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)

OTTIMIZZAZIONE DI PARAMETRI DI PROCESSO NELLA PRODUZIONE DI UNA PASTA PIGMENTO IN DISPERSIONE ACQUOSA

Relatore: ing. Fabrizio Bezzo
Correlatore: dott. Daniele Foletto

Laureando: **EMANUELE TOMBA**

ANNO ACCADEMICO 2005-2006

Riassunto

In questa Tesi è stata sviluppata una procedura di progettazione di esperimenti (*design of experiments*, DOE) per l'ottimizzazione dei parametri macchina relativamente al processo di produzione di una pasta pigmento in dispersione acquosa. Il primo passo del percorso seguito è stato la generazione di un numero limitato di punti sperimentali, servendosi di un disegno fattoriale generale a blocchi. Il passo successivo è stato l'ottenimento dei dati sperimentali relativi ai quattro output (risposte) scelte come oggetto di indagine per ognuno dei punti sperimentali. Servendosi delle informazioni raccolte, si è poi passati all'individuazione, attraverso elaborazioni statistiche, dei modelli matematici che descrivono l'andamento delle risposte stesse in tutto il dominio del piano sperimentale .
Per il controllo dei modelli ottenuti si sono sfruttati tre punti sperimentali addizionali (*check point*). Ultima fase della Tesi è stata l'ottimizzazione delle variabili di processo applicando una serie di vincoli su ciascuna delle quattro risposte considerate, con indice di importanza diverso per ciascun vincolo.

Il lavoro descritto in questa Tesi è stato svolto durante un tirocinio di 300 ore 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

Introduzione

Nella seguente Tesi si è cercato di verificare l'applicabilità di tecniche di *programmazione di esperimenti* (DOE) nella produzione di paste pigmento in dispersione acquosa, nell'ambito della razionalizzazione dell'attività di sviluppo di processo. Si è pertanto utilizzato un disegno fattoriale a blocchi per ottimizzare i parametri macchina in una lavorazione comunemente condotta in SAMIA s.a.s..

Dopo i buoni risultati ottenuti in precedenti lavori (Savegnago, 2004), finalizzati a esplorare e/o ottimizzare il piano sperimentale nella formulazione di miscele, il naturale passo successivo è stato l'estensione dell'impiego di tecniche DOE su scala industriale.

L'obiettivo finale è quello di giungere a gestire e sfruttare strutture complesse di dati in modo da estrarne informazioni in grado di abbattere i costi e migliorare le performance qualitative di un determinato processo.

Il lavoro ha anche permesso di valutare la versatilità dell'impianto pilota e la sua effettiva funzionalità nell'ottimizzazione di processo, anche se i dati monitorati e registrati, anche relativi a variabili passive, devono ancora essere appieno studiati e interpretati.

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, l'utilizzo di blocchi 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), Smith (2005), Todeschini ((2003) e al Manuale Unichim N.186 (1998); l'elaborazione dei dati per la parte sperimentale è stata condotta utilizzando il software Design-Expert (Vers. 7.0) di Stat-Ease®. L'attività di progettazione e analisi dei dati è stata svolta nel laboratorio chimico SAMIA s.a.s..

Capitolo 1

Presentazione dell'azienda

1.1 Nota introduttiva

Il distretto industriale di Arzignano, Chiampo, Montorso, Montebello e Zermeghedo, situato nella Valle del Chiampo in provincia di Vicenza, è oggi il maggiore polo conciario europeo. La concia nel Vicentino attualmente conta 1077 aziende, 11644 addetti e un fatturato 2002 di 3170 milioni di euro, con la maggior parte delle aziende dislocate in questo distretto.
La lavorazione del cuoio si divide 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 rifinizione, 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 rifinizione.

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 comprende:

- la Sede di Arzignano (VI), (oltre 14500 m^2 coperti) in cui sono concentrate le attività amministrative, commerciali, tecnico-produttive e di laboratorio;
- il deposito di S. Croce sull'Arno (PI), attraverso il quale si sviluppa la distribuzione dei prodotti standard e l'assistenza dei clienti in loco.

Inoltre è presente 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, emulsionamento, 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 t/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.

