $\varepsilon_{ijk}$  siano distribuiti normalmente e indipendentemente con varianza costante  $\sigma^2$ , allora ognuno dei rapporti di quadrati medi  $MS_A/MS_E$ ,  $MS_B/MS_E$  e  $MS_{AB}/MS_E$  è distribuito come una variabile casuale  $F$  rispettivamente con  $a-1$ ,  $b-1$  e  $(a-1)(b-1)$  gradi di libertà al numeratore e  $ab(n-1)$  gradi di libertà al denominatore e la regione critica è la coda destra della distribuzione  $F$  di Fisher. La procedura del test di solito è riassunta in una tabella d'analisi della varianza, come mostrato in Tabella 2.2

**Tabella 2.2** Tabella d'analisi della varianza per il fattoriale a due fattori, modello a effetti fissi.

Origine della variabilità	Somma dei quadrati	Gradi di libertà	Quadrati medi	$F_0$
A trattamenti	$SS_A$	$a - 1$	$MS_A = \frac{SS_A}{a - 1}$	$F_0 = \frac{MS_A}{MS_E}$
B trattamenti	$SS_B$	$b - 1$	$MS_B = \frac{SS_B}{b - 1}$	$F_0 = \frac{MS_B}{MS_E}$
Interazione	$SS_{AB}$	$(a - 1)(b - 1)$	$MS_{AB} = \frac{SS_{AB}}{(a - 1)(b - 1)}$	$F_0 = \frac{MS_{AB}}{MS_E}$
Errore	$SS_E$	$ab(n - 1)$	$MS_E = \frac{SS_E}{ab(n - 1)}$	
Totale	$SS_T$	$abn - 1$		

Per i calcoli relativi all'analisi della varianza, si ricorre all'utilizzo di software statistici o a fogli elettronici tipo excel.

### 2.3.2 Controllo d'adeguatezza del modello

Prima di accettare le conclusioni dell'analisi della varianza, si dovrebbe controllare l'adeguatezza del modello adottato, ossia verificare che siano soddisfatte alcune assunzioni: le osservazioni devono essere adeguatamente descritte dal modello e gli errori devono essere distribuiti normalmente e indipendentemente con valore atteso zero e varianza  $\sigma^2$  costante anche se ignota. Se queste assunzioni sono soddisfatte la procedura di analisi della varianza è un test esatto per verificare le ipotesi suddette. Lo strumento diagnostico principale è costituito dall'**analisi dei residui**. I residui del modello fattoriale a due fattori sono definiti come:

$$e_{ijk} = y_{ijk} - \hat{y}_{ijk} \quad (2.11)$$

dove  $\hat{y}_{ijk}$  è il **valore previsto** della singola osservazione. Poiché  $\hat{y}_{ijk} = \bar{y}_{ij}$  (le medie aritmetiche delle osservazioni nella  $ij$ -esima cella), la (1.9) diventa:

$$e_{ijk} = y_{ijk} - \bar{y}_{ij} \quad (2.12)$$

L'esame dei residui dovrebbe essere sempre eseguito per qualunque analisi della varianza. Il controllo diagnostico del modello può essere eseguito facilmente attraverso un'analisi grafica dei

residui. Se il modello è adeguato, i residui dovrebbero essere **privi di struttura**, cioè non dovrebbero mostrare alcun andamento sistematico evidente.

### 2.3.2.1 L'assunzione di normalità

Tale controllo potrebbe essere effettuato rappresentando i residui con un istogramma. Se l'assunzione è soddisfatta, questo grafico dovrebbe essere simile a quello che si avrebbe per un campione proveniente da una distribuzione normale centrata sullo zero. Sfortunatamente, spesso si verificano considerevoli fluttuazioni in presenza di campioni di dimensioni ridotte, pertanto l'apparente moderato allontanamento dalla normalità non è detto che implichi necessariamente gravi violazioni delle assunzioni.

E' invece di estrema utilità il tracciamento di un **grafico di probabilità normale** dei residui, in cui i residui sono riportati in ordine crescente rispetto alla loro frequenza cumulata osservata; se la distribuzione dell'errore è normale, il tracciato somiglierà ad un segmento di linea retta. Nel rappresentarlo conviene dare maggior enfasi ai valori centrali del grafico, piuttosto che agli estremi. In generale, modesti scostamenti dalla normalità sono di scarsa rilevanza nell'analisi della varianza ad effetti fissi. Poiché il test  $F$  ne è solo leggermente influenzato, se ne deduce che l'analisi della varianza è **robusta** per l'assunzione di normalità. Gli scostamenti dalla normalità, di solito, fanno sì che sia i veri livelli di significatività sia la potenza differiscano leggermente dai valori nominali e che la potenza del test si abbassi.

Un'anomalia molto comune, spesso riscontrata sui grafici di probabilità normale, è la presenza di un residuo in valore assoluto molto più grande degli altri; tale residuo abnorme è indicato come **valore anomalo** (*outlier*). La presenza di uno o più *outlier* può seriamente portare a distorsioni sui risultati dell'analisi della varianza e soprattutto nella formulazione del modello. A volte la presenza di *outlier* è dovuta ad errori di calcolo nella codifica dei dati o nella trascrizione. Se non è questo il caso, si devono analizzare attentamente le condizioni sperimentali corrispondenti alla prova in questione; se la risposta abnorme è un valore particolarmente desiderabile, l'*outlier* può essere più informativo del resto dei dati. Si deve avere molta cura a non rifiutare o scartare un'osservazione soltanto perché abnorme, a meno di non avere sostanziali motivi di natura non statistica per farlo. Si può altresì procedere a due analisi distinte, una comprende l'*outlier* e una senza. Un controllo grossolano può essere effettuato esaminando i **residui standardizzati**:

$$d_{ij} = \frac{e_{ij}}{\sqrt{MS_E}} . \quad (2.13)$$

Se gli errori  $\varepsilon_{ijk}$  sono normalmente e indipendentemente distribuiti, i residui standardizzati dovrebbero essere approssimativamente normali con valore atteso zero e varianza unitaria. Pertanto circa il 68 percento dei residui standardizzati dovrebbe cadere entro i limiti  $\pm 1$ , circa il 95 percento

dovrebbe cadere entro  $\pm 2$  e quasi tutti quanti dovrebbero cadere entro  $\pm 3$ . Un residuo maggiore di 3 o 4 deviazioni standard da zero è potenzialmente un *outlier*.

### 2.3.2.2 Grafico dei residui in sequenza temporale

Il tracciamento del grafico dei residui nell'ordine temporale della raccolta dei dati è utile per individuare una eventuale **correlazione** tra i residui. Una tendenza ad avere sequenze di residui positivi e alternate a sequenze di residui negativi indica una correlazione positiva, dal che se ne deduce che l'**assunzione di indipendenza** degli errori è stata violata. Questo è un problema potenzialmente grave, difficile da correggere, pertanto è di estrema importanza prevenirlo possibilmente all'atto della raccolta dati. Un'opportuna casualizzazione dell'esperimento è un importante passo per garantire l'indipendenza.

Talvolta il processo in esame può subire derive o fluttuazioni incontrollate; tutto ciò può dare luogo ad alterazioni della varianza dell'errore nel tempo. Questa condizione viene evidenziata nei grafici dei residui tracciati rispetto al tempo, i quali mostrano maggiore dispersione ad un capo rispetto all'altro. Una **varianza non costante** è un problema potenzialmente grave. Se l'assunzione di omogeneità della varianza è violata, il test  $F$  è distorto solo marginalmente nel modello a effetti fissi. L'approccio più comune nella trattazione di problemi con varianza non costante, qualora il problema si manifesti, è quello di applicare una trasformazione stabilizzatrice della varianza (vedi § 2.6.4) ed eseguire quindi l'analisi sui dati trasformati; le conclusioni tratte da tali analisi della varianza valgono per le popolazioni *trasformate*. Se la distribuzione teorica delle osservazioni è nota allo sperimentatore, tale informazione può essere sfruttata per la scelta della trasformazione. Nella pratica, molti sperimentatori scelgono la forma della trasformazione semplicemente provando alcune alternative ed osservando gli effetti di ciascuna trasformazione sul grafico dei residui rispetto alla risposta prevista (§ 2.3.2.3).

### 2.3.2.3 Grafico dei residui rispetto ai valori previsti

Se il modello scelto è corretto e se le assunzioni fatte sono soddisfatte, i residui dovrebbero essere privi di ogni struttura; in particolare essi non dovrebbero dipendere da alcuna altra variabile, compresa la risposta prevista. Un semplice controllo può essere effettuato riportando il grafico dei residui rispetto ai valori previsti  $\hat{y}_{ijk}$ . Questo grafico non dovrebbe mostrare alcuna tendenza manifesta. Un'anomalia che occasionalmente appare su questo grafico è la varianza non costante. Talvolta la varianza delle osservazioni cresce al crescere dell'entità delle osservazioni. Questa situazione si presenta quando l'errore, o il rumore di fondo dell'esperimento, è una percentuale costante della grandezza dell'osservazione (fatto che si verifica comunemente con molti strumenti di misurazione – l'errore è una percentuale della lettura sulla scala). Se il caso fosse questo, i residui dovrebbero crescere all'aumentare del valore  $y_{ijk}$  ed il grafico dei residui rispetto a  $\hat{y}_{ijk}$  dovrebbe assomigliare ad un imbuto o ad un megafono. Una varianza non costante si presenta anche in quelle situazioni in cui i dati seguono una distribuzione non normale, asimmetrica, poiché in certe distribuzioni asimmetriche la varianza tende ad essere funzione della media.

#### 2.3.2.4 Grafici dei residui rispetto ad altre variabili

Se i dati sono stati raccolti in corrispondenza di una qualunque altra variabile suscettibile di influenzare la risposta, si dovrebbero tracciare i grafici dei residui anche rispetto a queste variabili. Qualunque andamento non casuale in tali grafici dei residui implica che la variabile in questione è in grado di influenzare la risposta. Ciò consiglia di controllare tale variabile con maggiore cura in futuri esperimenti o di includerla esplicitamente nell'analisi.

#### *2.3.3 Scelta della dimensione campionaria*

Lo sperimentatore può utilizzare le **curve caratteristiche operative** o **curve O.C.** per determinare una dimensione campionaria appropriata (numero di replicazioni  $n$ ) per un piano fattoriale a due fattori. Le curve caratteristiche operative rappresentano la probabilità di non rifiutare l'ipotesi nulla nonostante questa sia falsa (errore di II specie o  $\beta$ ) in funzione di un parametro  $\Phi$  dipendente dagli effetti dei trattamenti e delle interazioni, per una certa dimensione campionaria. Nell'uso delle curve caratteristiche allo sperimentatore viene richiesto di dare uno specifico valore al parametro  $\Phi$ . Un modo molto efficace è trovare il più piccolo valore di  $\Phi^2$  che corrisponde ad una differenza prefissata tra qualunque coppia di medie di trattamento. Per esempio, se la differenza tra qualunque coppia di medie di riga è  $D$ , allora il valore minimo di  $\Phi^2$  è:

$$\Phi^2 = \frac{nbD^2}{2a\sigma^2} \quad (2.14)$$

mentre se la differenza tra qualunque coppia di medie di colonna è  $D$ , allora il valore minimo di  $\Phi^2$  è:

$$\Phi^2 = \frac{naD^2}{2b\sigma^2} \quad (2.15)$$

Infine, il valore minimo di  $\Phi^2$  corrispondente ad una differenza di  $D$  tra qualunque coppia di effetti di interazione è:

$$\Phi^2 = \frac{nD^2}{2\sigma^2[(a-1)(b-1)+1]} \quad (2.16)$$

Per usare queste equazioni, lo sperimentatore deve decidere un'appropriata differenza  $D$  tra le coppie di medie di trattamento per la quale rifiutare l'ipotesi nulla con alta probabilità e verificare il numero di replicazioni  $n$  per il quale si ha un rischio  $\beta$  limitato. Si noti che per utilizzare tali formule è richiesta una stima della deviazione standard: se vi fosse qualche dubbio, lo sperimentatore potrebbe ripetere la precedente procedura con altri valori di  $\sigma$  per determinare l'effetto di una stima errata di questo parametro sulla sensibilità del piano.

### 2.3.4 Un'osservazione per cella

Talvolta si incontrano esperimenti con una **singola replicazione**, cioè una sola osservazione per cella e il piano che si trova è chiamato fattoriale non replicato. In queste condizioni, però, mancano stime interne dell'errore ( o “errore puro”).

Un approccio all'analisi di un fattoriale non replicato consiste nell'assumere che certe iterazioni di ordine elevato siano trascurabili ed utilizzare la somma dei loro quadrati medi per stimare l'errore. Ciò equivale a ricorrere al **principio di rarità degli effetti**, osservando cioè che la maggior parte dei sistemi sono influenzati solo da alcuni effetti principali e dalle loro interazioni di basso ordine, mentre la maggior parte delle interazioni d'ordine elevato sono trascurabili ai fini pratici. Se ci sono due fattori e solo un'osservazione per cella, il modello è:

$$Y_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ij} \quad \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, b \end{cases} \quad (2.17)$$

Dall'esame delle medie attese delle somme dei quadrati si deduce che la varianza dell'errore  $\sigma^2$  non è stimabile; cioè l'effetto di interazione a due fattori  $(\tau\beta)_{ij}$  e l'errore sperimentale non possono essere separati in alcun modo. Di conseguenza non si possono fare test sugli effetti principali, salvo che l'interazione non sia zero. Se non è presente alcun effetto di interazione, allora  $(\tau\beta)_{ij} = 0$  per tutti gli  $i$  e  $j$  ed un modello plausibile è:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \quad \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, b \end{cases} \quad (2.18)$$

Se il modello è appropriato allora il quadrato medio del residuo è uno stimatore non distorto di  $\sigma^2$  e gli effetti principali possono essere sottoposti a test confrontando  $MS_A$  e  $MS_B$  con  $MS_{Residuo}$ .

## 2.4 Il piano composito centrale

Il CCD è un piano molto efficiente per accostare il modello del secondo ordine e consiste di tre gruppi di punti:

- a) punti appartenenti ad un fattoriale  $2^k$
- b) punti assiali ( chiamati anche “star points”)
- c) punti centrali

Occorre, quindi, specificare due parametri del piano: la distanza  $\alpha$  delle prove assiali dal centro del piano ed il numero di punti centrali.

La scelta di  $\alpha$  nel CCD è dettata principalmente dalla regione d'interesse. Se  $\alpha$  è uguale a uno, i punti assiali vengono disposti in posizione intermedia tra i due livelli dei fattori considerati.



I punti centrali di solito sono ripetuti in maniera da avere una buona stima dell'errore sperimentale. Si trasforma quindi in un disegno che richiede tre livelli per ciascun fattore.

Questa tipologia di piano è un modo eccellente per ottenere un'indicazione sulla curvatura.

Consente infatti di tenere bassa la dimensione e la complessità del piano ed ottenere allo stesso tempo una certa protezione nei confronti della curvatura.

Il più semplice tra questi piani è quello con due fattori ciascuno a tre livelli. In presenza di  $n$  repliche, ci saranno  $n3^2 - 1$  gradi di libertà in totale e  $3^2(n - 1)$  gradi di libertà per l'errore.

Un modello di regressione che lega la risposta  $Y$  ai regressori  $x_1$  e  $x_2$  è:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon \quad (2.19)$$

Ogni effetto può essere rappresentato da una componente lineare ed una quadratica, ognuna con un solo grado di libertà.

Le somme dei quadrati per gli effetti dei fattori e la loro interazione sono calcolate con il metodo esposto nel capitolo 2.3 .

## 2.5 Accostamento di modelli per regressione

Il problema della regressione riguarda l'ottenimento di modelli matematici che descrivano quantitativamente il comportamento di un sistema in funzione di alcuni fattori sperimentali (Montgomery e Peck, 1992). Lo scopo di ottenere un modello matematico di questo tipo può essere inquadrato principalmente in due esigenze:

- Comprendere quali siano le leggi che regolano il funzionamento di un sistema, il suo meccanismo.
- Utilizzare il modello matematico al posto della sperimentazione vera e propria per fare delle simulazioni o, in base al modello stesso, prevedere le migliori condizioni di funzionamento del sistema.

Da una parte, la forma della relazione ottenuta descrive la modalità con cui la descrizione del sistema si raccorda con la misura sperimentale (*fitting*), e, dall'altra il modello ottenuto, una volta verificata la sua qualità (*validazione*), consente di predire le risposte future di oggetti per i quali si conoscono soltanto le variabili che li descrivono ma non le misure sperimentali.

In modo più rigoroso, i metodi di regressione forniscono informazioni circa le relazioni quantitative tra una risposta  $y$  e un certo numero  $p$  di descrittori indipendenti  $x_1, \dots, x_p$  :

$$y = f(x_1, x_2, \dots, x_p)$$

Il problema generale della regressione si riconduce quindi a:

- stabilire il tipo di modello (la relazione  $f$ )
- determinare i parametri del modello
- valutare l'attendibilità del modello.

Sia  $\mathbf{X}$  la matrice dei dati con  $n$  righe (le osservazioni) e  $p$  colonne (le variabili),  $\mathbf{y}$  il vettore delle  $n$  risposte sperimentali,  $\mathbf{b}$  il vettore dei coefficienti del modello, di dimensione  $p'$ , dove  $p'$  è il numero dei parametri del modello.

Ad esempio un modello lineare in  $p$  variabili lineari con intercetta  $b_0$  è definito nel seguente modo:

$$y_i = b_0 + \sum_{j=1}^p b_j x_{ij}. \quad (2.20)$$

Un modello in  $p$  variabili è *lineare* se la risposta è una combinazione lineare delle variabili del modello, cioè se i coefficienti  $\mathbf{b}$  sono dei fattori moltiplicativi delle variabili. Un modello lineare in  $p$  variabili è quindi anche:

$$y_i = b_0 + \sum_{j=1}^p b_j f_j(x_{ij}). \quad (2.21)$$

In generale, dalla matrice originale dei dati  $\mathbf{X}$  si ottiene la matrice del modello  $\mathbf{X}_M$ . Questa matrice può contenere, oltre alle colonne della matrice  $\mathbf{X}$ , anche delle colonne aggiuntive che risultano da trasformazioni delle colonne originali, quali, ad esempio, termini quadratici ( $\mathbf{x}_1^2$ ,  $\mathbf{x}_2^2$ ) e termini misti ( $\mathbf{x}_1 \cdot \mathbf{x}_2$ ). In particolare, quando il modello prevede il termine  $b_0$  (l'intercetta), la matrice del modello viene costruita dalla matrice dei dati aggiungendo una colonna di 1, che indica la presenza nel modello di un termine costante.

### 2.5.1 Il metodo dei minimi quadrati ordinari

In termini matriciali, il problema della regressione lineare col metodo dei **minimi quadrati ordinari** (*Ordinary Least Squares, OLS*) è rappresentato dal seguente modello:

$$\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{e}, \quad (2.22)$$

dove  $\boldsymbol{\beta}$  è il vettore dei coefficienti veri da stimare,  $\mathbf{X}$  è la matrice del modello ed  $\mathbf{e}$  è il vettore degli errori; l'analisi dimensionale è la seguente:

$$(n, 1) = (n, p')(p', 1) + (n, 1),$$

con  $p' = p + 1$ .

In termini non matriciali si può scrivere esplicitamente la relazione tra la risposta  $y$  dell'*i-esimo* oggetto e la sua descrizione delle variabili indipendenti con la combinazione lineare:

$$y_i = b_0 + b_1 \cdot x_{i1} + b_2 \cdot x_{i2} + \dots + b_p \cdot x_{ip}.$$

La soluzione consiste nel determinare il vettore dei coefficienti  $\mathbf{b}$ . I seguenti passaggi algebrici portano alla soluzione cercata:

$$\begin{aligned} \mathbf{y} &= \mathbf{X} \cdot \mathbf{b} \\ \mathbf{X}^T \cdot \mathbf{y} &= \mathbf{X}^T \cdot \mathbf{X} \cdot \mathbf{b} \\ (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y} &= (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{X} \cdot \mathbf{b}. \end{aligned} \quad (2.23)$$

Poiché  $(\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{X} = \mathbf{I}$ , la soluzione **OLS** risulta:

$$\mathbf{b} = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y}. \quad (2.24)$$

L'analisi dimensionale è la seguente:

$$(p', 1) = (p', p')(p', n)(n, 1),$$

con  $p' = p + 1$ .