Come già anticipato, 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

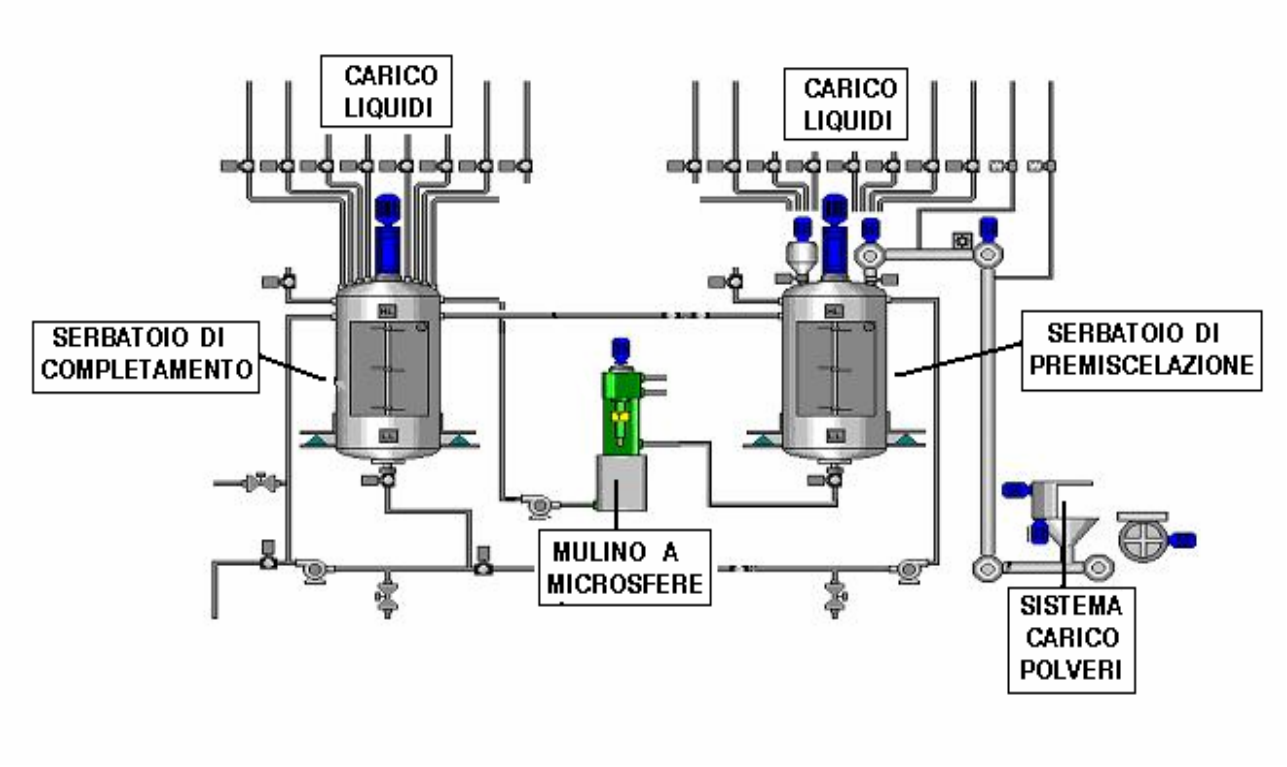

Figura 1.1 *Schema di impianto per la produzione di dispersioni di pigmento*

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), a 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 microsfere 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 raffinante 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. É 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 $x_1, x_2, ..., x_p$ sono controllabili, mentre altre variabili $z_1, z_2, ..., z_q$, 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 $z_1, z_2, ..., z_q$ 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. L'approccio generale per pianificare e condurre l'esperimento è detto **strategia sperimentale**; ve ne sono diverse a disposizione dello sperimentatore, alcune delle quali sono:

- Approccio **a tentativi**
- Approccio **un-fattore-alla-volta**
- Approccio **fattoriale**.

L'approccio a tentativi consiste nello scegliere una combinazione arbitraria dei fattori coinvolti, condurre l'esperimento ed osservare cosa accade. Questa strategia, usata frequentemente da ingegneri e ricercatori, spesso funziona, magari anche ragionevolmente bene, quando gli sperimentatori dispongono di un rilevante bagaglio di conoscenze tecniche o

teoriche del sistema che stanno studiando e di una notevole esperienza pratica. Ci sono però almeno due svantaggi nell'approccio a tentativi. Primo, si supponga che il tentativo iniziale non produca i risultati sperati; ora lo sperimentatore deve fare un altro tentativo per tentare di indovinare la corretta combinazione dei livelli dei fattori. Questi tentativi potrebbero continuare a lungo senza alcuna garanzia di successo. Secondo, si supponga che il tentativo iniziale produca un risultato accettabile. Lo sperimentatore sarà tentato di non procedere oltre con le prove, sebbene non ci sia alcuna garanzia che sia stata individuata la soluzione migliore.