La matrice  $\mathbf{X}^T \cdot \mathbf{X}$  che compare nell'espressione matriciale di  $\mathbf{b}$  introdotta precedentemente viene comunemente chiamata **matrice di informazione**. La sua inversa, che è di estrema importanza sia per la teoria del disegno sperimentale che per i metodi di regressione, viene chiamata invece **matrice di dispersione** (o *matrice varianza-covarianza*). La matrice di dispersione ha un'importanza fondamentale nella teoria del disegno sperimentale; infatti, gli elementi lungo la sua diagonale principale sono proporzionali, attraverso l'errore sperimentale, all'incertezza che si ha nella stima dei coefficienti del modello di regressione. In pratica gli elementi diagonali della matrice di dispersione rendono conto della varianza dei coefficienti del modello. La somma di tali elementi, ovvero la traccia della matrice, divisa per il numero  $p$  dei coefficienti, dà la varianza media dei coefficienti. Invece, i termini fuori dalla diagonale contengono l'informazione riguardante la covarianza dei coefficienti del modello, cioè come si comporta un coefficiente se si commette un certo errore nella stima di un altro. I termini extradiagonali informano quindi sull'indipendenza relativa delle stime dei coefficienti. Valori elevati indicano che la stima dei relativi coefficienti non è indipendente e quindi la funzione delle variabili corrispondenti non è interpretabile correttamente sulla base dei coefficienti stessi. Dato che la matrice di dispersione dipende da come sono stati scelti gli esperimenti e non presuppone che gli stessi siano già stati eseguiti, ci si rende conto che, se da un lato è importante operare bene manualmente per ridurre l'errore puramente sperimentale ( $s^2$ ), dall'altro è importante progettare bene gli esperimenti in modo da operare sugli elementi diagonali della matrice di dispersione ( $d_{ii}$ ). La precisione relative dei coefficienti è tuttavia nota a priori, prima che vengano effettuati gli esperimenti. La raccolta dei dati permette di stimare l'errore sperimentale, da cui l'incertezza sulla stima dei coefficienti. La

matrice di dispersione ha un significato teorico molto importante in quanto i suoi valori definiscono, dato un certo livello di confidenza, un iperelissoide nello spazio dei fattori. Questo iperelissoide contiene i valori possibili, a quel certo livello di confidenza, dei coefficienti del modello. Questa regione, detta regione di probabilità congiunta (*joint probability region*), viene definita tenendo conto della mutua influenza dei fattori presenti. Il volume della *joint probability region* delimitata dall'iperelissoide è proporzionale alla varianza puramente sperimentale ed al determinante della matrice di dispersione. Quindi dato che il volume della regione è proporzionale all'incertezza sulla conoscenza dei coefficienti, per ottenere un'elevata precisione nella stima dei coefficienti del modello di regressione si devono scegliere gli esperimenti in modo che il determinante di  $(\mathbf{X}^T \cdot \mathbf{X})^{-1}$  sia minimo. Un tale piano viene detto *D-ottimale*. La matrice di dispersione permette anche di calcolare gli elementi della **matrice di correlazione dei coefficienti di regressione**. Gli elementi diagonali di questa matrice sono pari a 1, mentre gli elementi al di fuori sono calcolati dagli elementi della matrice di dispersione tramite la formula:

$$\rho_{ij} = \frac{c_{ij}}{\sqrt{c_{ii} \cdot c_{jj}}}, \quad (2.25)$$

dove  $c_{ii}$  sono gli elementi della diagonale di  $(\mathbf{X}^T \cdot \mathbf{X})^{-1}$  mentre  $c_{ij}$  sono gli elementi extradiagonali. La matrice di correlazione dei coefficienti di regressione indica la correlazione presente tra i coefficienti. Per questo presenta tutti i termini sulla diagonale pari a 1 (ogni coefficiente è correlato a sè stesso) ed è simmetrica. I termini extradiagonali indicano le eventuali interazioni. Se questi elementi sono pari a 0 i coefficienti di regressione sono indipendenti l'uno con l'altro, o ortogonali. Se sono presenti, queste correlazioni, positive o negative a seconda del segno degli elementi extradiagonali, tenderanno a offuscare le interpretazioni sui risultati del modello. La matrice di correlazione dei coefficienti di regressione non va confusa con la matrice di correlazione dei fattori, che indica invece quanto le variabili indipendenti e le loro interazioni sono correlate le une alle altre.

Dalla matrice di dispersione è possibile ricavare la **matrice di influenza** (o *matrice dei leverage* o **hat matrix**). Il vettore delle risposte calcolate  $\hat{y}$  si determina come:

$$\hat{y} = \mathbf{X} \cdot \mathbf{b}. \quad (2.26)$$

Sostituendo al vettore dei coefficienti  $\mathbf{b}$  l'espressione ricavata in precedenza, si ottiene un'importante relazione:

$$\hat{y} = \mathbf{X} \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y} = \mathbf{H} \cdot \mathbf{y}, \quad (2.27)$$

dove la matrice  $\mathbf{H}$ , di dimensione  $n \cdot n$ , è la matrice d'influenza che mette in relazione le risposte calcolate con quelle sperimentali.  $\mathbf{H}$  è pertanto definita come

$$\mathbf{H} = \mathbf{X} \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T. \quad (2.28)$$

Poichè dipende solo da  $\mathbf{X}$ , può essere determinata prima dell'esecuzione degli esperimenti. Gli elementi interessanti di questa matrice sono gli elementi diagonali  $h_{ii}$ , detti anche funzioni di varianza o *leverage*, per i quali valgono le seguenti proprietà:

- il leverage per un punto del piano sperimentale cade sempre nell'intervallo:

$$\frac{1}{n} \leq h_{ii} \leq \frac{1}{r_i}, \quad (2.29)$$

dove  $n$  è il numero totale di osservazioni ed  $r_i$  è il numero di replicazioni. Un metodo per scegliere i punti da replicare è quindi quello di considerare i punti con il più alto leverage, dal momento che il leverage rappresenta il peso che tali punti hanno sulla risposta predetta.

- Il leverage è una misura standardizzata della distanza dell'  $i$ -esimo punto del piano dal centro dello spazio dei dati: un valore elevato indica che l' $i$ -esima osservazione è distante dal centro del modello.
- Una superficie di risposta sarà portata a passare attraverso i punti aventi i maggiori valori di leverage. Quando  $h_{ii} = 1$ , il valore predetto sarà uguale al valore osservato, e la superficie di risposta passerà per quel punto.
- Le varianze dei valori predetti sono proporzionali ai valori di leverage tramite  $\sigma^2$ :

$$\text{var}(\hat{Y}_i | x_i) = \sigma^2 \cdot h_{ii} = \sigma^2 \cdot \mathbf{x}_i^T \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{x}_i, \quad (2.30)$$

Un'importante estensione di questa equazione applicabile ai punti  $\mathbf{x}_0$  che non sono necessariamente nel disegno sperimentale è la seguente:

$$\text{var}(\hat{Y}_i | x_0) = \sigma^2 \cdot \mathbf{x}_0^T \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{x}_0 = \sigma^2 \cdot h_{00}. \quad (2.31)$$

Il simbolo  $h_{00}$  è usato per rappresentare un qualsiasi punto nella regione sperimentale. E' quindi possibile calcolare **l'errore standard di predizione** per qualsiasi punto:

$$\text{s.e.}(\hat{Y}_i | x_0) = s \cdot \sqrt{h_{00}}. \quad (2.32)$$

Il leverage esprime quindi il contributo all'incertezza della risposta calcolata mediante il modello di regressione. Mentre per gli  $n$  oggetti utilizzati nella costruzione del modello i valori di  $h$  sono sempre compresi tra  $1/n$  e 1, nell'applicazione del modello a nuovi oggetti, il valore di  $h$  può anche essere molto maggiore di 1: si ha così a disposizione un indice che misura il grado di estrapolazione del modello. E' evidente quindi che nell'applicare a fini

predittivi il modello ottenuto a nuovi campioni, un alto valore di leverage (ad esempio maggiore di un valore di controllo  $h^*$  oltre il quale il dato può essere considerato *influyente* nel determinare i parametri del modello, con  $h^* > 3p'/n$ ) dovrebbe suggerire una certa cautela nell'accettare il valore predetto della risposta. In generale, è preferibile avere punti che abbiano circa la stessa influenza nel determinare il modello di regressione. Questo si può ottenere soltanto con punti i cui valori delle variabili che li descrivono sono ottenuti mediante un disegno sperimentale controllato.

- $\sum_{i=1}^n h_{ii} = p$  ovvero  $\text{tr}(\mathbf{H}) = p$ . Ciò significa che tralasciando  $\sigma^2$  la somma delle varianze di predizione sui punti del disegno è uguale al numero dei parametri del modello. Questa proprietà ha implicazioni importanti. Per un dato modello, la varianza di predizione totale sui punti del disegno è una costante al di là del numero di dati raccolti. Questo significa che aumentando il numero di punti del piano, la varianza di predizione totale sarà maggiormente diffusa tra i punti, portando ad un più basso valore della varianza media di predizione sui punti. Questa proprietà rende quindi più credibile la scelta di modelli meno complessi durante la fase di costruzione. Se tuttavia la diminuzione del numero dei termini riduce  $p$ , allo stesso modo  $s^2$  diventa preponderante a causa della mancanza di adattamento e comincerà ad aumentare.

### 2.5.2 I parametri di valutazione dei modelli di regressione

Per ogni modello di regressione si assume come situazione di riferimento o *modello di ordine zero*, la quantità riferita al valor medio della risposta, detta somma totale dei quadrati ( $SS_T$ ):

$$SS_T = \sum_i (y_i - \bar{y})^2. \quad (2.33)$$

Un modello di regressione è tanto migliore quanto più piccola è la somma dei quadrati dovuta all'errore ( $SS_E$ ):

$$SS_E = \sum_i (y_i - \hat{y})^2, \quad (2.34)$$

ottenuta dalla differenza tra ciascun valore sperimentale della risposta e la risposta calcolata. Nello stesso tempo, un modello di regressione è tanto migliore quanto più grande è la somma dei quadrati del modello ( $SS_R$ ):

$$SS_R = \sum_i (\hat{y}_i - \bar{y})^2. \quad (2.35)$$

Vale pertanto la seguente relazione:

$$SS_T = SS_R + SS_E. \quad (2.36)$$

E' importante notare che in un processo sequenziale di costruzione del modello, la parte sinistra della precedente equazione rimane costante. Nel valutare cioè funzioni polinomiali di diverso grado (lineari, quadratiche, cubiche, etc.), quello che varia è la distribuzione di questa quantità tra i due termini di variabilità spiegata dal modello ( $SS_R$ ) e variabilità casuale ( $SS_E$ ). Modelli di ordine inferiore spiegano una percentuale di variabilità inferiore. Quando siano presenti molte variabili e le loro combinazioni, nasce il problema di scegliere il miglior modello tra quelli possibili. Design Expert basa questa scelta sul test  $F$ , valutando l'ipotesi che la risposta sia invariata alla presenza o meno di termini che sono stati aggiunti al modello. Per questo la statistica  $F$  assume la forma:

$$F = \frac{(SS_{R_{ridotto}} - SS_{R_{completo}}) / r}{SS_{R_{completo}} / (n - p)}, \quad (2.37)$$

dove *ridotto* si riferisce all'utilizzo per la stima di  $y$  di un modello di regressione di ordine inferiore,  $r$  rappresenta la differenza tra il numero di parametri dei modelli completo e ridotto rispettivamente, mentre  $n$  è il numero di esperimenti e  $p$  il numero di parametri del modello completo (intercetta compresa se presente). Utilizzando questo test si valuta se il miglioramento introdotto dall'aggiunta di parametri al modello porti o meno ad un miglioramento significativo del modello stesso.

Design Expert suggerisce il miglior modello scegliendo quello che dà i migliori risultati per quanto riguarda particolari grandezze statistiche indicatrici della qualità sia di fitting che di previsione del modello stesso. Tali grandezze sono le medesime utilizzate per la validazione del modello. La grandezza utilizzata abitualmente per valutare la qualità di un modello di regressione è il **coefficiente di correlazione multipla  $R^2$** .

Per definizione esso è dato dalla formula:

$$R^2 = 1 - \frac{\sum_{i=1,n} (\hat{y}_i - y_i)^2}{\sum_{i=1,n} (y_i - \bar{y})^2} = 1 - \frac{SS_E}{SS_T} = \frac{SS_R}{SS_T}, \quad (2.38)$$

che si può leggere come il rapporto tra la varianza spiegata dal modello e la varianza contenuta nei risultati sperimentali, rapporto che, moltiplicato per cento, rappresenta la percentuale di varianza spiegata dal modello. Questa grandezza che, come detto, viene abitualmente utilizzata per descrivere se il modello si adegua bene ai dati sperimentali, in realtà soffre di un grande inconveniente, e cioè che cresce sempre all'aumentare del numero di variabili utilizzate nel modello di regressione, anche se le variabili introdotte non hanno alcun legame con la risposta studiata. Questo è dovuto al fatto che tale indice statistico non tiene conto del numero di gradi di libertà del modello in rapporto al numero di gradi di libertà forniti dagli esperimenti effettuati.

Una prima modifica a questo indice è stata apportata con l'introduzione del cosiddetto coefficiente di correlazione multipla **aggiustato** (*adjusted*) che corrisponde ad una definizione molto simile a quella fornita per l' $R^2$  classico, con l'introduzione di una pesatura a numeratore e denominatore coincidente col numero di gradi di libertà con cui ciascuna grandezza viene determinata:

$$R_{adj}^2 = 1 - \frac{\left(\sum_{i=1}^n (\hat{y}_i - y_i)^2\right)/(n-p)}{\left(\sum_{i=1}^n (y_i - \bar{y})^2\right)/(n-1)}, \quad (2.39)$$

dove come al solito  $n$  è il numero delle prove e  $p$  il numero di parametri presenti nel modello. Quando le variabili aggiunte non portano nuova informazione, la somma dei quadrati dei residui a numeratore non diminuisce di molto, per cui l'effetto del rapporto  $(n-1)/(n-p)$  fa diminuire questo indice. Il parametro  $R_{adj}^2$  presenta infatti un massimo per la complessità ottimale del modello e ridiscende quando l'aggiunta di una variabile al modello non è adeguatamente compensata da un significativo aumento di  $R$ . Il punto in cui  $R_{adj}^2$  è massimo coinciderà anche con il punto in cui la differenza tra  $R_{adj}^2$  e  $R$  è al minimo.

Entrambi questi indici sono ottenuti in condizioni cosiddette di fitting, cioè la valutazione della qualità del modello è basata sulla sua capacità di adeguarsi bene ai risultati sperimentali. Per ottenere dei parametri che misurino la capacità predittiva del modello ottenuto è necessario utilizzare le tecniche di **validazione**. Questo per avere un modello che sia in grado di predire efficacemente la risposta corrispondente a nuove condizioni sperimentali.

Una validazione delle capacità predittive di un modello di regressione può essere calcolata facilmente mediante gli algoritmi di cross-validazione (*cross-validation*). Il metodo più usato è quello definito *leave-one-out*: si supponga di avere  $n$  esperimenti. Al primo passo si lascia fuori il primo esperimento e si calcola il modello di regressione in sua assenza. Quindi si utilizza il modello appena calcolato per predire la risposta per il primo esperimento, che era stato lasciato fuori nella fase di calcolo del modello. Si ottiene così un valore stimato per la risposta che può essere interpretato come valore predetto. Da questo valore si può calcolare l'errore come differenza tra il valore stimato e quello vero. A questo punto si reintroduce il primo esperimento e si opera allo stesso modo sul secondo esperimento. Nuovamente viene calcolato un modello di regressione, questa volta in assenza del secondo esperimento, e si usa questo modello per stimare la risposta dell'esperimento lasciato fuori. Si può calcolare da qui una nuova differenza tra il valore vero della risposta e il valore predetto. Questo procedimento può essere ripetuto per ciascun esperimento presente nel set di dati e le differenze calcolate tra le risposte predette e quelle sperimentali sono utilizzate nell'espressione:



$$PRESS = \sum_{i=1}^n (y_i - \hat{y}_{i/i})^2, \quad (2.40)$$

dove *PRESS* sta per *PR*edictive *E*rroR *S*um of *S*quares mentre  $\hat{y}_{i/i}$  indica il valore predetto dal modello per l'*i*-esimo campione che non è stato considerato per calcolare il modello. Utilizzando *PRESS* al posto di  $SS_E$  nell'espressione di  $R^2$  si ottiene la percentuale di varianza spiegata dal modello in predizione:

$$R_{pred}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{PRESS}{SS_T}. \quad (2.41)$$

Si può dimostrare che questo nuovo coefficiente di correlazione multipla cross-validato non cresce necessariamente all'aumentare del numero di variabili del modello, anzi, è molto sensibile all'introduzione di variabili che portano soltanto rumore e non nuova informazione. Infatti  $R_{pred}^2$  può addirittura assumere valori negativi quando il set di variabili utilizzato non abbia alcuna relazione con la risposta studiata.

La differenza tra il valore osservato  $y_i$  e il valore predetto  $\hat{y}_{i/i}$  come riportata nella formula usata per calcolare *PRESS*, viene definita residuo *PRESS*. E' dimostrato che i residui *PRESS* possono essere facilmente calcolati dai residui ordinari attraverso l'espressione:

$$e_{i,-i} = \frac{e_i}{1 - h_{ii}}, \quad (2.42)$$

dove  $h_{ii}$  è il leverage dell'*i*-esimo punto. Un punto con un alto leverage sarà causa di una grande differenza tra il residuo ordinario e il residuo *PRESS*, indicando che quel particolare punto sperimentale ha una larga influenza sulla regressione. Sebbene le informazioni sull'influenza dei punti possono essere desunte dall'esame dei leverages, gli effetti sui residui possono essere abbastanza impressionanti.

### 2.5.3 Diagnostica del modello di regressione

#### 2.5.3.1 Residui ridotti

Questo paragrafo riassume alcune procedure diagnostiche (Montgomery e Peck, 1992). E' stato fatto notare che la diagnostica si basa sostanzialmente sullo studio dei residui, perché sono i residui ( $e_i$ ), che sono misurabili e che assumono il ruolo di surrogati degli errori concettuali ( $\varepsilon_i$ ). Molti tra quanti operano su modelli preferiscono lavorare con residui **ridotti**, piuttosto che sui residui

ordinari, in quanto i residui ridotti spesso forniscono più informazioni di quelli ordinari. Un tipo di residui ridotti è dato dai **residui standardizzati**, definiti dalla (2.11). Questi residui standardizzati, caratterizzati da media zero e varianza approssimativamente unitaria, tornano particolarmente utili nella ricerca degli **outlier** o valori anomali. Gli outlier dovrebbero essere esaminati con cura, poiché essi possono rappresentare sia un semplice errore di lettura dei dati, sia un'evenienza più preoccupante, quale l'esistenza di una regione nello spazio dei regressori in cui il modello accostato in realtà approssima in modo non adeguato la superficie di risposta vera.