L'approccio un-fattore-alla-volta consiste nello scegliere un valore iniziale, o insieme di livelli **di base**, per ciascun fattore; quindi far variare in successione i livelli di ciascun fattore nel proprio campo di variazione, mantenendo gli altri fattori costanti al loro livello base. Dopo aver eseguito tutte le prove, di solito si costruisce un certo numero di grafici, con l'intento di evidenziare come la variabile di risposta sia influenzata dal variare di ciascun 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. Molti non se ne rendono conto, pertanto gli esperimenti un-fattore-alla-volta sono frequentemente eseguiti in pratica. Gli esperimenti un-fattore-alla-volta sono sempre meno efficienti di altri, eseguiti seguendo metodi basati su un approccio statistico alla progettazione.

L'approccio corretto, per condurre esperimenti con più fattori, consiste nel realizzare un esperimento **fattoriale**, secondo una strategia sperimentale in cui i fattori variano *congiuntamente* invece che uno alla volta.

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 replicazioni 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*. Se il fattore è presente con più di due livelli, tale procedura può essere modificata poiché ci sono altri modi di definire l'effetto di un fattore.

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 (Figura 2.1), 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.

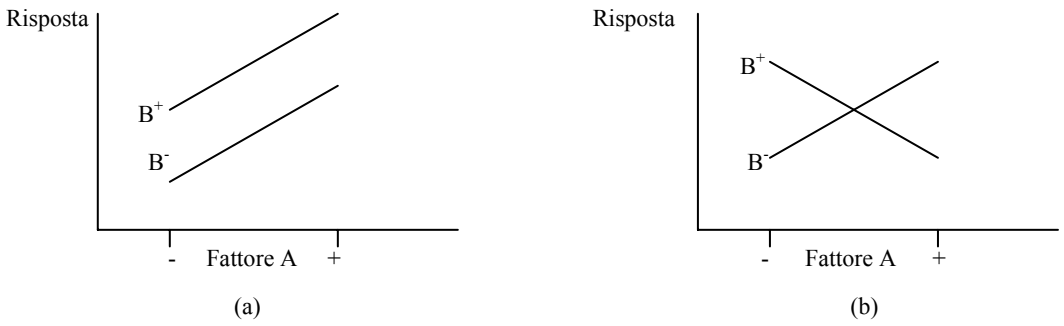

Figura 2.1 *Esperimenti fattoriali senza interazione* (a) *e con interazione* (b)

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:

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

dove y è la risposta, β 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 ε 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,…,a) ed il fattore B è al j-esimo livello (j = 1,2,…,b) per la k-esima replicazione (k = 1,2,…,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**. La **casualizzazione** è la pietra angolare su cui si fonda l'uso dei metodi statistici nella pianificazione sperimentale. Con tale termine si intende che sia l'allocazione del materiale sperimentale sia l'ordine con cui vengono eseguite le singole prove, vengono stabiliti in modo casuale. I metodi statistici richiedono che le osservazioni (o meglio gli errori) siano variabili casuali indipendenti; la casualizzazione di regola rende valida questa assunzione. Inoltre casualizzando in modo appropriato le prove sperimentali, vengono "mediati" anche gli effetti di fattori estranei eventualmente presenti.

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

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

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**:

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \qquad \begin{cases} i = 1,2,...,a \\ j = 1,2,...,b \\ k = 1,2,...,n \end{cases} \tag{2.1}$$

dove μ è l'effetto medio generale, τ_i è l'effetto dell'*i*-esimo livello del fattore di riga *A*, β_j è l'effetto del *j*-esimo livello del fattore di colonna *B*, $(\tau\beta)_{ij}$ è l'effetto dell'interazione tra τ_i e β_j e ε_{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**:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \qquad \begin{cases} i = 1,2,...,a \\ j = 1,2,...,b \\ k = 1,2,...,n \end{cases} \tag{2.2}$$