Il processo di standardizzazione trasforma la scala dei residui in quanto li divide per la loro deviazione standard media approssimata. In alcuni sistemi di dati i residui possono avere deviazioni standard notevolmente differenti tra loro. Per tener conto di ciò si considerino i residui di un modello accostato scritti convenientemente in forma matriciale come:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}. \quad (2.43)$$

Considerando che  $\hat{\mathbf{y}} = \mathbf{H} \cdot \mathbf{y}$ , risulta che la matrice di covarianza dei residui è:

$$\text{Cov}(\mathbf{e}) = \sigma^2 \cdot (\mathbf{I} - \mathbf{H}). \quad (2.44)$$

La matrice  $\mathbf{I} - \mathbf{H}$  in genere non è diagonale; quindi i residui hanno varianze differenti e sono correlati. Pertanto la varianza dell' $i$ -esimo residuo è:

$$V(e_j) = \sigma^2 \cdot (1 - h_{ii}), \quad (2.45)$$

dove  $h_{ii}$  è il leverage del punto  $i$ -esimo. Poiché  $0 \leq h_{ii} \leq 1$ , usando i valori quadratici medi dei residui  $MS_E$  per stimare la varianza, in realtà la si sovrastima. Inoltre poiché  $h_{ii}$  è una misura della posizione dell' $i$ -esimo punto nello spazio delle  $x$ , la varianza di  $e_i$  dipende da dove si trova il punto  $x_i$ . In generale, residui prossimi al centro dello spazio delle  $x$  hanno varianza maggiore di quelli relativi a posizioni più distanti. Le violazioni delle assunzioni del modello sono più probabili nei punti più distanti e può essere arduo riconoscere queste violazioni dall'esame di  $e_i$  (o di  $d_i$ ) perché i loro residui di solito saranno più piccoli.

Per tener conto di questa disuniformità della varianza quando si trasforma la scala dei residui, si fa riferimento ai **residui studentizzati** (o internamente studentizzati, *internally studentized*):

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2 \cdot (1 - h_{ii})}} \quad i = 1, 2, \dots, n, \quad (2.46)$$

con  $\hat{\sigma} = \sqrt{MS_E}$ . I residui studentizzati hanno varianza costante pari a 1 indipendentemente dalla posizione di  $x_i$ , quando la forma del modello è corretta. In molte situazioni la varianza dei residui si stabilizza, in particolare per grandi insiemi di dati ed in questi casi le differenze tra residui

standardizzati e studentizzati saranno piccole. Quindi i residui standardizzati e studentizzati spesso producono la stessa informazione. Tuttavia poiché un punto con un residuo grande e un grande valore di  $h_{ii}$  può influenzare pesantemente l'accostamento ai minimi quadrati, in genere si raccomanda l'esame dei residui studentizzati.

Poiché nei residui studentizzati è consueto usare  $MS_E$  come stima di  $\sigma^2$ , ci si riferisce a ciò come ad una trasformazione interna della scala del residuo (da qui *internally studentized*), poiché  $MS_E$  è una stima generata internamente, ottenuta dall'accostamento del modello a tutte le  $n$  osservazioni. Un altro approccio potrebbe essere usare una stima di  $\sigma^2$  basata su un insieme di dati rimuovendo l' $i$ -esima osservazione. Indicando la stima di  $\sigma^2$  così ottenuta con  $S_{(i)}^2$  si può mostrare che:

$$S_{(i)}^2 = \frac{(n-p) \cdot MS_E - e_i^2 / (1-h_{ii})}{n-p-1}. \quad (2.47)$$

La stima di  $\sigma^2$  è usata al posto di  $MS_E$  per ottenere un residuo studentizzato esterno (*externally studentized*), di solito chiamato **R-Student**, dato da:

$$t_i = \frac{e_i}{\sqrt{S_{(i)}^2 (1-h_{ii})}} \quad i = 1, 2, \dots, n. \quad (2.48)$$

In molti casi  $t_i$  differirà poco dal residuo studentizzato  $r_i$ . Tuttavia, se l' $i$ -esima osservazione è influente, allora  $S_{(i)}^2$  può differire significativamente da  $MS_E$  e quindi l'**R-Student** sarà più sensibile a questo punto.

### 2.5.3.2 Diagnostici di influenza

Gli elementi diagonali della matrice **H**, cioè i leverages, identificano nello spazio delle  $x$  punti potenzialmente influenti a causa della loro posizione. Nel misurare tale influenza è opportuno tener conto sia della posizione del punto sia della variabile di risposta. Cook (1977, 1979) suggerisce di usare una misura del quadrato della distanza tra la stima ai minimi quadrati basata su tutti gli  $n$  punti  $\hat{\beta}$  e quella ottenuta tralasciando l' $i$ -esimo punto,  $\hat{\beta}_{(i)}$ . Questa misura di distanza, che prende il nome di **distanza di Cook**, può essere espressa come:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' \cdot X' \cdot X \cdot (\hat{\beta}_{(i)} - \hat{\beta})}{p \cdot MS_E} \quad i = 1, 2, \dots, n. \quad (2.49)$$

Un ragionevole livello di soglia per  $D_i$  è unitario, vale a dire che di solito consideriamo osservazioni per cui  $D_i > 1$  come influenti. La statistica  $D_i$  in realtà è calcolata da:

$$D_i = \frac{r_i^2}{p} \cdot \frac{V[\hat{y}(x_i)]}{V(e_i)} = \frac{r_i^2}{p} \cdot \frac{h_{ii}}{(1-h_{ii})} \quad i = 1, 2, \dots, n. \quad (2.50)$$

Si noti che, a parte la costante  $p$ ,  $D_i$  è il prodotto del quadrato dell' $i$ -esimo residuo studentizzato per  $h_{ii}/(1-h_{ii})$ ; si può dimostrare che questo rapporto è la distanza del vettore  $\mathbf{x}_i$  dal baricentro dei dati rimanenti. Quindi  $D_i$  è costituito da una componente che riflette quanto bene il modello accosti l' $i$ -esima osservazione  $y_i$  ed una componente che misura quanto disti quel punto dai dati restanti. Ognuna delle componenti, o entrambe, possono dar luogo ad un valore elevato di  $D_i$ .

Una misura della differenza tra valori calcolati e valori predetti è definita dalla seguente espressione:

$$DFFIT_i = \hat{y}_i - \hat{y}_{i/i} = r_i \cdot \left( \frac{h_{ii}}{1-h_{ii}} \right), \quad (2.51)$$

mentre la corrispondente misura normalizzata è data dall'espressione:

$$DFFITs_i = \frac{\hat{y}_i - \hat{y}_{i/i}}{s_{(i)} \cdot \sqrt{h_{ii}}} = r_{i/i} \cdot \left( \frac{h_{ii}}{1-h_{ii}} \right)^{1/2}. \quad (2.52)$$

Il valore di  $DFFITs$  è uguale al numero di errori standard stimati di  $\hat{y}_i$  che il valore fittato modifica quando l' $i$ -esimo punto non viene considerato nell'analisi. Valori di controllo per la rilevazione di un punto la cui influenza è anomala sono  $2 \cdot \sqrt{p'/n}$  e  $3 \cdot \sqrt{p'/n}$ . Un altro parametro utile per valutare le differenze tra i coefficienti di regressione calcolati con tutti i dati e senza l' $i$ -esimo dato è definito come:

$$DFBETA_i = \hat{\mathbf{b}} - \hat{\mathbf{b}}_{i/i} = \frac{(\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{x}_i \cdot r_i}{1-h_{ii}}. \quad (2.53)$$

Il corrispondente parametro scalato, relativo a ciascuna variabile, è

$$DFBETAS_{ij} = \frac{\hat{b}_j - \hat{b}_{j(i)}}{s_{(i)} \cdot \sqrt{d^{jj}}} = \frac{\left\{ (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{x}_i \right\}_j \cdot r_i}{s_{(i)} \cdot \sqrt{d^{jj}}} = \frac{\left\{ (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{x}_i \right\}_j \cdot r_{i/i}}{\sqrt{d^{jj}} \cdot \sqrt{1-h_{ii}}}, \quad (2.54)$$

dove il termine  $\left\{ (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{x}_i \right\}_j$  rappresenta il  $j$ -esimo elemento del vettore  $(\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{x}_i$ .

Valori di controllo di questo parametro sono  $2/\sqrt{n}$  e  $3/\sqrt{n}$ .

### 2.5.4 Le trasformazioni della risposta: il metodo di Box - Cox

Il problema della disomogeneità della varianza nella risposta di un esperimento programmato è uno scostamento dalle assunzioni fatte per l'ANOVA. Le trasformazioni della variabile di risposta sono un metodo appropriato per stabilizzarne la varianza. Le trasformazioni si usano inoltre per rendere la distribuzione della variabile di risposta più vicina alla distribuzione normale e migliorare l'accostamento del modello ai dati. Tali trasformazioni vengono spesso effettuate per tentativi scegliendo quella che produce il grafico più soddisfacente dei residui rispetto alla risposta prevista (§2.3.2.3). In una trasformazione risulta tuttavia essere molto utile la famiglia delle trasformazioni di potenze  $y^* = y^\lambda$ , dove  $\lambda$  è il parametro di trasformazione che deve essere determinato. Box e Cox (1964) hanno dimostrato come il parametro ed effetti dei trattamenti). La procedura di calcolo consiste nell'eseguire, per differenti valori di  $\lambda$ , un'analisi della varianza standard su:

$$y^{(\lambda)} = \begin{cases} y^\lambda - 1 / \lambda \hat{y}^{\lambda-1} & \lambda \neq 0 \\ \hat{y} \ln y & \lambda = 0 \end{cases}, \quad (2.55)$$

dove  $\hat{y} = \ln^{-1}[(1/n) \sum \log y]$  è la media geometrica delle osservazioni. La stima di massima verosimiglianza di  $\lambda$  è il valore per cui la somma dei quadrati dell'errore che si trova dall'analisi della varianza di  $\hat{y}^{(\lambda)}$  ( $SS_E(\lambda)$ ) è minima. Quello che solitamente viene fatto è tracciare un grafico di  $SS_E(\lambda)$  in funzione di  $\lambda$ , leggendo quindi il valore che minimizza  $SS_E(\lambda)$ . Design Expert riporta  $\ln(SS_E(\lambda))$ , oltre che un intervallo di confidenza al 95% per  $\lambda$ .

Tale intervallo è determinato calcolando:

$$SS^* = SS_E(\lambda) \cdot \left( 1 + \frac{t_{\alpha/2, \nu}^2}{\nu} \right), \quad (2.56)$$

( $\nu$  numero di gradi di libertà) e tracciando una linea parallela all'asse  $\lambda$  in corrispondenza del valore di  $SS^*$ . Le intersezioni con la curva  $SS_E(\lambda)$  rappresentano i limiti dell'intervallo. Il software riporta sul grafico oltre a tali limiti, il valore corrente di  $\lambda$ , pari a 1 in assenza di trasformazioni.

Se l'intervallo calcolato non include  $\lambda = 1$ , una trasformazione della risposta può essere utile. Una volta scelto un valore di  $\lambda$ , si possono analizzare i dati usando  $\hat{y}^{(\lambda)}$  come risposta, a meno che  $\lambda$  non sia zero, nel qual caso si può usare  $\ln y$ .

## 2.6 Risposte multiple: le funzioni di desiderabilità

Una procedura numerica utile all'ottimizzazione di risposte multiple sta nell'uso di tecniche di ottimizzazione simultanee, trattate da Derringer e Suich (1980), impiegando funzioni di desiderabilità. L'approccio generale consiste nel convertire dapprima ciascuna risposta  $y_i$  in una funzione individuale di desiderabilità  $d_i$  che assume valori nell'intervallo  $0 \leq d_i \leq 1$ , dove  $d_i = 1$  se la risposta  $y_i$  corrisponde al suo obiettivo, mentre  $d_i = 0$  se la risposta è esterna alla regione di accettabilità. Le variabili operative sono quindi scelte in modo da massimizzare la desiderabilità generale, data dalla media geometrica delle desiderabilità individuali:

$$D = (d_1 \cdot d_2 \cdot \dots \cdot d_m)^{1/m}, \quad (2.57)$$

che tiene conto di tutte le  $m$  risposte. La ragione per cui si usa la media geometrica anziché quella aritmetica sta nel fatto che se almeno una desiderabilità individuale è uguale a zero, la desiderabilità generale sarà uguale a zero, cioè basta che una risposta sia fuori dai limiti accettati perché l'intera situazione sia inaccettabile.

Le funzioni individuali di desiderabilità sono strutturate come mostrato in seguito: se l'obiettivo  $T$  per la risposta  $y$  è un valore massimo, si ha:

$$d = \begin{cases} 0 & y < L \\ \left( \frac{y-L}{T-L} \right)^r & L \leq y \leq T \\ 1, & y > T \end{cases} \quad (2.58)$$

quando il peso  $r$  è pari a 1, la funzione di desiderabilità è lineare. Scegliendo  $r > 1$  si pone più importanza allo stare vicino al valore obiettivo dal momento che si ottiene una famiglia di curve concave verso l'alto. Scegliendo  $0 < r < 1$  si pone meno importanza all'obiettivo ma si dà importanza all'intervallo attorno all'obiettivo, ottenendo una famiglia di curve concave verso il basso. La scelta di  $r$  così come quella dei limiti superiore e inferiore è soggettiva.

Se l'obiettivo  $U$  per la risposta  $y$  è un valore minimo, si ha:

$$d = \begin{cases} 1 & y < T \\ \left( \frac{U-y}{U-T} \right)^r & T \leq y \leq U \\ 0, & y > U \end{cases} \quad (2.59)$$

$$y > U$$

Se l'obiettivo è porre  $y$  il più vicino possibile ad un target, la funzione bilaterale, assumendo che l'obiettivo sia posizionato tra i limiti inferiore ( $L$ ) e superiore ( $U$ ), è definita come:

$$d = \begin{cases} 0 & y < L \\ \left(\frac{y-L}{T-L}\right)^{t_1} & L \leq y \leq T \\ \left(\frac{U-y}{U-T}\right)^{t_2} & T \leq y \leq U \\ 0, & y > U \end{cases} \quad (2.60)$$

L'ulteriore possibilità offerta da Design Expert è data dal porre l'intervallo come obiettivo. Un'ulteriore possibilità che da Design Expert è quella di poter dare a ciascuna risposta una determinata **importanza**  $t$ . L'equazione per il calcolo della desiderabilità generale viene così modificata:

$$D = (d_1^{t_1} \cdot d_2^{t_2} \cdot \dots \cdot d_m^{t_m})^{1/(t_1+t_2+\dots+t_m)}, \quad (2.61)$$

dove  $t_1$  è l'importanza relativa della risposta 1,  $t_2$  è l'importanza relativa della risposta 2, e così via. I valori di  $t_i$  possono essere scelti su Design Expert in un intervallo che va da 1 a 5.





# Capitolo 3

## Parte Sperimentale

L'attività sperimentale è consistita nella progettazione di due piani sperimentali (uno utilizzato per le prove in laboratorio, l'altro per le prove in impianto pilota in produzione), nell'esecuzione degli esperimenti e raccolta dati, nell'analisi delle risposte e ottimizzazione delle condizioni operative. L'intero percorso, dalla fase di pianificazione all'ottimizzazione, è stato supportato dall'utilizzo del software Design-Expert (versione 7.0.10) di Stat-Ease®.

### 3.1 Descrizione dell'apparecchiatura

Gli impianti pilota su cui si è condotta la parte sperimentale sfruttano il principio dell'alto potere di dispersione ottenibile con corpi macinanti di piccolo diametro sottoposti, insieme alla massa da disperdere, a un movimento tale da creare elevati attriti e impatti.

#### 3.1.1 Impianto Laboratorio

L'impianto pilota (Figura 3.1), su cui è stata condotta la parte sperimentale in laboratorio, è costituito da un sistema di macinazione DISPERMAT®, formato da una tazza della capacità di 1000 ml, con camicia di raffreddamento, albero motore, girante in nylon, variatore di giri.



**Figura 3.1** Apparecchiatura su cui sono state condotte le prove

All'interno della tazza, per tutte le prove, sono stati introdotti 260 grammi dell'impasto da disperdere e 520 grammi di sfere di vetro con diametro pari a due millimetri.

### 3.1.2 Impianto Produzione

L'impianto pilota (Figura 3.2) è costituito da un sistema di macinazione (mulino a microsferi), 2 contenitori carrellati (capacità 1000 l), un agitatore a cowless, il tutto corredato con un sistema di supervisione e controllo che permette di lavorare in modalità automatica, semiautomatica e manuale (foto).



*Figura 3.2 Impianto pilota su cui è stata condotta l'analisi*

Il mulino consta essenzialmente di un cilindro di macinazione (capacità 60 l), riempito di sfere di ossido di zirconio (diametro 2 mm), un albero a giranti multiple collegato tramite cinghia di trasmissione a un motore (M1), una pompa volumetrica a disco cavo oscillante (G1).

La foto dell'apparecchiatura e i relativi dati tecnici sono riportati in Figura 3.2 e Tabella 3.1.

In Figura 3.3 si riporta il particolare del sinottico.

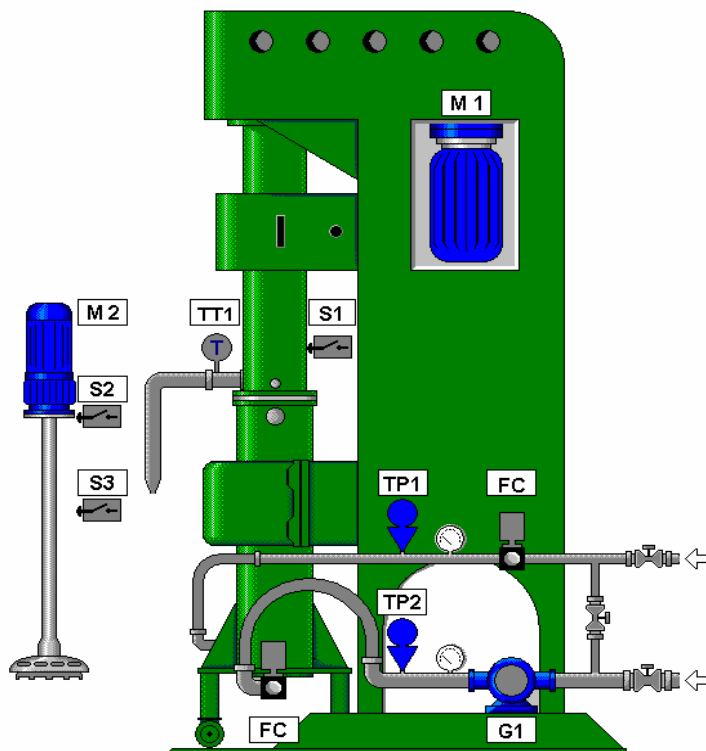


Figura 3.3 Impianto pilota: particolare del sinottico

<b>DATI TECNICI MACCHINA MS60P TS</b>		
Peso totale: 3300 Kg. Peso cilindro: 300 Kg.		
Motore principale con inverter (M1) Rapporto motore/albero 3.2	Kw 30	Giri/min (con max rapporto di puleggia) 1465
Impostabile da 780 giri/min (40 Hz) a 1375 giri/min (70 Hz). Assorbimento massimo raggiungibile dal motore principale impostabile da 0 a 99 A.		
Pompa: tipo disco cavo oscillante (G1) Motore pompa	Kw 0.75	
impostabile da 0 a 1575 giri/min. (tolleranza ±10)		
Motore agitatore (M2)	Kw 1.85	Giri/min 1390
Variatore: Bonfiglioli AS 20P P90		

Tabella 3.1 Dati tecnici dell'impianto pilota

## 3.2 Progettazione del piano sperimentale

Il lavoro è consistito nella progettazione di due piani sperimentali da effettuarsi l'uno in laboratorio chimico e l'altro in produzione.

La tipologia scelta per entrambi è stato un piano composito centrale (CCD) con replica del punto centrale, in cui sono state considerate due variabili:

- il tempo di permanenza all'interno della tazza in laboratorio e di ricircolo in produzione, fattore quantitativo su tre livelli (45-56'30"-68 minuti in laboratorio e 240-360-480 minuti in produzione)
- la velocità della girante, fattore quantitativo su tre livelli ( 4000-4500-5000 giri/minuto in laboratorio e 980-1055-1130 giri/minuto in produzione)

I livelli del piano sono stati scelti partendo dai valori standard (240 minuti e 980 giri/minuto) che si utilizzano attualmente nella lavorazione dell'impasto e sono stati aumentati in maniera tale da studiare una zona sperimentale in cui sono state portate al limite le condizioni di processo attualmente adottate

Le condizioni imposte, per poter confrontare le superfici di risposta generate dai due piani, sono state determinate da esperienze pregresse aziendali che permettessero il trasferimento delle condizioni operative dal laboratorio alla produzione.

Nella progettazione del piano si è considerato un unico lotto di materia prima necessario all'esecuzione di tutte le prove dei due piani.

Sono state analizzate pertanto tutte le possibili combinazioni dei livelli dei due fattori, ciascuna delle quali ripetuta una sola volta in produzione mentre, in laboratorio, sono state eseguite due prove per ogni punto del piano e si è considerata la media aritmetica relativa ad ognuno. Gli esperimenti sono stati svolti in una sequenza completamente casualizzata generata dal software. Nelle Tabelle 3.2 e 3.3 sono riportati i punti sperimentali ordinati secondo l'ordine standard.

*Tabella 3.2 Riassunto del piano sperimentale in laboratorio*

		Fattore 1	Fattore 2		
Ordine standard	Ordine di esecuzione	Tempo minuti	Velocità Giri/min	Fattore 1 codificato	Fattore 2 codificato
1	4	45	4000	-1	-1
2	10	68	4000	1	-1
3	7	45	5000	-1	1
4	9	68	5000	1	1
5	5	45	4500	-1	0
6	8	68	4500	1	0
7	2	56,5	4000	0	-1
8	3	56,5	5000	0	1
9	6	56,5	4500	0	0
10	1	56,5	4500	0	0

**Tabella 3.3** Riassunto del piano sperimentale in produzione

		Fattore 1	Fattore 2		
Ordine standard	Ordine esecuzione di	Tempo minuti	Velocità Giri/min	Fattore 1 codificato	Fattore 2 codificato
1	1	240	980	-1	-1
2	3	480	980	1	-1
3	8	240	1130	-1	1
4	6	480	1130	1	1
5	5	240	1055	-1	0
6	2	480	1055	1	0
7	4	360	980	0	-1
8	10	360	1130	0	1
9	9	360	1055	0	0
10	7	360	1055	0	0

### 3.3 Esecuzione delle prove: raccolta dati e risposte

In produzione, ciascuna prova è consistita nel far passare attraverso il mulino a microsferi 300 kg di prodotto da raffinare pompato da una bacinella tramite la pompa volumetrica (G1) all'interno del cilindro contenente le microsferi, è portato a ricircolo settando i parametri macchina secondo quanto riportato in Tabella 3.3. In laboratorio, ogni prova è stata realizzata utilizzando 260 g di impasto e riempiendo la tazza con 520 g di corpi sferici in vetro del diametro di due millimetri.

L'impasto raffinato ottenuto da tutte le prove è stato successivamente completato con acqua ed addensante secondo ricetta standard.

Le variabili di risposta analizzate sono state la resa coloristica e la viscosità misurata in  $mPa \cdot s$ .

La resa coloristica (nel prosieguo della trattazione denominata semplicemente resa) è stata determinata applicando il metodo interno SAMIA IL02 ( $CV\% = 0,4$ ), tramite spettrofotometro per letture nel visibile in riflessione, X-RITE 8200B e relativo software di calcolo.

La viscosità è stata misurata in accordo con la procedura interna SAMIA IL20 ( $CV\% = 1,5$ ), tramite un sistema costituito da viscosimetro HAAKE VT500, termostato HAAKE F3, bagno termostatico HAAKE C, sensore HAAKE SV-E.

Le misure sono state condotte alla temperatura di  $25^{\circ}C$ , la velocità del sensore è stata impostata a 45.3 rpm.

Si è riportato il quadro sinottico delle prove effettuate con le rispettive risposte nelle Tabelle 3.4 e 3.5.

Le risposte valutate nel piano relativo al laboratorio sono state la media tra due repliche.

**Tabella 3.4** Riassunto delle prove sperimentali e delle risposte ottenute in laboratorio.

		Factor 1	Factor 2	Response 1	Response 2
<b>Std</b>	<b>Run</b>	<b>A:TEMPO</b>	<b>B:VELOCITA'</b>	<b>resa</b>	<b>viscosità</b>
		min	giri/min		<i>mPa · s</i>
10	1	56,5	4500	100,6	405,5
7	2	56,5	4000	100,8	604
8	3	56,5	5000	101,6	407,5
1	4	45	4000	100,6	432,5
5	5	45	4500	100,8	430,5
9	6	56,5	4500	101,8	452,5
3	7	45	5000	100,8	451,5
6	8	68	4500	103,4	690
4	9	68	5000	104,1	1190
2	10	68	4000	101,2	1410

**Tabella 3.5** Riassunto delle prove sperimentali e delle risposte ottenute in produzione

		Factor 1	Factor 2	Response 1	Response 2
<b>Std</b>	<b>Run</b>	<b>A:TEMPO</b>	<b>B:VELOCITA'</b>	<b>resa</b>	<b>viscosità</b>
		min	giri/min		<i>mPa · s</i>
1	1	240	980	102,4	487
6	2	480	1055	106,9	1110
2	3	480	980	106,7	957
7	4	360	980	105,3	940
5	5	240	1055	103,6	586
4	6	480	1130	106,5	1110
10	7	360	1055	105,4	933
3	8	240	1130	101,9	735
9	9	360	1055	105	852
8	10	360	1130	101,9	1100

Si è successivamente passati all'analisi delle risposte, effettuata con il supporto del software. La procedura, dopo una prima valutazione del piano sperimentale, si è articolata attraverso la scelta del modello da accostare ai dati, la successiva analisi della varianza, il controllo dell'adeguatezza del modello e la determinazione della superficie di risposta.

### 3.4 Valutazione del piano sperimentale

Prima di procedere con l'analisi delle risposte, si è valutato il disegno sperimentale, che è risultato uguale per entrambe le situazioni studiate: attraverso il software è stata generata una risposta casuale normalmente distribuita con varianza unitaria ed è stato accostato ad essa un modello quadratico che consentisse di valutare l'errore standard lungo il dominio sperimentale.

In Tabella 3.6 è mostrato il riassunto relativo ai gradi di libertà.

**Tabella 3.6** *Gradi di libertà associati al piano sperimentale*

**2 Factors: A, B**

**Design Matrix Evaluation for Response Surface Quadratic Model**

**No aliases found for Quadratic Model**

**Degrees of Freedom for Evaluation**

Model	5
Residuals	4
Lack Of Fit	3
Pure Error	1
Corr Total	9

“No aliases found” si riferisce al fatto che per il modello selezionato (di secondo ordine), non si sono ottenuti effetti sovrapposti che non fossero stimabili. Nella Tabella 3.7 sono presenti i risultati della valutazione del modello:

**Tabella 3.7** *Valutazione del disegno sperimentale*

Term	StdErr**	VIF	Ri-Squared	Power at 5% alpha level for effect of		
				1/2 Std. Dev.	1 Std. Dev.	2 Std. Dev.
A	0.41	1.00	0.0000	7.7 %	15.9 %	46.3 %
B	0.41	1.00	0.0000	7.7 %	15.9 %	46.3 %
A <sup>2</sup>	0.65	1.03	0.0278	9.2 %	21.9 %	63.4 %
B <sup>2</sup>	0.65	1.03	0.0278	9.2 %	21.9 %	63.4 %
AB	0.50	1.00	0.0000	6.8 %	12.2 %	33.6 %

\*\*Basis Std. Dev. = 1.0

In essa sono riportate interessanti statistiche: l'indice VIF (*variance inflation factor*) misura quanto cresce la varianza dei coefficienti del modello selezionato a seguito della mancanza di ortogonalità nella matrice di dispersione. Più specificamente l'errore standard di un coefficiente del modello cresce proporzionalmente alla radice quadrata di VIF. Se un coefficiente è ortogonale ai rimanenti

termini del modello, il suo VIF è unitario. Questo valore o uno superiore, come è suggerito in tabella, indica che è presente una certa correlazione tra i coefficienti. VIF è legato a *R-squared* dalla formula:

$$VIF = \frac{1}{(1 - R - squared)} \quad (3.1)$$

*R-squared* è il coefficiente di correlazione multipla che indica quanto il coefficiente per quel termine è correlato agli altri. I termini *R-squared* dovrebbero essere vicini a 0 per indicare piccola correlazione, come avviene nel caso in esame.

Con il termine “*power*” che compare nell’output del software in Tabella 3.7 si intende la capacità del disegno sperimentale di individuare quanto i termini specifici siano statisticamente significativi. Per piani fattoriali generali, la potenza (*power*) del disegno è definita come la probabilità di risolvere due termini, all’interno di uno stesso effetto principale, se la differenza tra questi cade entro ½, 1 or 2 deviazioni standard.

L’output afferma che questo disegno ha dal 33,6% al 63,4% di probabilità di identificare come statisticamente significativo uno qualsiasi dei termini considerati entro due deviazioni standard. Se l’effetto è grande solo quanto una deviazione standard, la probabilità di determinare la significatività si riduce al 21,9 o al 12,2%. Esiste la probabilità del 5% che un termine che appaia essere statisticamente significativo, in realtà non lo sia e l’effetto sia dovuto solo all’errore casuale.

In Tabella 3.8 sono riportati i leverage dei diversi punti sperimentali. Si è assunto come valore di controllo per i leverage l’unità. È possibile osservare quindi che nessun punto ha avuto un’influenza elevata relativamente agli altri sul disegno. I punti che hanno presentato più alto leverage sono quelli ai confini del dominio sperimentale, dove il “rumore di fondo” (Figura 3.3) è maggiore .

**Tabella 3.8** Valori di leverage per i punti sperimentali

Std	Leverage	Point Type
1	0.7976	Fact
2	0.7976	Fact
3	0.7976	Fact
4	0.7976	Fact
5	0.5238	Axial
6	0.5238	Axial
7	0.5238	Axial
8	0.5238	Axial
9	0.3571	Center
10	0.3571	Center
<b>Average =</b>	<b>0.6000</b>	



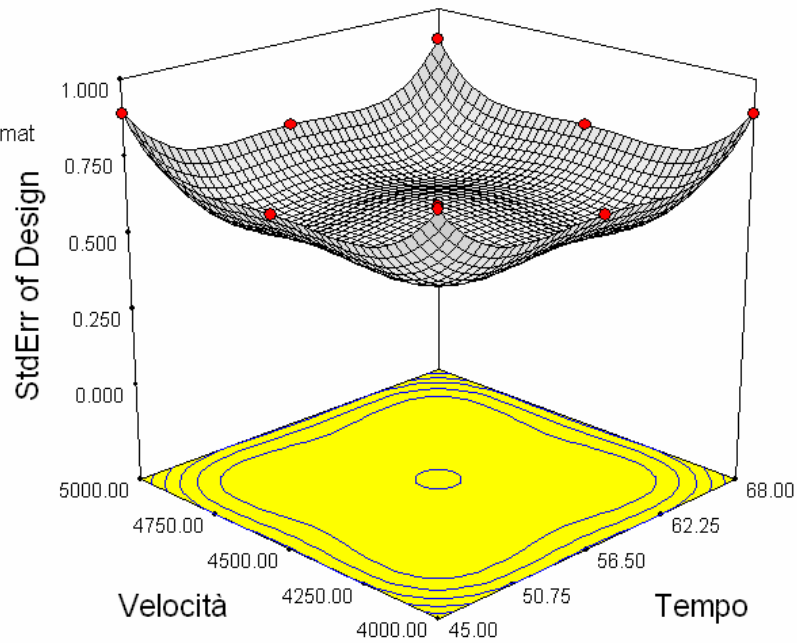
Design-Expert® Software

StdErr of Design



X1 = A: TEMPO dispermat

X2 = B: VELOCITA' dispermat



**Figura 3.4** *Errore standard nel dominio sperimentale*

La forma della superficie in Figura 3.4 dipende solo dai punti del disegno sperimentale e dal polinomio utilizzato per il fit. La forma reale sarà funzione della deviazione standard, che dipende dalle risposte osservate. Per generare la superficie in Figura 3.4 è stata considerata una deviazione standard pari a uno. La forma ideale, come si osserva in figura, è una superficie simmetrica. Un'altra caratteristica desiderabile è un errore relativamente basso in prossimità del centro.

## 3.5 Analisi delle risposte

### 3.5.1 Resa-Laboratorio

La prima risposta presa in considerazione è stata la resa coloristica ottenuta in laboratorio. Sulla base dei dati raccolti è stato suggerito di utilizzare un modello lineare oppure un modello di secondo ordine in cui sia contenuto solo il termine d'interazione tra i due fattori come indicato in Tabella 3.9.

**Tabella 3.9** Suggerimento del modello per la risposta "resa"

Response 1      resa      Transform:      None

\*\*\* WARNING: The Cubic Model is Aliased! \*\*\*

Sequential Model Sum of Squares [Type I]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Mean vs Total	1.032E+005	1	1.032E+005			
<u>Linear vs Mean</u>	<u>9.46</u>	<u>2</u>	<u>4.73</u>	<u>8.13</u>	<u>0.0149</u>	<u>Suggested</u>
<u>2FI vs Linear</u>	<u>1.81</u>	<u>1</u>	<u>1.81</u>	<u>4.80</u>	<u>0.0710</u>	<u>Suggested</u>
Quadratic vs 2FI	1.08	2	0.54	1.83	0.2729	
Cubic vs Quadratic	0.35	2	0.18	0.42	0.7020	Aliased
Residual	0.83	2	0.41			
Total	1.032E+005	10	10318.43			

"Sequential Model Sum of Squares [Type I]": Select the highest order polynomial where the additional terms are significant and the model is not aliased.

Source	Std. Dev.	R-Squared	Adjusted R-Squared	Predicted R-Squared	PRESS	
<u>Linear</u>	<u>0.76</u>	<u>0.6991</u>	<u>0.6131</u>	<u>0.3041</u>	<u>9.41</u>	<u>Suggested</u>
<u>2FI</u>	<u>0.61</u>	<u>0.8328</u>	<u>0.7492</u>	<u>0.6430</u>	<u>4.83</u>	<u>Suggested</u>
Quadratic	0.54	0.9127	0.8035	0.5212	6.48	
Cubic	0.64	0.9387	0.7241	-0.5083	20.40	Aliased

"Model Summary Statistics": Focus on the model maximizing the "Adjusted R-Squared" and the "Predicted R-Squared".

Si è notato come il modello cubico fosse "aliased" ovvero con tutti gli effetti non stimabili perché sovrapposti. Per farlo, sarebbe stato necessario aumentare il numero di livelli per ogni variabile. Il suggerimento del miglior modello viene fatto sulla base del test  $F$ , confrontando tra loro modelli a cui in successione vengono aggiunti termini. L'attenzione è stata in seguito focalizzata sul modello che massimizzasse gli indici statistici di fitting e di predizione (§2.6.2). In particolare sono stati considerati migliori i modelli che fornissero il miglior risultato per quanto riguarda  $R_{adj}^2$  e  $R_{pred}^2$ .

Per l'analisi dei dati si è scelto di utilizzare il modello di secondo ordine in maniera tale da indagare le interazioni tra i fattori, vista la superiore capacità di spiegare la variabilità dei dati e la migliore capacità predittiva rispetto al modello lineare.

Si è passati pertanto all'analisi della varianza (riportata in Tabella 3.10), per verificare la significatività dei fattori. Si è visto che erano significativi entrambi i fattori, mentre l'interazione fra i due presentava un p-value superiore al 5% che la rendeva quindi non indicativa. Nella parte inferiore dell'output sono state riportate le prestazioni del modello fittato, introdotte in precedenza.

La grandezza C.V.% è il coefficiente di variazione definito come:

$$C.V.\% = \frac{\sqrt{MS_E}}{\bar{y}} \cdot 100 \tag{3.2}$$

Con esso si esprime la variabilità non spiegata o residua dei dati come percentuale della risposta media. Si è osservato come in questo caso sia molto basso. La statistica "Adeq Precision", calcolata dividendo la differenza tra la massima e la minima risposta prevista per la deviazione standard di tutte le risposte previste, misura sostanzialmente un rapporto segnale/rumore. Per questa statistica si preferiscono valori elevati. Solitamente viene assunto 4 come valore di controllo: valori superiori indicano che il modello ha buone capacità previsionali. In questo caso la situazione è stata ottimale essendo calcolato un valore di "Adeq Precision" pari a nove.

**Tabella 3.10 ANOVA per la risposta "resa-laboratorio"**

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	11.26	3	3.75	9.96	0.0096	significant
A-TEMPO disperme	6.93	1	6.93	18.40	0.0052	
B-VELOCITA' dispe	2.52	1	2.52	6.69	0.0414	
AB	1.81	1	1.81	4.80	0.0710	
Residual	2.26	6	0.38			
Lack of Fit	1.57	5	0.31	0.45	0.8039	not significant
Pure Error	0.70	1	0.70			
Cor Total	13.53	9				

The Model F-value of 9.96 implies the model is significant. There is only a 0.96 % chance that a "Model F-Value" this large could occur due to noise.

Std. Dev.	0.61	R-Squared	0.8328
Mean	101.57	Adj R-Squared	0.7492
C.V. %	0.60	Pred R-Squared	0.6430
PRESS	4.83	Adeq Precision	9.001

Vengono quindi calcolati i coefficienti di regressione per il modello (Tabella 3.11):

**Tabella 3.11** *Calcolo dei coefficienti di regressione*

Factor	Coefficient		Standard Error	95% CI		VIF
	Estimate	df		Low	High	
Intercept	101.57	1	0.19	101.10	102.05	
A-TBMPO dispermat	1.07	1	0.25	0.46	1.69	1.00
B-VELOCITA' dispermat	0.65	1	0.25	0.035	1.26	1.00
AB	0.67	1	0.31	-0.079	1.42	1.00

Final Equation in Terms of Coded Factors:

$$\begin{aligned} \text{resa} = & \\ & +101.57 \\ & +1.07 * A \\ & +0.65 * B \\ & +0.67 * A * B \end{aligned}$$

Per ogni coefficiente è stato indicato anche l'intervallo di confidenza al 95% calcolato come:

$$\hat{\beta} - t_{0,025,N-p} se(\hat{\beta}) \leq \beta \leq \hat{\beta} + t_{0,025,N-p} se(\hat{\beta})$$

dove  $se$  è l'errore standard di ciascun coefficiente, calcolato come  $\sqrt{\frac{MS_E}{n}}$ , con  $n$  numero di prove. Viene inoltre calcolata l'equazione finale in termini di fattori adimensionali.

In Tabella 3.12 sono riportate le equazioni ottenute per il modello in cui sono stati utilizzati i fattori con le loro effettive unità di misura:

**Tabella 3.12** *Equazioni del modello di regressione per la risposta "resa-laboratorio"*

Final Equation in Terms of Actual Factors:

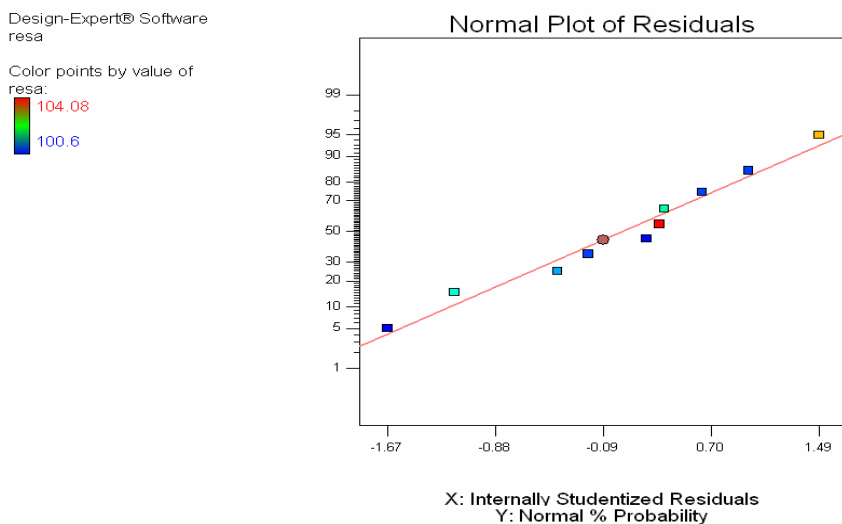
$$\begin{aligned} \text{resa} = & \\ & +120.19267 \\ & -0.43283 * \text{TBMPO dispermat} \\ & -5.31138\text{E-}003 * \text{VELOCITA' dispermat} \\ & +1.16957\text{E-}004 * \text{TBMPO dispermat} * \text{VELOCITA' dispermat} \end{aligned}$$

Nella fase successiva è stato previsto il controllo diagnostico sul modello verificando le assunzioni di normalità e indipendenza degli errori e l'omogeneità della varianza.