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 verificare l'eguaglianza di effetti di trattamenti di riga o di colonna. Ciò viene fatto attraverso la formulazione di **ipotesi** precise. Un'ipotesi statistica è un'affermazione sui parametri di un modello. L'ipotesi riflette alcune congetture che ci si pongono sul problema. In questo caso le ipotesi saranno del tipo:

$$\begin{aligned} &H_0 :\ \tau_1 = \tau_2 \ldots = \tau_a = 0 \\ &H_1 : \text{almeno un } \tau_i \neq 0, \end{aligned} \qquad (2.3)$$

dove H_0 è detta ipotesi nulla mentre H_1 è l'ipotesi alternativa. Le ipotesi per verificare l'eguaglianza di effetti di trattamento di colonna sono invece del tipo:

$$\begin{aligned} &H_0 :\ \beta_1 = \beta_2 \ldots = \beta_b = 0 \\ &H_1 : \text{almeno un } \beta_j \neq 0. \end{aligned} \qquad (2.4)$$

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

$$\begin{aligned} &H_0 :\ (\tau\beta)_{ij} = 0 \text{ per tutti } i, j \\ &H_1 : \text{almeno un } (\tau\beta)_{ij} \neq 0. \end{aligned} \qquad (2.5)$$

Queste ipotesi vengono verificate usando **l'analisi della varianza (ANOVA)**.

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:

$$
\begin{aligned}
y_{i..} &= \sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk} & \bar{y}_{i..} &= \frac{y_{i..}}{bn} & i &= 1,2,\ldots,a \\
y_{.j.} &= \sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk} & \bar{y}_{.j.} &= \frac{y_{.j.}}{an} & j &= 1,2,\ldots,b \\
y_{ij.} &= \sum_{k=1}^{n} y_{ijk} & \bar{y}_{ij.} &= \frac{y_{ij.}}{n} & i &= 1,2,\ldots,a; \quad j = 1,2,\ldots,b \\
y_{...} &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk} & \bar{y}_{...} &= \frac{y_{...}}{abn}
\end{aligned}
\tag{2.6}
$$

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 ordinario per misurare la variabilità in modo naturale. Essa può essere scritta come:

$$
\begin{aligned}
&\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{...})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}[(\bar{y}_{i...}-\bar{y}_{...})+(\bar{y}_{.j.}-\bar{y}_{...})+(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})+(y_{ijk}-\bar{y}_{ij.})]^2 = \\
&= bn\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})^2 + an\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})^2 + n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij.})^2,
\end{aligned}
\tag{2.7}
$$

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 "colonne" 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 sul lato destro della (2.7), si deduce che ci devono essere almeno due replicazioni per ottenere una somma di quadrati dell'errore. La (2.7) può essere scritta in simboli come:

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E. \tag{2.8}$$

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 (2.8) 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:

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

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

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

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

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 di σ^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 (2.1) sia adeguato e che i termini di errore ε_{ijk} siano distribuiti normalmente e indipendentemente con varianza costante σ^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.

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 σ^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:

$$e_{ijk} = y_{ijk} - \hat{y}_{ijk} \tag{2.9}$$

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 (2.9) diventa:

$$e_{ijk} = y_{ijk} - \bar{y}_{ij.} \tag{2.10}$$

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.

É 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 i veri livelli di significatività 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. 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 comprendente l'*outlier* e una senza. Un controllo grossolano può essere effettuato esaminando i **residui standardizzati**:

$$d_{ij} = \frac{e_{ij}}{\sqrt{MS_E}} \ . \tag{2.11}$$

Se gli errori ε_{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 ± 1, circa

il 95 percento dovrebbe cadere entro ± 2 e quasi tutti quanti dovrebbero cadere entro ± 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 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 prevenire il problema possibilmente all'atto della raccolta dei 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). Viene scelta la trasformazione che produce il grafico dei residui con una forma più soddisfacente.

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. La curva caratteristica operativa è un grafico che rappresenta la probabilità di accettare l'ipotesi nulla nonostante questa sia falsa, ossia di assumere erroneamente che gli effetti dei fattori sulla risposta osservata siano nulli (errore di II specie o β). Tale valore di probabilità è funzione di un parametro Φ dipendente dagli effetti dei trattamenti e delle interazioni, che misura quanto l'ipotesi nulla sia lontana dalla realtà, per una certa dimensione campionaria. Queste curve sono utilizzate come guida nella scelta del numero di replicazioni, in modo tale da rendere il piano capace di riconoscere eventuali differenze tra i trattamenti (potenza del piano). Nell'uso delle curve caratteristiche allo sperimentatore viene richiesto di dare uno specifico valore al parametro Φ. Un modo molto efficace è trovare il più piccolo valore di Φ^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 Φ^2 è:

$$\Phi^2 = \frac{nbD^2}{2a\sigma^2} \tag{2.12}$$

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

$$\Phi^2 = \frac{naD^2}{2b\sigma^2} \tag{2.13}$$

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

$$\Phi^2 = \frac{nD^2}{2\sigma^2[(a-1)(b-1)+1]} \tag{2.14}$$

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 β 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 σ per determinare l'effetto di una stima errata di questo parametro sulla sensibilità del piano.