Attraverso la diagnostica si sono presi in esame i residui analizzando il loro comportamento.

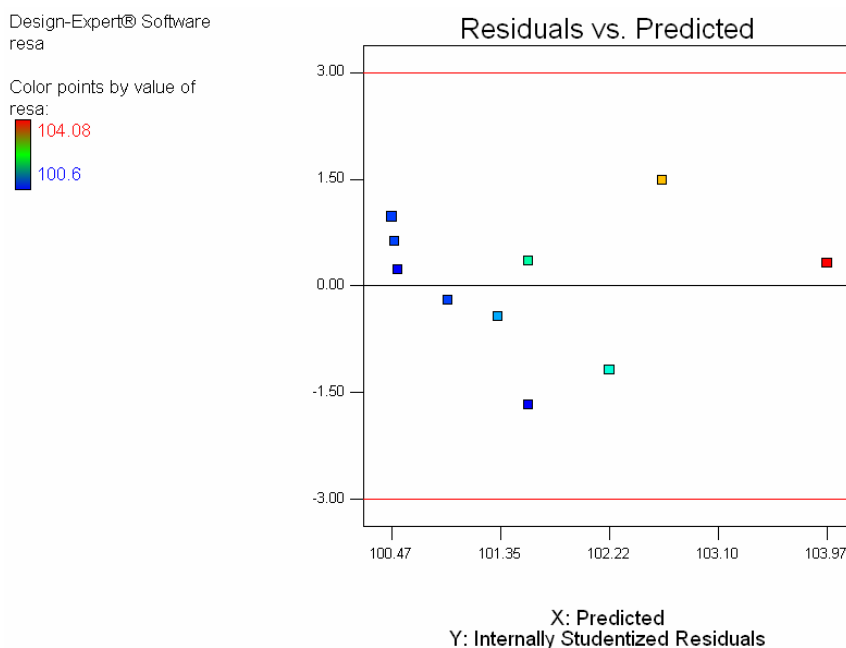
L'assunzione di normalità può essere verificata valutando il grafico di probabilità normale in Figura 3.5. L'idealità è data dalla retta segnata in rosso, indicativa della distribuzione normale. Dal grafico si è notato che non c'erano motivi per dubitare dell'assenza di normalità.

Non è stata rilevata la presenza di outlier ovvero elementi che si discostassero dall'andamento rettilineo dei punti nel grafico. La ragione per cui sono stati riportati in ascissa i residui internamente studentizzati sta nel fatto che in tal maniera si è tenuto conto anche della posizione del punto nel dominio sperimentale, cosa che non avveniva per i residui.



**Figura 3.5** Grafico di probabilità normale dei residui per la risposta “resa-laboratorio”

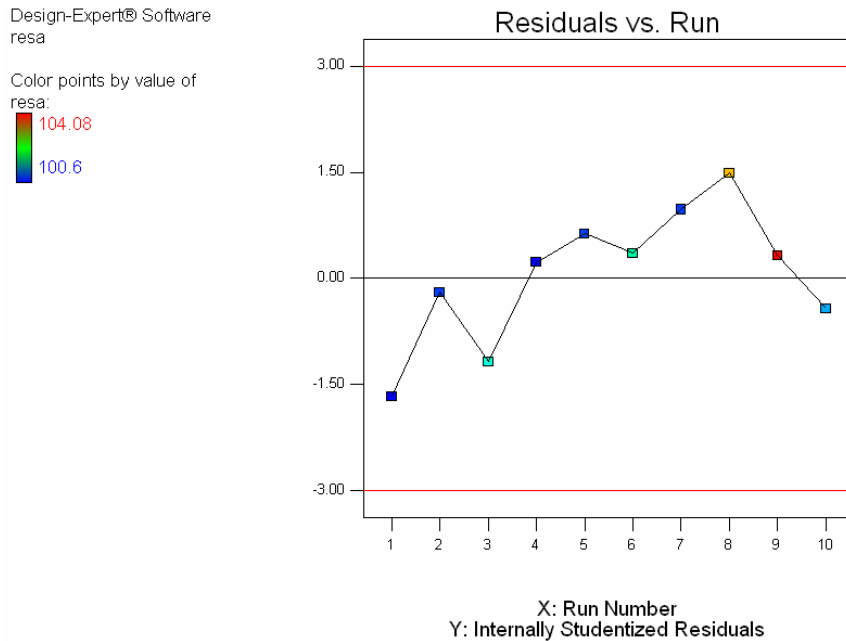
Sono stati riportati nel grafico successivo i residui studentizzati rispetto ai valori predetti. Il suo scopo è stato quello di evidenziare un eventuale disomogeneità della varianza. Infatti, un qualsiasi andamento regolare era indice di una varianza non costante. Nel caso in esame (Figura 3.6) non viene evidenziato alcun andamento particolare e si è osservato, inoltre, che la maggioranza dei punti sta all'interno di più o meno una deviazione standard.



**Figura 3.6** Grafico dei residui rispetto ai valori previsti per la risposta “resa-laboratorio”

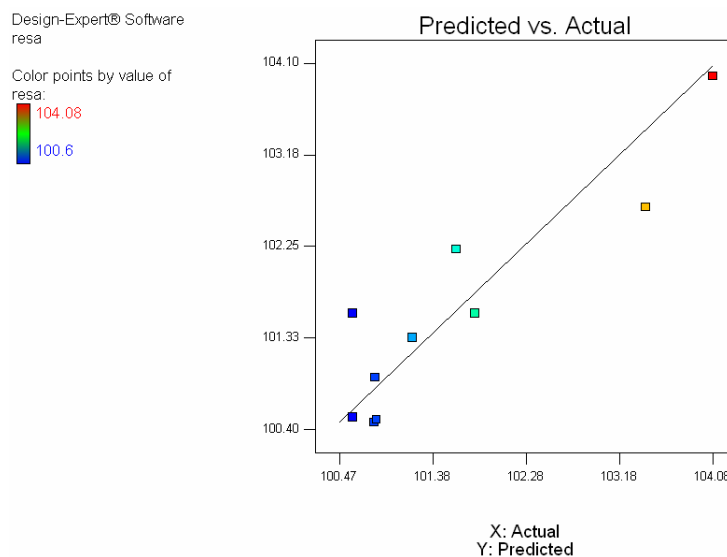
In Figura 3.7 è stato proposto il grafico dei residui studentizzati in funzione della sequenza sperimentale in cui può essere evidenziata una possibile correlazione tra i residui e quindi la mancanza di indipendenza.

Si è notato che non erano presenti andamenti particolari o sistematici, ovvero non si sono riscontrate sequenze di residui positivi alternate a sequenze di residui negativi, ma si è osservato, piuttosto, un andamento altalenante.



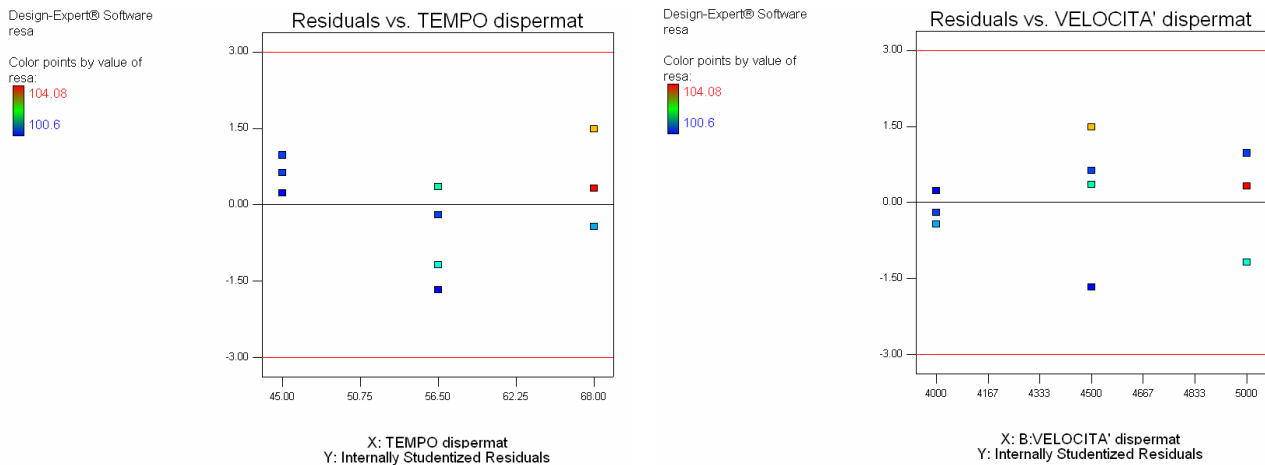
**Figura 3.7** Grafico dei residui rispetto la sequenza sperimentale

E' stato riportato quindi il grafico dei valori predetti contro i valori osservati: in esso si è mostrata la presenza o meno di valori che non fossero facilmente predetti dal modello. L'andamento rettilineo costituisce l'idealità (Figura 3.8). Si è constatato che l'andamento dei punti non segue esattamente la retta, ma che nello scostamento massimo dei punti non viene superato comunque il valore di una deviazione standard e mezzo.



**Figura 3.8** Grafico dei valori predetti rispetto ai valori osservati

In Figura 3.9 sono stati presi in considerazione i residui relativi a ciascun livello per i due fattori utilizzati. In entrambi i fattori si è rilevata una variabilità minore rispetto il livello più basso per essi considerato.

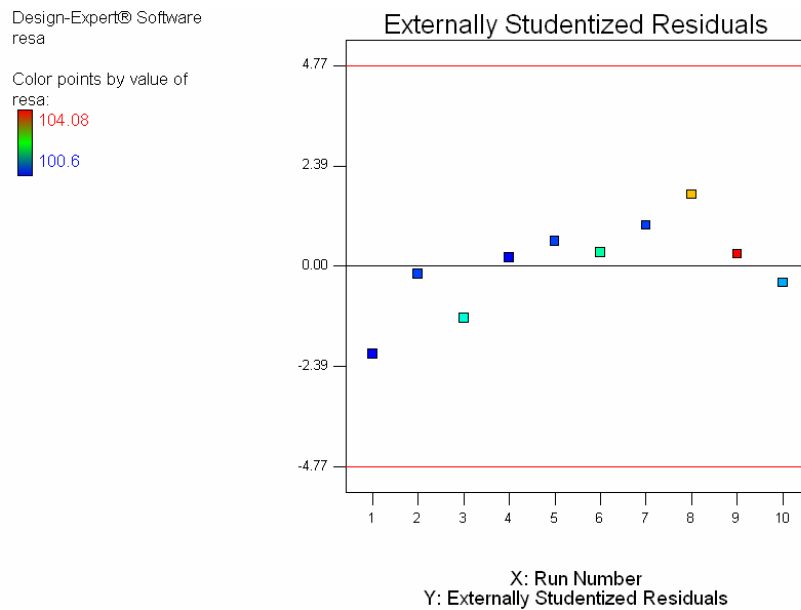


**Figura 3.9** Grafici dei residui rispetto ai livelli dei fattori

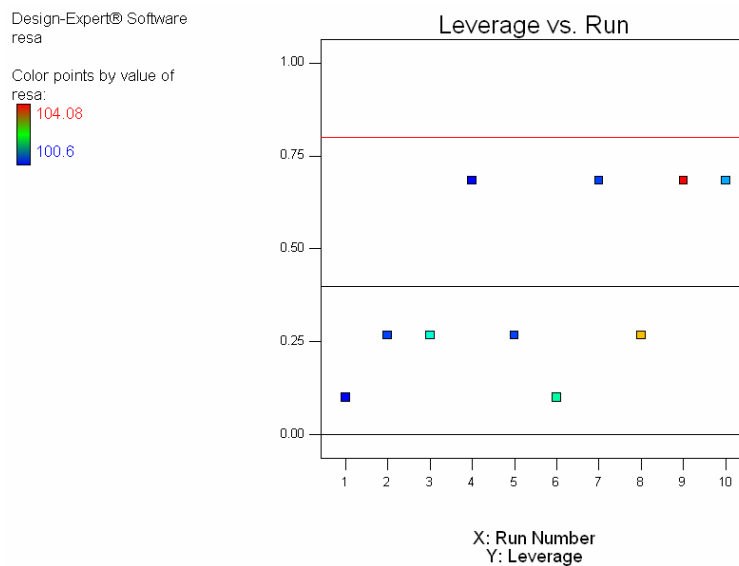
E' stato necessario, a questo punto, analizzare i grafici relativi alle misure di influenza (residui esternamente studentizzati, leverage, *DFFITs* e distanza di Cook).

I residui esternamente studentizzati costituiscono una misura di quanto il valore predetto differisce dal valore osservato quando il punto non è considerato nell'analisi. Il grafico dei residui esternamente studentizzati (Figura 3.10) è indicativo sia della presenza di valori non facilmente predetti dal modello, sia della presenza di punti particolarmente influenti. In figura sono stati riportati anche i limiti oltre i quali il valore è da considerare anomalo e come può essere visto non vi è la presenza di anomalie.

In Figura 3.11 è riportato il grafico dei leverage. Come precedentemente osservato non si sono presentati punti con leverage elevato, ovvero punti il cui valore superasse il doppio della media dei valori di leverage considerati, e ,dunque, non si sono ottenuti punti che avessero grande influenza sulla scelta del modello.



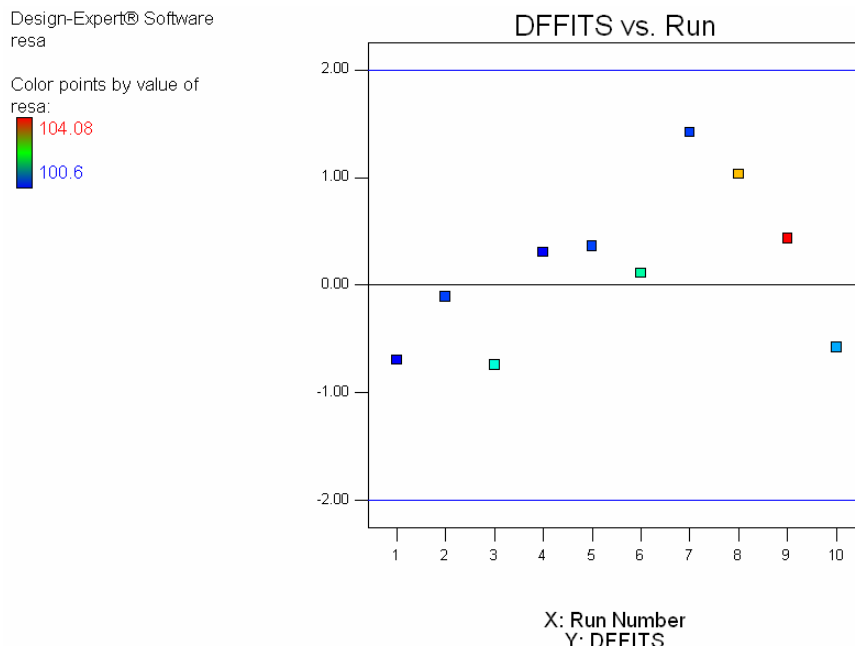
**Figura 3.10** Grafico dei residui esternamente studentizzati rispetto la sequenza temporale



**Figura 3.11** Grafico dei leverage rispetto la sequenza sperimentale.

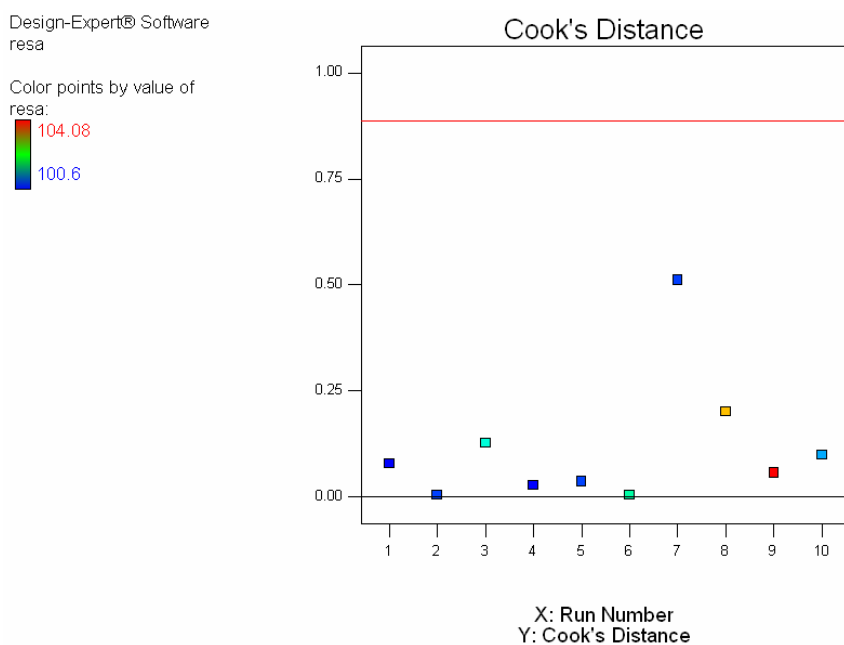
Ne grafico (Figura 3.12) del DFFITS si è misurata l'influenza dell'  $i$ -esima osservazione sul valore predetto, ed è riportata la differenza studentizzata tra il valore predetto e quello osservato con l' $i$ -esima osservazione e quello predetto senza l' $i$ -esima osservazione. Può essere accertata, anche qui, come non vi fosse la presenza di particolari irregolarità.





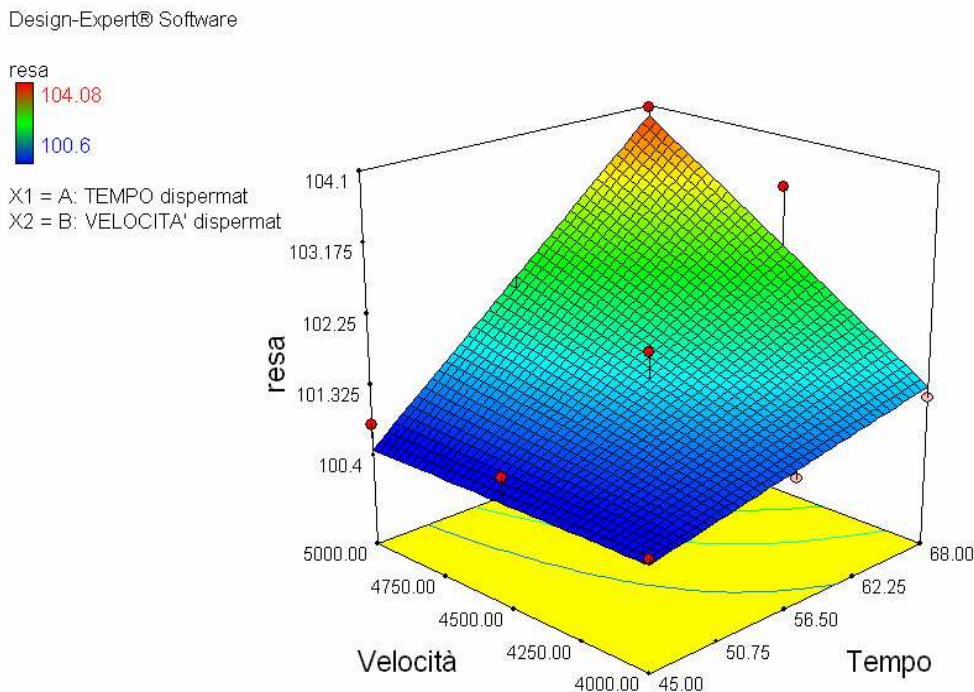
**Figura 3.12** Grafico dei DFFITS rispetto la sequenza sperimentale

L'ultimo grafico preso in esame (Figura 3.13) è stato quello associato alla distanza di Cook. Si è trattata di una misura della distanza al quadrato tra due stime: quella dei minimi quadrati per i coefficienti di regressione, basata su tutte le n osservazioni, e quella che si è ottenuta eliminando l'i-esimo punto.