2.3.4 L'assunzione d'assenza d'interazione nel modello a due fattori

Talvolta lo sperimentatore ritiene che sia appropriato un modello a due fattori senza interazione, ossia:

$$Y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk} \qquad \begin{cases} i = 1,2,...,a \\ j = 1,2,...,b \\ k = 1,2,...,n \end{cases} . \tag{2.15}$$

Si dovrebbe tuttavia essere molto cauti nello scartare termini di interazione, poiché la presenza d'interazioni significative può avere effetti marcati sull'interpretazione dei dati. Tuttavia dall'analisi dei residui, qualunque comportamento sistematico in queste quantità suggerisce la presenza di interazione.

2.3.5 Un'osservazione per cella

Talvolta si incontrano esperimenti a due fattori con una **singola replicazione**, cioè una sola osservazione per cella. Se ci sono due fattori e solo un'osservazione per cella, il modello è:

$$Y_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ij} \qquad \begin{cases} i = 1,2,...,a \\ j = 1,2,...,b \end{cases} . \tag{2.16}$$

Dall'esame delle medie attese delle somme dei quadrati si deduce che la varianza dell'errore σ^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:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \qquad \begin{cases} i = 1,2,...,a \\ j = 1,2,...,b \end{cases} . \tag{2.17}$$

Se il modello è appropriato allora il quadrato medio del residuo è uno stimatore non distorto di σ^2, cioè una statistica il cui valore atteso dovrebbe essere uguale al valore del parametro stimato. Gli effetti principali possono quindi essere sottoposti a test confrontando MS_A e MS_B con $MS_{Residuo}$.

2.4 Piano fattoriale generale

I risultati del piano fattoriale a due fattori possono essere estesi al caso generale, in cui ci sono a livelli del fattore A, b livelli del fattore B, c livelli del fattore C e così via, disposti in un esperimento fattoriale. In generale ci saranno $a \cdot b \cdot c \cdot \ldots \cdot n$ osservazioni totali se ci sono n replicazioni dell'esperimento completo. Si noti ancora una volta che, se tutte le interazioni possibili sono incluse nel modello, è necessario avere almeno due replicazioni ($n \geq 2$) per determinare una somma di quadrati dovuta all'errore.

Se tutti i fattori dell'esperimento sono fissi, si possono facilmente formulare e valutare le ipotesi sugli effetti principali e sulle interazioni. Per un modello ad effetti fissi, le statistiche test per ciascun effetto principale e ciascuna interazione possono essere costruite dividendo il corrispondente quadrato medio dell'effetto o dell'interazione per il quadrato medio dell'errore. Tutti questi test F sono variabili casuali indipendenti che seguono la distribuzione F di Fisher. Ognuno di essi è pertanto confrontato con il valore $F_{\alpha,\upsilon_1,\upsilon_2}$ che rappresenta il percentile della distribuzione F con v_1 gradi di libertà al numeratore e v_2 al denominatore che lascia alla propria destra un'area pari ad α (valori tabulati). Il numero dei gradi di libertà per ogni effetto principale è dato dal numero dei livelli del fattore meno uno ed il numero di gradi di libertà per un'interazione è dato dal prodotto del numero di gradi di libertà associati con le componenti individuali dell'interazione.

Per esempio si consideri il **modello d'analisi della varianza a tre fattori**:

$$Y_{ijkl} = \mu + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{ij} + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + (\tau\beta\gamma)_{ijk} + \varepsilon_{ijkl} \qquad \begin{cases} i = 1,2,\ldots,a \\ j = 1,2,\ldots,b \\ k = 1,2,\ldots,c \\ l = 1,2,\ldots,n \end{cases} . \qquad (2.18)$$

Assumendo che A, B e C siano fissi, la **tabella d'analisi della varianza** è mostrata in Tabella 2.3. I test F sugli effetti principali e sulle interazioni seguono direttamente dai valori attesi dai quadrati medi.