**Figura 3.13** Grafico della distanza di Cook rispetto alla sequenza sperimentale

Dalla diagnostica è risultato che il modello può essere utilizzato per descrivere la risposta. Si è stati in grado, quindi, di descrivere l'andamento della risposta in funzione dei fattori considerati. Ciò viene effettuato generando la superficie di risposta riportata in Figura 3.14



**Figura 3.14** Superficie di risposta per la risposta “resa-laboratorio”

Dalla superficie si è potuto osservare che, nonostante l'impiego di un modello di secondo ordine, non si è riscontrata una curvatura marcata a riprova del fatto, peraltro già noto dall'analisi della varianza, che l'interazione tra i fattori non è stata significativa.

Dal grafico si è evidenziato come, entro il dominio sperimentale, la resa tendesse ad aumentare con il protrarsi del tempo di permanenza all'interno del sistema di raffinazione e a numero di giri più elevato.

Il risultato era aspettato: prolungando il tempo di permanenza dell'impasto nella macchina si incrementa la raffinazione. Nello stesso modo, aumentando il numero di giri, lo sforzo di taglio e gli impatti tra le microsferi e la massa da disperdere cresce, determinando anche in questo caso una migliore raffinazione.

### 3.5.2 Resa-Produzione

La procedura per determinare la significatività del modello, analizzare diagnosticamente i dati e definire la superficie di risposta è stata la medesima adottata in precedenza.

In questo caso si è suggerito di applicare un modello lineare ( Tabella 3.13 ).

**Tabella 3.13** Suggerimento del modello per la risposta “resa-produzione”

Response 1            resa            Transform:            None

\*\*\* WARNING: The Cubic Model is Aliased! \*\*\*

Sequential Model Sum of Squares [Type I]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Mean vs Total	1.093E+005	1	1.093E+005			
<u>Linear vs Mean</u>	<u>27.53</u>	<u>2</u>	<u>13.76</u>	<u>12.77</u>	<u>0.0046</u>	<u>Suggested</u>
2FI vs Linear	0.014	1	0.014	0.011	0.9182	
Quadratic vs 2FI	3.56	2	1.78	1.79	0.2780	
Cubic vs Quadratic	3.63	2	1.81	10.53	0.0868	Aliased
Residual	0.34	2	0.17			
Total	1.093E+005	10	10934.42			

\*Sequential Model Sum of Squares [Type I]: Select the highest order polynomial where the additional terms are significant and the model is not aliased.

Model Summary Statistics

Source	Std. Dev.	R-Squared	Adjusted R-Squared	Predicted R-Squared	PRESS	
<u>Linear</u>	<u>1.04</u>	<u>0.7849</u>	<u>0.7234</u>	<u>0.5665</u>	<u>15.20</u>	<u>Suggested</u>
2FI	1.12	0.7853	0.6779	0.3029	24.45	
Quadratic	1.00	0.8868	0.7463	-0.2029	42.19	
Cubic	0.42	0.9902	0.9558	-0.1119	39.00	Aliased

\*Model Summary Statistics\*: Focus on the model maximizing the "Adjusted R-Squared" and the "Predicted R-Squared".

D’altro canto, il modello che si è preso in considerazione è quadratico, dal momento che, dall’analisi della varianza, risultava essere significativo come si può vedere in Tabella 3.14 .

Dall’esame dei dati riportati è stato visto come il fattore tempo influenzasse notevolmente la resa e che gli altri termini del modello non fossero significativi.

Gli indici relativi alla capacità di fitting sono risultati elevati mentre quello relativo alla predizione è stato negativo. Ciò significa che la media dei valori osservati si è dimostrata più predittiva di qualsiasi altro modello accostabile ai dati.

Le prestazioni del modello e il calcolo delle equazioni sono riportate in Tabella 3.15 e 3.16.

**Tabella 3.14** ANOVA per la risposta “resa-produzione”

Source	Sum of Squares	df	Mean Square	F Value	p-value	Prob > F	
Model	31.10	5	6.22	6.27	0.0499		significant
A-TEMPO ricircolo	24.68	1	24.68	24.87	0.0076		
B-VEL girante	2.84	1	2.84	2.86	0.1659		
AB	0.014	1	0.014	0.015	0.9099		
A <sup>2</sup>	0.53	1	0.53	0.54	0.5050		
B <sup>2</sup>	3.38	1	3.38	3.40	0.1389		
Residual	3.97	4	0.99				
Lack of Fit	3.90	3	1.30	19.00	0.1667		not significant
Pure Error	0.068	1	0.068				
Cor Total	35.07	9					

The Model F-value of 6.27 implies the model is significant. There is only a 4.99% chance that a "Model F-Value" this large could occur due to noise.

Std. Dev.	1.00	R-Squared	0.8868
Mean	104.55	Adj R-Squared	0.7453
C.V. %	0.95	Pred R-Squared	-0.2029
PRESS	42.19	Adeq Precision	7.785

**Tabella 3.15** Determinazione dei coefficienti di regressione per la risposta “resa-produzione”

Factor	Coefficient		Standard Error	95% CI		VIF
	Estimate	df		Low	High	
Intercept	104.99	1	0.60	103.33	106.64	
A-TEMPO ricircolo	2.03	1	0.41	0.90	3.16	1.00
B-VEL girante	-0.69	1	0.41	-1.82	0.44	1.00
AB	0.060	1	0.50	-1.32	1.44	1.00
A <sup>2</sup>	0.48	1	0.65	-1.33	2.29	1.03
B <sup>2</sup>	-1.20	1	0.65	-3.01	0.61	1.03

**Tabella 3.16** Equazioni del modello per la risposta “resa-produzione”

Final Equation in Terms of Actual Factors:

$$\begin{aligned}
 \text{resa} = & \\
 & -122.60041 \\
 & -0.013988 * \text{TEMPO ricircolo} \\
 & +0.43963 * \text{VEL girante} \\
 & +6.66667\text{E-}006 * \text{TEMPO ricircolo} * \text{VEL girante} \\
 & +3.31349\text{E-}005 * \text{TEMPO ricircolo}^2 \\
 & -2.13841\text{E-}004 * \text{VEL girante}^2
 \end{aligned}$$

Dalla diagnostica del modello riportata in Figura 3.15, non sono risultate anomalie. Nel grafico di probabilità normale si è mostrato un andamento rettilineo, e non si sono riscontrati particolari andamenti nei grafici relativi ai residui.

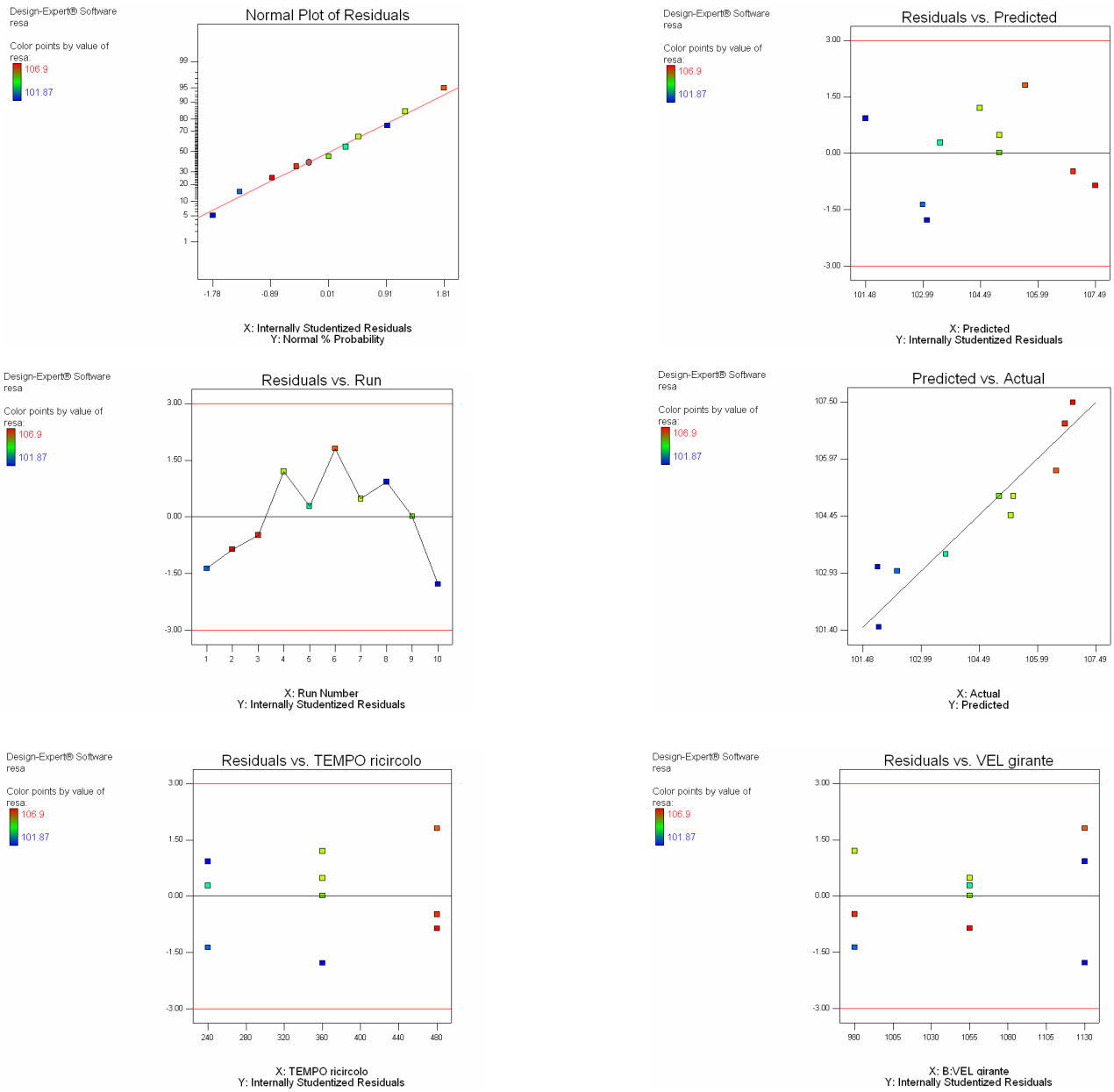
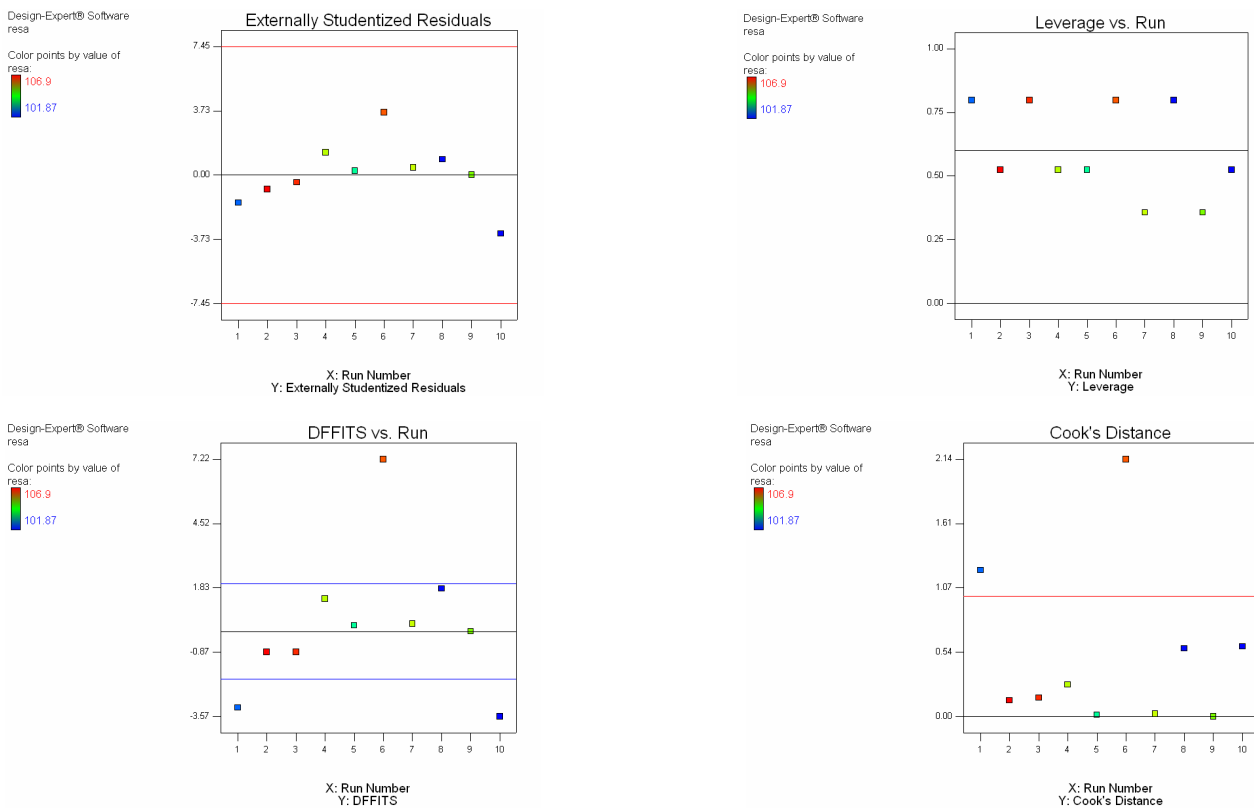


Figura 3.15 Diagnostica del modello per la risposta “resa-produzione”

Anche nei grafici sull'influenza (Figura 3.16) non sono stati rinvenuti evidenti punti che influenzino il modello.



**Figura 3.16** Grafici delle misure di influenza per la risposta "tempo"

Dopo aver validato il modello e dopo aver verificato l'adeguatezza a descriverne i dati sperimentali, si passa alla costruzione della superficie di risposta (Figura 3.17).

Dal grafico è emersa una situazione particolare in cui la resa massima si è ottenuta per valori intermedi dei fattori considerati. L'andamento trovato, differente da quello ottenuto in laboratorio, non è stata una sorpresa poiché esso è stato riscontrato anche in studi effettuati precedentemente.

Questo comportamento può essere spiegato, infatti, considerando che le condizioni con cui si è deciso di lavorare sono state molto esasperate e, quindi, a causa della velocità spinta della girante, il pacco di sfere all'interno del mulino tendeva a divenire solidale con il sistema girante, diminuendo la capacità di raffinazione del pigmento.

Design-Expert® Software



X1 = A: TEMPO ricircolo  
X2 = B: VEL girante

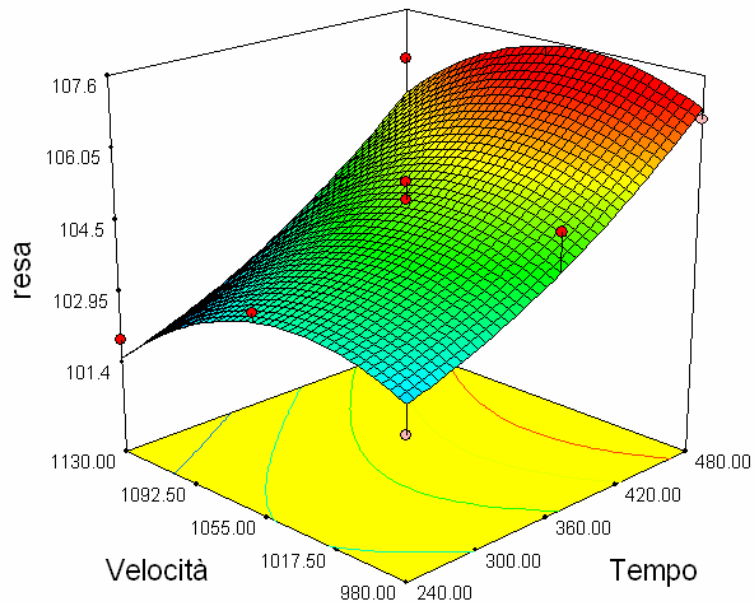


Figura 3.17 Superficie di risposta per la risposta “resa-produzione”

### 3.5.3 Viscosità-Laboratorio

Per questa risposta si è preferito utilizzare una trasformazione inversa dal momento che si è notato un aumento della deriva della risposta con la sequenza degli esperimenti e una mancanza di normalità da parte dei dati come è dato in Figura 3.18.

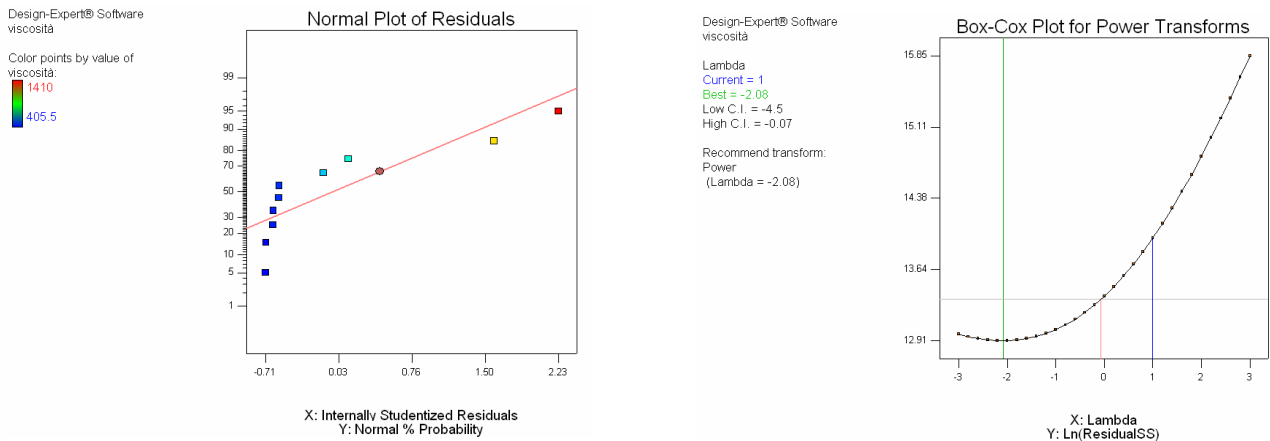


Figura 3.18 Grafico dei residui rispetto la sequenza sperimentale e della distribuzione normale

Si propone di utilizzare un modello lineare (Tabella 3.17)

**Tabella 3.17** Suggerimento del modello per la risposta “ viscosità-laboratorio”

Response 2 viscosità Transform: Inverse  
 \*\*\* WARNING: The Cubic Model is Aliased! \*\*\*

Sequential Model Sum of Squares [Type I]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Mean vs Total	3.472E-005	1	3.472E-005			
<u>Linear vs Mean</u>	<u>2.587E-006</u>	<u>2</u>	<u>1.294E-006</u>	<u>6.59</u>	<u>0.0246</u>	<u>Suggested</u>
2FI vs Linear	1.304E-008	1	1.304E-008	0.057	0.8185	
Quadratic vs 2FI	9.949E-007	2	4.974E-007	5.43	0.0725	
Cubic vs Quadratic	3.297E-007	2	1.648E-007	8.96	0.1004	Aliased
Residual	3.881E-008	2	1.841E-008			
Total	3.869E-005	10	3.869E-006			

"Sequential Model Sum of Squares [Type I]": Select the highest order polynomial where the additional terms are significant and the model is not aliased.

Model Summary Statistics

Source	Std. Dev.	R-Squared	Adjusted R-Squared	Predicted R-Squared	PRESS	
<u>Linear</u>	<u>4.431E-004</u>	<u>0.6531</u>	<u>0.5539</u>	<u>0.2420</u>	<u>3.003E-006</u>	<u>Suggested</u>
2FI	4.763E-004	0.6564	0.4845	-0.7589	6.968E-006	
Quadratic	3.027E-004	0.9075	0.7919	0.0513	3.758E-006	
Cubic	1.357E-004	0.9907	0.9582	0.8377	6.431E-007	Aliased

"Model Summary Statistics": Focus on the model maximizing the "Adjusted R-Squared" and the "Predicted R-Squared".

Si è scelto un modello quadratico poiché dall'analisi della varianza ( Tabella 3.18) si è desunto che esso è significativo, sebbene solo il fattore tempo presenti un p-value accettabile, mentre gli altri termini siano sopra al valore limite. La capacità di fitting si è evidenziata elevata, ma quella predittiva si è dimostrata molto scarsa.



**Tabella 3.18** ANOVA per la risposta “viscosità-laboratorio”

Response 2 viscosità  
 Transform: Inverse  
 ANOVA for Response Surface Quadratic Model  
 Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	3.595E-006	5	7.190E-007	7.85	0.0340	significant
A-TEMPO dispeme	2.472E-006	1	2.472E-006	26.98	0.0065	
B-VELOCITA' dispe	1.154E-007	1	1.154E-007	1.26	0.3245	
AB	1.304E-008	1	1.304E-008	0.14	0.7251	
A <sup>2</sup>	5.828E-007	1	5.828E-007	6.36	0.0652	
B <sup>2</sup>	2.557E-007	1	2.557E-007	2.79	0.1701	
Residual	3.665E-007	4	9.162E-008			
Lack of Fit	3.337E-007	3	1.112E-007	3.39	0.3752	not significant
Pure Error	3.281E-008	1	3.281E-008			
Cor Total	3.962E-006	9				

The Model F-value of 7.85 implies the model is significant. There is only a 3.40% chance that a "Model F-value" this large could occur due to noise.

Std. Dev.	3.027E-004	R-Squared	0.9075
Mean	1.863E-003	Adj R-Squared	0.7919
C.V. %	16.24	Pred R-Squared	0.0513
PRESS	3.758E-006	Adeq Precision	7.722

Successivamente sono stati determinati i coefficienti del modello di regressione (Tabella 3.19). Si é notato che l’intervallo di confidenza al 95% per il coefficiente relativo ai termini che non fossero significativi include 0, a conferma del fatto che il fattore velocità e la sua interazione hanno avuto un’influenza limitata sulla risposta in esame.

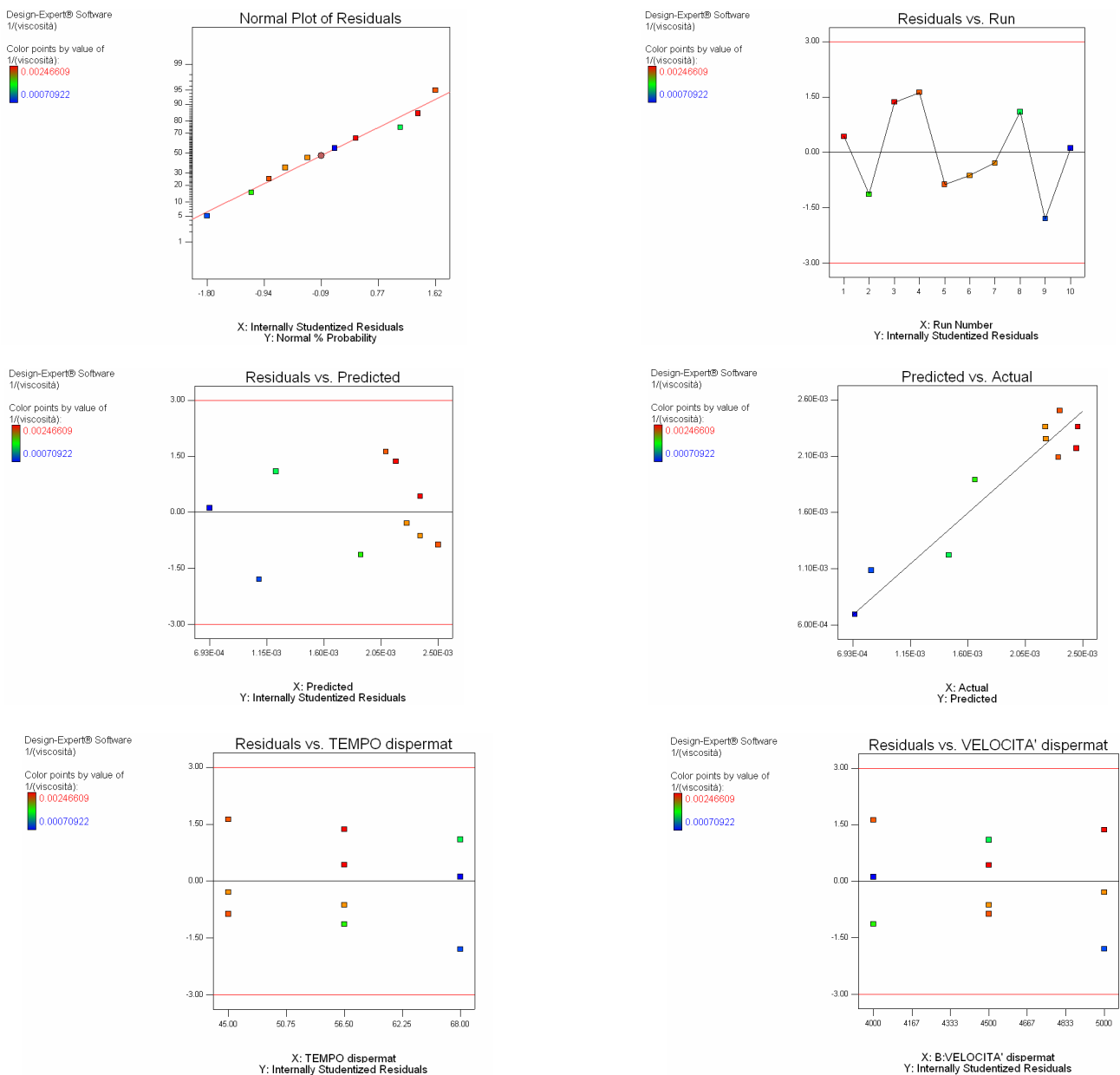
**Tabella 3.19** Calcolo delle equazioni del modello per la risposta “viscosità-laboratoio”

Factor	Coefficient		Standard Error	95% CI		VIF
	Estimate	DF		Low	High	
Intercept	2.362E-003	1	1.809E-004	1.860E-003	2.864E-003	
A-Tempo	-6.418E-004	1	1.236E-004	-9.849E-004	-2.987E-004	1.00
B-Velocità	1.387E-004	1	1.236E-004	-2.044E-004	4.818E-004	1.00
A <sup>2</sup>	-4.998E-004	1	1.982E-004	-1.050E-003	5.038E-005	1.03
B <sup>2</sup>	-3.311E-004	1	1.982E-004	-8.812E-004	2.191E-004	1.03
AB	5.710E-005	1	1.513E-004	-3.631E-004	4.773E-004	1.00

### Final Equation in Terms of Actual Factors:

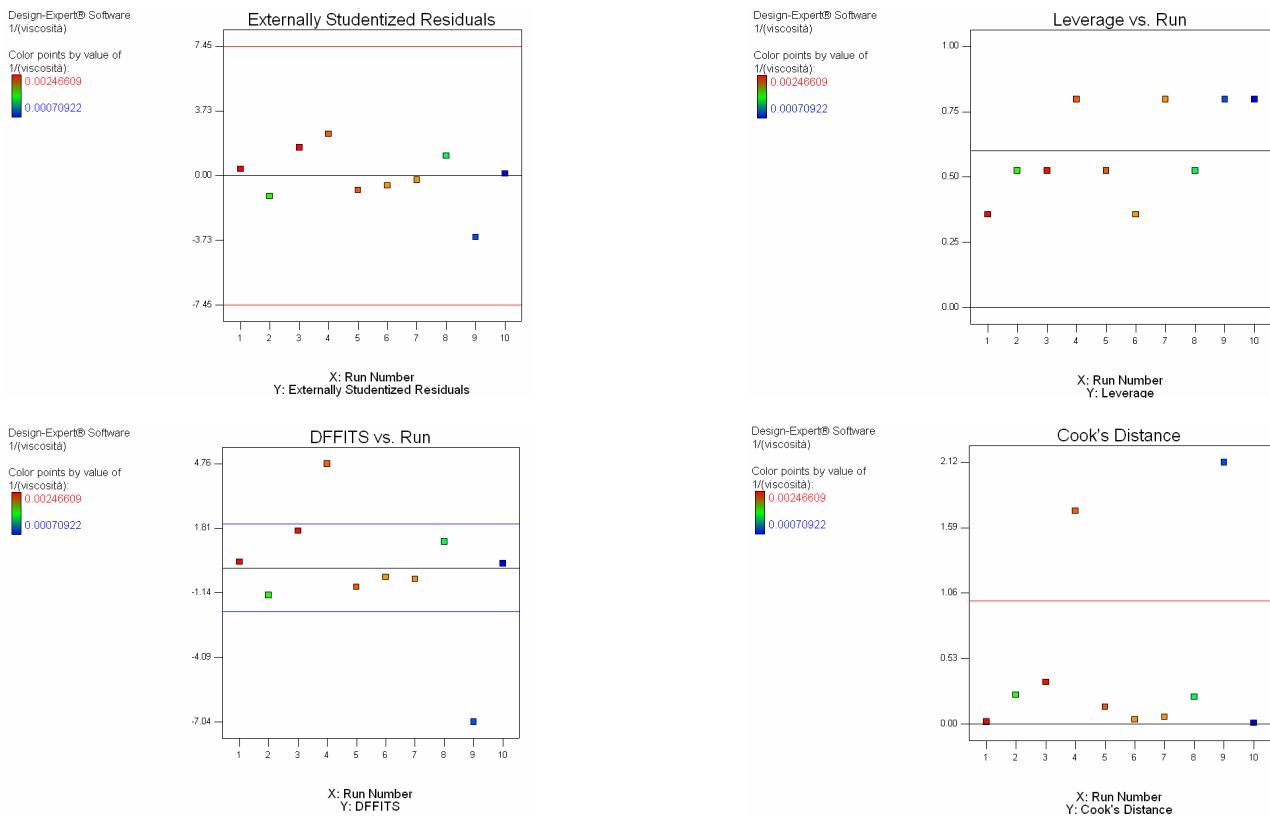
$$\begin{aligned}
 1.0/(\text{Viscosità}) = & \\
 & -0.032088 \\
 & +3.26539\text{E-}004 * \text{Tempo} \\
 & +1.16344\text{E-}005 * \text{Velocità} \\
 & -3.77912\text{E-}006 * \text{Tempo}^2 \\
 & -1.32423\text{E-}009 * \text{Velocità}^2 \\
 & +9.93111\text{E-}009 * \text{Tempo} * \text{Velocità}
 \end{aligned}$$

Si è passati quindi alla fase di diagnostica per verificare le assunzioni fatte sul modello. Dai grafici di Figura 3.19 si è osservato che i residui hanno seguito la distribuzione normale e non sono stati riscontrati andamenti sistematici o anomalie, segno di indipendenza dei residui e omogeneità della varianza.



**Figura 3.19** Diagnostica del modello per la risposta “Viscosità-laboratorio”

Anche nelle misure di influenza (Figura 3.20), non sono stati messi in evidenza punti influenti sul modello.



**Figura 3.20** Grafici delle misure di influenza per la risposta "Viscosità-laboratorio"

Dopo aver verificato l'adeguatezza del modello e l'influenza dei punti, si è passati alla costruzione della superficie di risposta (Figura 3.21). Come riscontrato nell'analisi della varianza, il grafico ha evidenziato l'influenza del tempo sulla viscosità.

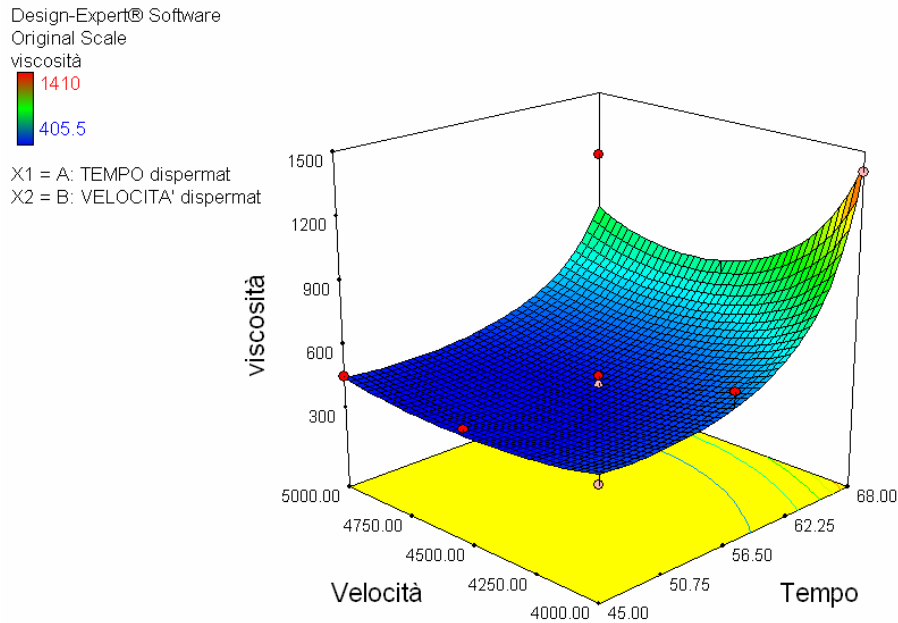


Figura 3.21 Superficie di risposta per la risposta “Viscosità-laboratorio”

### 3.5.4 Viscosità-Produzione

In Tabella 3.18 si è indicato di impiegare un modello lineare.

Tabella 3.18 Suggerimento del modello per la risposta “Viscosità-produzione”

Response 2 viscosità Transform: None  
\*\*\* WARNING: The Cubic Model is Aliased! \*\*\*

Sequential Model Sum of Squares [Type I]

Source	Sum of Squares	df	Mean Square	F Value	p-value	
Mean vs Total	7.762E+006	1	7.762E+006			
<u>Linear vs Mean</u>	<u>3.648E+005</u>	<u>2</u>	<u>1.824E+005</u>	<u>19.82</u>	<u>0.0013</u>	<u>Suggested</u>
2FI vs Linear	2256.25	1	2256.25	0.22	0.6572	
Quadratic vs 2FI	41470.43	2	20735.21	4.01	0.1107	
Cubic vs Quadratic	3980.83	2	1990.42	0.24	0.8075	Aliased
Residual	16700.82	2	8350.41			
Total	8.191E+006	10	8.191E+005			

\*Sequential Model Sum of Squares [Type I]\*: Select the highest order polynomial where the additional terms are significant and the model is not aliased.

Model Summary Statistics

Source	Std. Dev.	R-Squared	Adjusted R-Squared	Predicted R-Squared	PRESS	
<u>Linear</u>	<u>95.92</u>	<u>0.8499</u>	<u>0.8071</u>	<u>0.6753</u>	<u>1.394E+005</u>	<u>Suggested</u>
2FI	101.78	0.8552	0.7828	0.4499	2.361E+005	
Quadratic	71.91	0.9518	0.8916	0.6693	1.419E+005	
Cubic	91.38	0.9611	0.8249	-3.4174	1.896E+006	Aliased

\*Model Summary Statistics\*: Focus on the model maximizing the “Adjusted R-Squared” and the “Predicted R-Squared”.

Dall'analisi della varianza ( Tabella 3.19) si è constatato che anche il modello quadratico è significativo e che gli indici di fitting e di predizione sono risultati soddisfacenti.

**Tabella 3.19 ANOVA per la risposta “viscosità-produzione”**

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	4.085E+005	5	81708.07	15.80	0.0097	significant
A-TEMPO circolare	3.124E+005	1	3.124E+005	60.41	0.0015	
B-VEL girante	52453.50	1	52453.50	10.14	0.0334	
AB	2256.25	1	2256.25	0.44	0.5450	
A <sup>2</sup>	40700.01	1	40700.01	7.87	0.0485	
B <sup>2</sup>	3720.01	1	3720.01	0.72	0.4441	
Residual	20681.65	4	5170.41			
Lack of Fit	17401.15	3	5800.38	1.77	0.4933	not significant
Pure Error	3280.50	1	3280.50			
Cor Total	4.292E+005	9				

The Model F-value of 15.80 implies the model is significant. There is only a 0.97 % chance that a "Model F-Value" this large could occur due to noise.

Std. Dev.	71.91	R-Squared	0.9518
Mean	881.00	Adj R-Squared	0.8916
C.V. %	8.16	Pred R-Squared	0.6693
PRESS	1.419E+005	Adeq Precision	11.550

Si sono determinati quindi i coefficienti del modello di regressione e l'equazione del modello (Tabella 3.20).

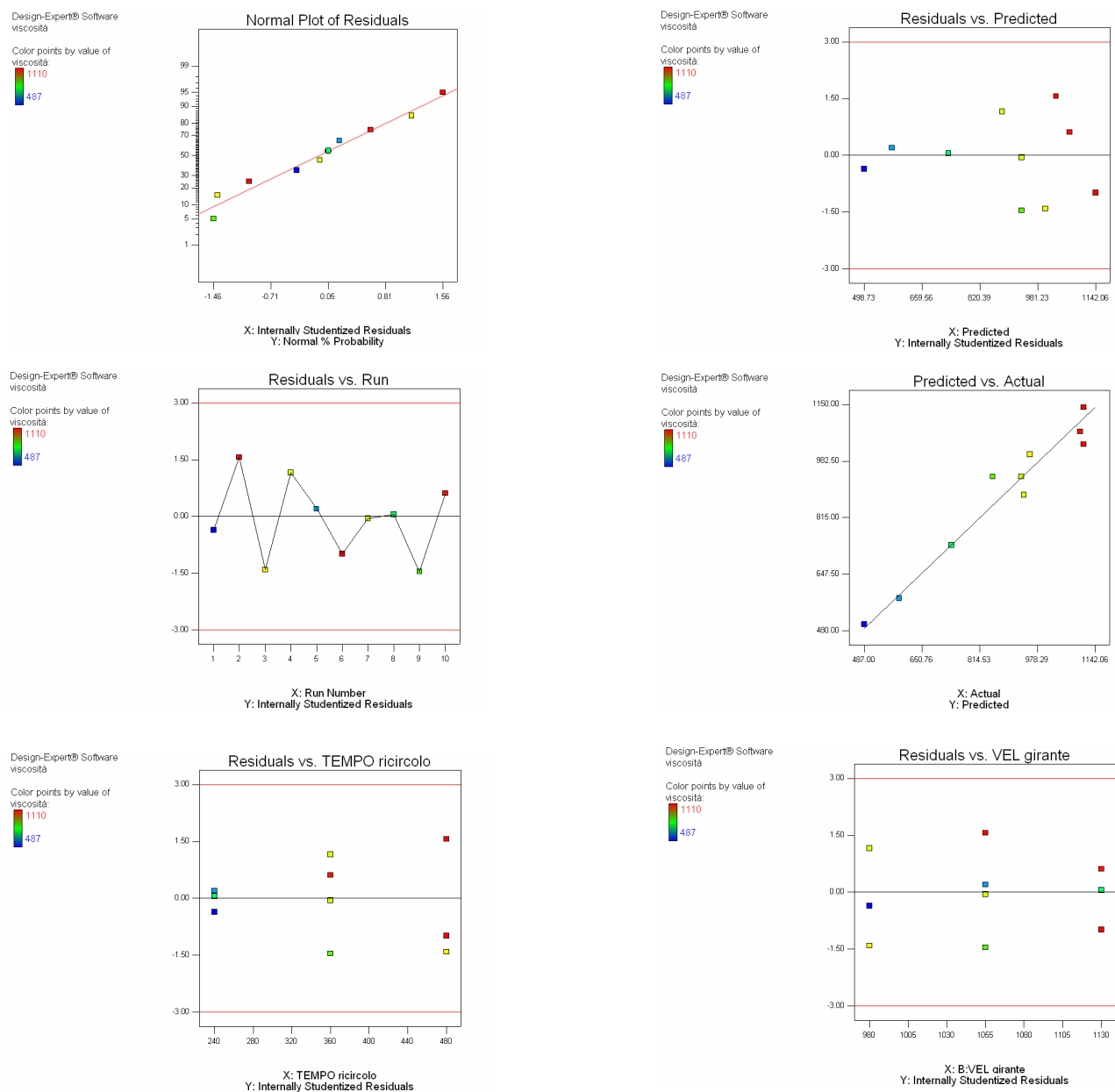
**Tabella 3.20** Calcolo delle equazioni del modello per la risposta “tempo”

Factor	Coefficient		Standard Error	95% CI		VIF
	Estimate	DF		Low	High	
Intercept	936.29	1	42.97	816.98	1055.59	
A-Tempo	228.17	1	29.36	146.66	309.67	1.00
B-Velocità	93.50	1	29.36	12.00	175.00	1.00
A <sup>2</sup>	-132.07	1	47.07	-262.77	-1.38	1.03
B <sup>2</sup>	39.93	1	47.07	-90.77	170.62	1.03
AB	-23.75	1	35.95	-123.57	76.07	1.00

### Final Equation in Terms of Actual Factors:

$$\begin{aligned}
 \text{Viscosità} = & \\
 & +4646.37032 \\
 & +11.28899 * \text{Tempo} \\
 & -12.78098 * \text{Velocità} \\
 & -9.17163\text{E-}003 * \text{Tempo}^2 \\
 & +7.09841\text{E-}003 * \text{Velocità}^2 \\
 & -2.63889\text{E-}003 * \text{Tempo} * \text{Velocità}
 \end{aligned}$$

Nella diagnostica non si è valutata nessuna anomalia rispetto alle assunzioni di normalità, indipendenza e costanza della varianza. (Figura 3.24)



**Figura 3.24** Diagnostica del modello per la risposta "Viscosità-produzione"

Anche nelle misure di influenza non si sono evidenziate situazioni particolari ( Figura 3.25).