Tabella 2.3 *Tabella d'analisi della varianza per il modello ad effetti fissi a tre fattori.*

Origine della variabilità	Somma dei quadrati	Gradi di libertà	Quadrati medi	Valori attesi dei quadrati medi	F_0
A	SS_A	$a-1$	MS_A	$\sigma^2+\frac{bcn\sum\tau_i^2}{a-1}$	$F_0=\frac{MS_A}{MS_E}$
B	SS_B	$b-1$	MS_B	$\sigma^2+\frac{acn\sum\beta_j^2}{b-1}$	$F_0=\frac{MS_B}{MS_E}$
C	SS_C	c - 1	MS_C	$\sigma^2+\frac{abn\sum\gamma_k^2}{c-1}$	$F_0=\frac{MS_C}{MS_E}$
AB	SS_{AB}	$(a$ - 1$)(b-1)$	MS_{AB}	$\sigma^2+\frac{cn\sum\sum(\tau\beta)_{ij}^2}{(a-1)(b-1)}$	$F_0=\frac{MS_{AB}}{MS_E}$
AC	SS_{AC}	$(a-1)(c-1)$	MS_{AC}	$\sigma^2+\frac{bn\sum\sum(\tau\gamma)_{ik}^2}{(a-1)(c-1)}$	$F_0=\frac{MS_{AC}}{MS_E}$
BC	SS_{BC}	$(b$ - 1$)(c-1)$	MS_{BC}	$\sigma^2+\frac{an\sum\sum(\beta\gamma)_{jk}^2}{(b-1)(c-1)}$	$F_0=\frac{MS_{BC}}{MS_E}$
ABC	SS_{ABC}	$(a$-1$)(b$-1$)(c$-1)	MS_{ABC}	$\sigma^2+\frac{n\sum\sum\sum(\tau\beta\gamma)_{ijk}^2}{(a-1)(b-1)(c-1)}$	$F_0=\frac{MS_{ABC}}{MS_E}$
Errore	SS_E	$abc(n$ - 1$)$	MS_E	σ^2	
Totale	SS_T	$abcn$ - 1			

Anche per i calcoli dell'analisi della varianza si ricorre solitamente ad un software statistico.

2.5 L'uso dei blocchi in un piano fattoriale

I piani fattoriali sono stati presentati ed analizzati nel contesto di un esperimento completamente casualizzato; talvolta però la casualizzazione completa di tutte le prove è poco praticabile, se non impossibile. Per esempio, la presenza di **fattori di disturbo** può richiedere che l'esperimento sia eseguito a **blocchi**. In generale, si definisce come fattore di disturbo un fattore che quasi certamente produce sulla risposta un effetto, che non interessa però allo sperimentatore. La casualizzazione è la tecnica di pianificazione usata per proteggere le analisi statistiche dalla presenza di un fattore di disturbo "nascosto", né noto né controllato; in pratica lo sperimentatore non sa neppure dell'esistenza di quel fattore ed i relativi livelli possono variare, sempre a sua insaputa, anche nel corso dell'esperimento. In altri casi, il fattore di disturbo è noto ma non è controllabile. Se almeno si potesse osservare quale valore il fattore di disturbo assume durante ciascuna prova dell'esperimento, se ne potrebbe tenere

conto nell'analisi statistica usando la tecnica dell'**analisi della covarianza**. Quando l'origine della variabilità di disturbo è nota e controllabile, si può usare la tecnica di pianificazione chiamata tecnica dei **blocchi**, per eliminare sistematicamente l'effetto in questione dai confronti statistici fra trattamenti. In questo caso infatti l'errore sperimentale rifletterà sia gli errori casuali sia la variabilità determinata dal disturbo. La tecnica dei blocchi permette di rendere minimo l'errore sperimentale, cioè di rimuovere dall'errore sperimentale la variabilità aggiuntiva data dal fattore di disturbo.