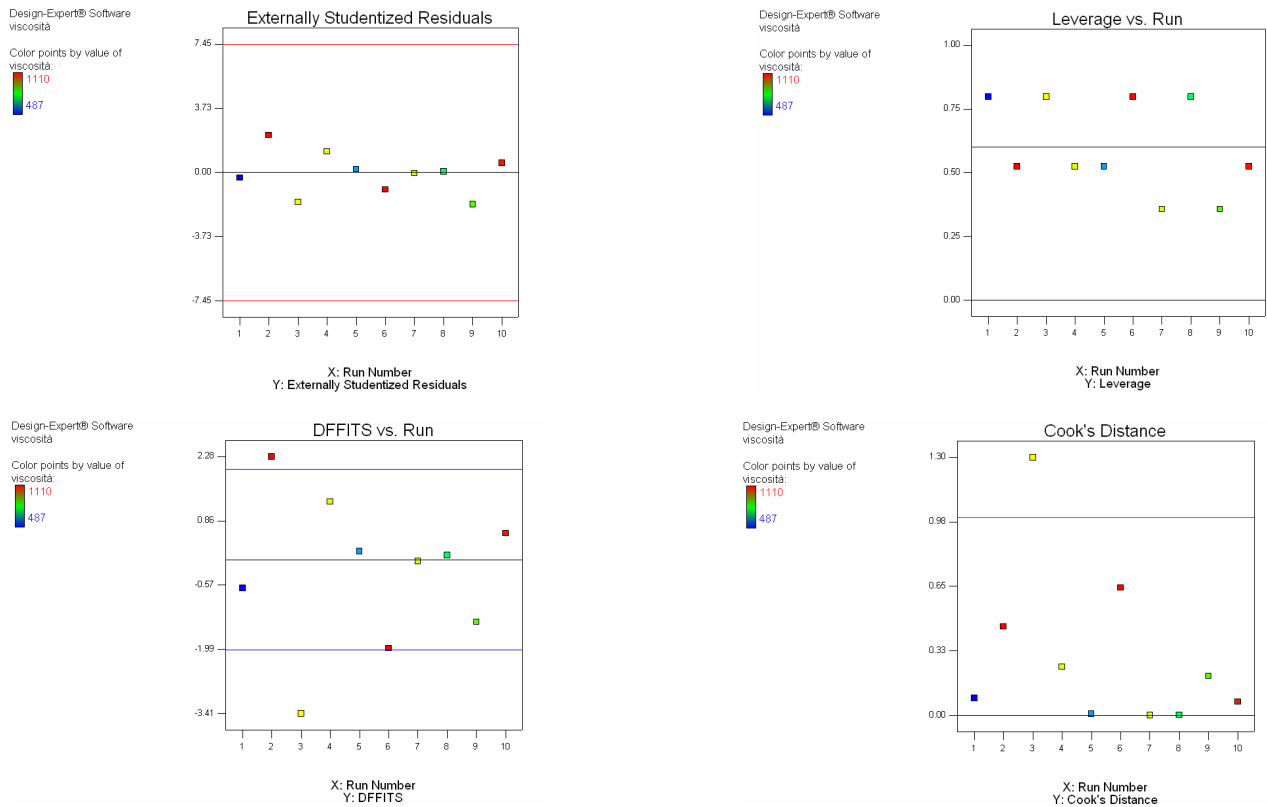


Figura 3.25 Grafici delle misure di influenza per la risposta "Viscosità-laboratorio"

Nella superficie di risposta realizzata (Figura 3.26) si è visto una analoga influenza del tempo rispetto all'esperimento effettuato in laboratorio, ma in questo caso si è evinto che la viscosità maggiore si ha per tempi elevati e velocità della girante massima.

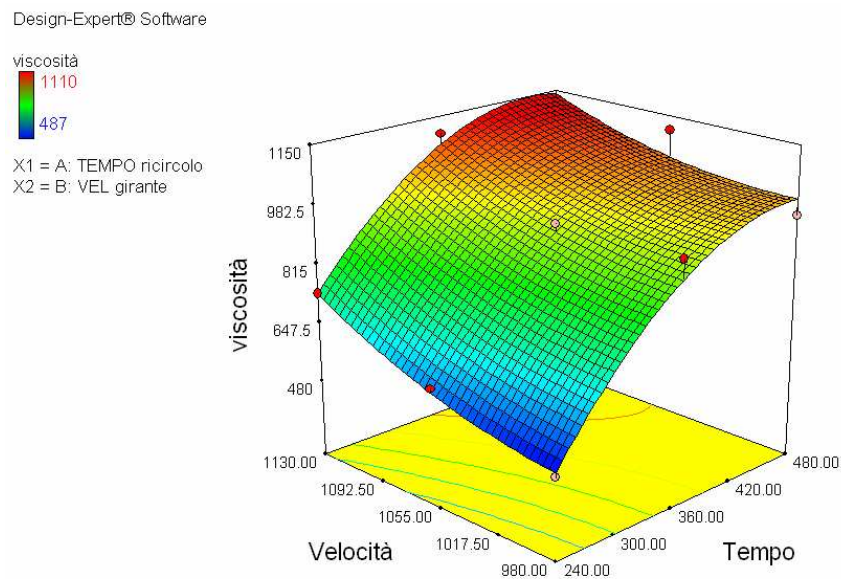


Figura 3.22 Superficie di risposta per la risposta "Viscosità-produzione"

### 3.6 Confronto fra le superfici di risposta

Uno degli obiettivi di questo studio è stato il confronto tra le superfici di risposta ottenute con gli esperimenti in laboratorio e produzione: si sono scelti due valori come target per le risposte relative all'esperimento del laboratorio e sono state riportate le condizioni operative migliori che massimizzassero la funzione di desiderabilità associata alle risposte stesse.

Dopo aver considerato come si riportano nel processo di industrializzazione, in maniera empirica, le condizioni operative di laboratorio, si è valutata la superficie di risposta ottenuta per l'esperimento di produzione relativamente ai valori di target ottenuti. I rispondenti predetti valori sono stati confrontati con i valori impostati come target.

Si è scelto come obiettivo una resa pari a 103,5 e una viscosità che variasse tra 600 e 1000  $mPa \cdot s$ .

I valori calcolati per i fattori che massimizzano la funzione di desiderabilità (Tabella 3.21) sono:

Tempo = 67,63min

Velocità = 4841 giri/min

**Tabella 3.21**

Constraints						
Name	Goal	Lower Limit	Upper Limit	Lower Weight	Upper Weight	Importance
Tempo	is in range	45	68	1	1	3
Velocità	is in range	4000	5000	1	1	3
Resa	is target = 103.5	100.6	104.08	1	1	3
Viscosità	is target = 800	405.5	1410	1	1	3

Solutions						
Number	Tempo	Velocità	Resa	Viscosità	Desirability	
1	<u>68.00</u>	<u>4822.45</u>	<u>103.5</u>	<u>762.753</u>	<u>0.975</u>	<u>Selected</u>

Con la metodologia empirica sopra descritta, si è trovato che al tempo di raffinazione di 67,63 minuti corrisponde, in produzione, un tempo di ricircolo di 338 minuti.

Con lo stesso metodo pratico, la velocità di 4841 giri al minuto equivale a 1094 giri al minuto.

A questo punto si sono introdotti i dati dei fattori del modello di produzione e si sono ottenuti come valori predittivi 103,9 per la resa e 952  $mPa \cdot s$  per la viscosità.

Dal confronto si evince la bontà della metodica adoperata per determinare le variabili operative nel processo di industrializzazione.



### 3.7 Check point

Per convalidare le conclusioni dell'analisi e verificare la bontà dei modelli ottenuti, sono state effettuate tre prove di controllo (*check point*). In particolare per la scelta delle condizioni operative si è tenuto conto di esplorare zone centrali del dominio sperimentale, per verificare sperimentalmente il modello e indagare la presenza di eventuali curvature e zone di confine nel dominio sperimentale in cui l'errore standard lungo il piano, cioè il rumore di fondo, è superiore. (come si può ravvisare in Figura 3.4). Sono stati quindi scelti i seguenti livelli per i fattori considerati:

- check point 1 : tempo di permanenza = 63 minuti  
velocità della girante = 4800 giri/min
- check point 2 : tempo di permanenza = 51 minuti  
velocità della girante = 4200 giri/min
- check point 3 : tempo di permanenza = 48 minuti  
velocità della girante = 4900 giri/min

**Tabella 3.22** Risultati ottenuti per i singoli check-points

#### Check Point 1 :

Resa = 102,5

Viscosità = 444

Factor	Name	Level	Low Level	High Level	Std. Dev.			
A	Tempo	63.00	45.00	68.00	0.000			
B	Velocità	4800.00	4000.00	5000.00	0.000			
	<b>Prediction</b>	<b>SE Mean</b>	<b>95% CI low</b>	<b>95% CI high</b>	<b>SE Pred</b>	<b>95% PI low</b>	<b>95% PI high</b>	
<b>Resa</b>	102.798	0.30	102.06	103.54	0.68	101.12	104.47	
<b>Viscosità</b>	631.363		483.62	909.07		365.23	2326.87	

#### Check Point 2 :

Resa = 100,2

Viscosità = 328

Factor	Name	Level	Low Level	High Level	Std. Dev.			
A	Tempo	51.00	45.00	68.00	0.000			
B	Velocità	4200.00	4000.00	5000.00	0.000			
	<b>Prediction</b>	<b>SE Mean</b>	<b>95% CI low</b>	<b>95% CI high</b>	<b>SE Pred</b>	<b>95% PI low</b>	<b>95% PI high</b>	
<b>Resa</b>	100.863	0.29	100.16	101.57	0.68	99.20	102.52	
<b>Viscosità</b>	479.115		391.61	616.98		309.21	1063.51	

### Check Point 3 :

Resa = 100,1

Viscosità = 457

Factor	Name	Level	Low Level	High Level	Std. Dev.			
A	Tempo	48.00	45.00	68.00	0.000			
B	Velocità	4900.00	4000.00	5000.00	0.000			
	<b>Prediction</b>	<b>SE Mean</b>	<b>95% CI low</b>	<b>95% CI high</b>	<b>SE Pred</b>	<b>95% PI low</b>	<b>95% PI high</b>	
<b>Resa</b>	100.899	0.38	99.97	101.83	0.72	99.13	102.67	
<b>Viscosità</b>	408.364		331.07	532.74		274.54	796.70	

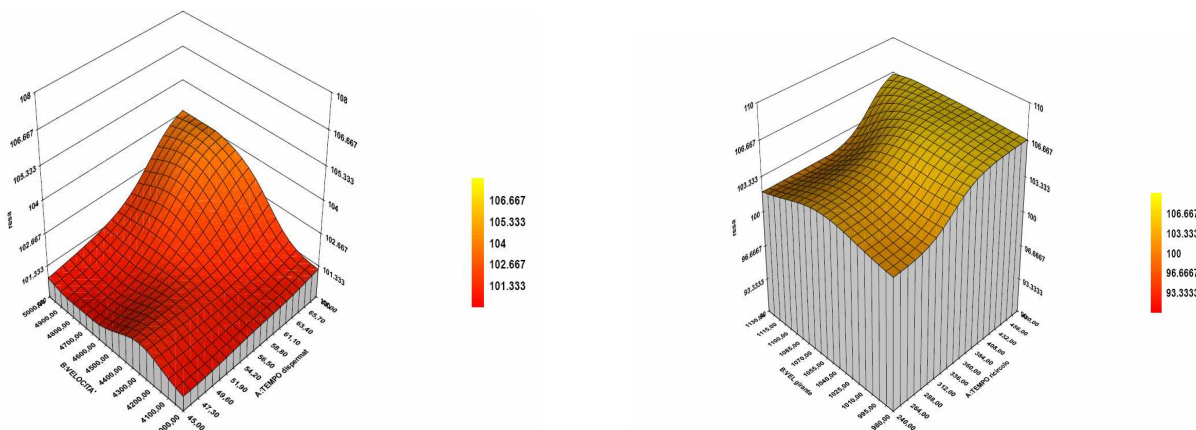
Dall'esame dei dati in Tabella 3.22 si può concludere che è confermata la capacità predittiva del modello. In ogni caso infatti i risultati sperimentali delle prove di conferma sono risultati all'interno del range di valori attesi con intervallo di confidenza del 95%. Da notare a questo proposito che nell'output del software si presentano due tipi di *standard error* e due tipi di intervalli di confidenza, denominati "95% CI" e "95% PI" dove *CI* sta per intervallo di confidenza, mentre *PI* per intervallo di predizione. Il primo tipo si riferisce all'intervallo entro il quale ci si aspetta di trovare la risposta media di un gruppo di prove; il secondo tipo si riferisce invece a valori individuali, cioè è l'intervallo che conterrà il valore vero di una singola osservazione con una probabilità del 95%.

### 3.8 Valutazione dei risultati con reti neurali

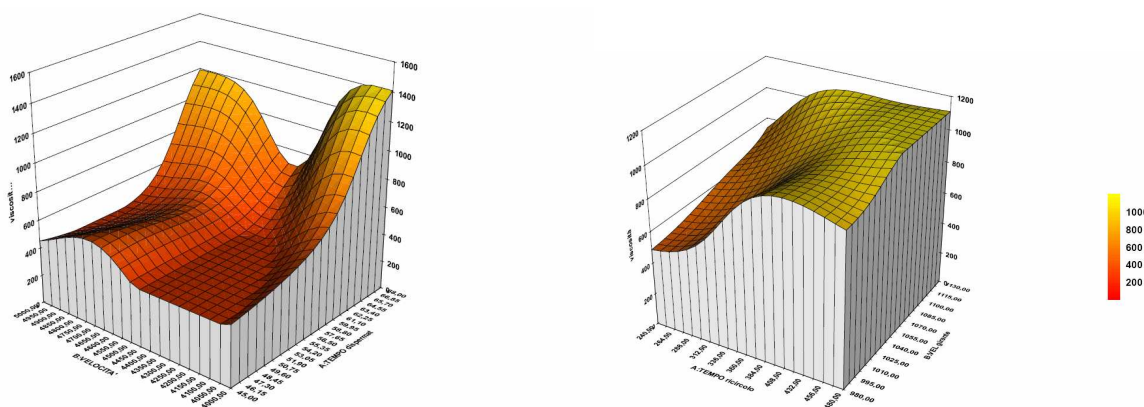
Si sono trattati i dati di input e di output per entrambi gli esperimenti con una metodologia di analisi basata su reti neurali. Il software di riferimento è stato NeuroShell 2<sup>®</sup> di Ward System Group, Inc.<sup>®</sup>. La struttura della rete utilizzata è una struttura multistrato ( 3 slab). L'architettura e l'algoritmo sono stati valutati in base all'analisi dei residui sulla risposta presente.

Le superfici di risposta ottenute (Figura 3.23 e 3.24) sono:

*Figura 3.23* Superfici di risposta per le risposte “Resa laboratorio e produzione”



*Figura 3.24* Superfici di risposta per le risposte “Viscosità laboratorio e produzione”



La valutazione della capacità predittiva, è stata effettuata tramite i tre check points relativi alla resa e alla viscosità delle prove effettuate in laboratorio, e si sono ottenuti i seguenti risultati ( Tabella 3.22 ).

**Tabella 3.22** Risultati dei check-points con reti neurali

	<b>TEMPO (min)</b>	<b>VELOCITA' (giri/min)</b>	<b>Actual Resa</b>	<b>Actual Viscosità</b>	<b>Network (Resa)</b>	<b>Network (Viscosità)</b>
Check 1	63	4800	102,5	444	103,7	892,9
Check 2	51	4200	100,2	328	101,0	405,5
Check 3	48	4900	100,1	457	100,8	462,5

Nella colonna “Actual” sono riportate le rese e le viscosità ricavate dalle prove, mentre in quella “Network” ci sono quelle calcolate dalla rete.

I risultati ottenuti dimostrano la bontà predittiva della rete soprattutto per la resa. La viscosità, invece, poiché risulta molto legata all’effetto di addensante associativo, presenta valori difficilmente prevedibili. I check points e la similitudine tra le superfici di risposta ottenute con i due differenti metodi, offrono una ulteriore conferma della validità di entrambi i metodi applicati.



## Conclusioni

Le considerazioni che si possono trarre al termine del lavoro si possono così riassumere: si è avuta conferma di quanto già osservato in altri esperimenti relativamente alla risposta della resa coloristica in funzione delle variabili operative considerate; particolarmente importante è l'evidente aumento dell'agitazione che non conduce necessariamente, all'interno del dominio sperimentale, a un continuo aumento della risposta. Per la viscosità, in questo esperimento adattabile ad un modello matematico (al contrario di quanto successo in precedenti lavori), l'andamento riscontrato è da considerare come "pura serie di dati scientificamente rilevanti" e necessita di ulteriori esperimenti per diverse paste pigmento, da confrontare ed analizzare. Peraltro la valutazione dei dati con metodologia basata su reti neurali, sia per le superfici di risposta ottenute che per la capacità predittiva verificata, alimenta i dubbi sull'influenza delle variabili operative sulla viscosità.

Altra importante considerazione riguarda il confronto tra esperimento condotto in "impianto pilota" in laboratorio e la fase di industrializzazione. Delle differenze tra le superfici di risposta ottenute il formulatore deve prenderne atto e può farne tesoro per le proprie attività; la metodologia empirica utilizzata per la "trasposizione" delle condizioni operative da laboratorio in produzione si è verificata essere effettivamente valida per l'esperimento condotto. Ultime considerazioni riguardano la metodologia di analisi basate sulle reti neurali, utilizzate come conferma ultima in questo esperimento. A differenza dei piani sperimentali trattati in questa tesi, il metodo, pur dimostrando buone capacità di fitting, presenta una limitata capacità predittiva; questa situazione si può spiegare a causa della bassa disponibilità di dati che si sono utilizzati nell'addestramento della rete la quale, quindi, risulta essere poco affinata.



## Riferimenti bibliografici

- Montgomery, D. C. (2005). *Progettazione e analisi degli esperimenti*, McGraw-Hill, Milano.
- Montgomery, D. C. , Ranger G.C. e Hubele N.F. (2006). *Statistica per ingegneria*, Egea, Milano.
- Savegnago, A. (2004). Progettazione di miscele per la produzione di vernici all'acqua. *Tesi di laurea in ingegneria chimica*, DIPIC, Università di Padova.
- Soliani, L. (2006). *Manuale di statistica per la ricerca e la professione*, Uninova, Parma.
- Todeschini, R. (1998). *Introduzione alla chemiometria*, EdiSES, Napoli.
- Tomba, E. (2006). Ottimizzazione di parametri di processo nella produzione di una pasta pigmento in dispersione acquosa. *Tesi di laurea in ingegneria chimica*, DIPIC, Università di Padova.