Si consideri un esperimento fattoriale con due fattori *A* e *B* ed *n* replicazioni. Il modello statistico lineare che descrive questo piano è (2.1). Si supponga ora che per eseguire questo esperimento sia richiesta una particolare materia prima. Questa materia prima è disponibile in lotti di dimensioni insufficienti per consentire di eseguire con lo stesso lotto tutte le $a \cdot b \cdot n$ combinazioni dei trattamenti. Tuttavia, se un lotto contiene abbastanza materiale per $a \cdot b$ osservazioni, allora un piano alternativo consiste nell'eseguire ciascuna delle *n* replicazioni usando lotti separati di materia prima. Conseguentemente, i lotti di materia prima rappresentano una restrizione alla casualizzazione o un **blocco**, e una singola replicazione di un esperimento fattoriale completo è eseguita all'interno di ciascun blocco. Il modello degli effetti per questo nuovo piano è:

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \delta_k + \varepsilon_{ijk} \qquad \begin{cases} i = 1,2,...,a \\ j = 1,2,...,b \\ k = 1,2,...,n \end{cases} \tag{2.19}$$

dove δ_k è l'effetto del *k*-esimo blocco. Naturalmente all'interno di un blocco l'ordine in cui le combinazioni dei trattamenti sono eseguite è completamente casualizzato. Il modello (2.19) assume che l'interazione tra blocchi e trattamenti sia trascurabile; nel caso in cui queste interazioni esistano, è però impossibile separarle dall'errore. In realtà il termine d'errore in questo modello è costituito dalle interazioni $(\tau\delta)_{ik}$, $(\beta\delta)_{jk}$ e $(\tau\beta\delta)_{ijk}$. L'analisi della varianza è descritta nella Tabella 2.4. L'impianto è molto simile a quello di un piano fattoriale, con la somma dei quadrati dell'errore diminuita della somma dei quadrati dei blocchi. Nei calcoli si trova la somma dei quadrati dei blocchi come la somma dei quadrati tra gli *n* totali dei blocchi $y_{..k}$.

Tabella 2.4 *Analisi della varianza per un fattoriale a due fattori a blocchi completi casualizzati*

Origine della variabilità	**Somma dei quadrati**	**Gradi di libertà**	**Valore atteso dei quadrati medi**	**F_0**
Blocchi	$\frac{1}{ab}\sum_k y_{..k}^2 - \frac{y_{...}^2}{abn}$	$n-1$	$\sigma^2 + ab\sigma_\delta^2$	
A	$\frac{1}{bn}\sum_i y_{i..}^2 - \frac{y_{...}^2}{abn}$	$a-1$	$\sigma^2 + \frac{bn\sum \tau_i^2}{a-1}$	$\frac{MS_A}{MS_E}$
B	$\frac{1}{an}\sum_j y_{.j.}^2 - \frac{y_{...}^2}{abn}$	$b-1$	$\sigma^2 + \frac{an\sum \beta_j^2}{b-1}$	$\frac{MS_B}{MS_E}$
AB	$\frac{1}{n}\sum_i\sum_j y_{ij.}^2 - \frac{y_{...}^2}{abn} - SS_A - SS_B$	$(a-1)(b-1)$	$\sigma^2 + \frac{n\sum\sum (\tau\beta)_{ij}^2}{(a-1)(b-1)}$	$\frac{MS_{AB}}{MS_E}$
Errore	Sottrazione	$(ab-1)(n-1)$	σ^2	
Totale	$\sum_i\sum_j\sum_k y_{ijk}^2 - \frac{y_{...}^2}{abn}$	abn - 1		

Nel precedente esempio la casualizzazione è stata limitata all'interno di un lotto di materia prima. In pratica una varietà di fenomeni può introdurre vincoli nella casualizzazione, quali il tempo, il personale e così via. Per esempio se non si può eseguire l'intero esperimento fattoriale in un giorno, allora lo sperimentatore potrebbe eseguire una replicazione completa il primo giorno, una seconda replicazione il secondo, e così via; ogni giorno verrebbe così a formare un blocco.

Dalla Tabella 2.4 si può notare come non venga effettuato il confronto tra le medie dei blocchi; l'interesse potrebbe essere anche quello di confrontare tali medie, oltre a verificare l'eguaglianza tra le medie dei trattamenti, poiché, se queste medie non differissero molto, l'uso dei blocchi potrebbe non essere necessario in futuri esperimenti. Tuttavia si ricordi che la casualizzazione è stata applicata ai trattamenti solo *entro* i blocchi, cioè i blocchi ne rappresentano una restrizione. Il problema dell'effetto che può avere la casualizzazione ristretta sulla statistica $F_0 = MS_{Blocchi} / MS_E$ è stato trattato in vari modi. Box, Hunter e Hunter (1978) fanno rilevare che l'usuale test F dell'analisi della varianza può essere giustificato sulla base anche della sola casualizzazione, senza l'uso diretto dell'assunzione di normalità. Essi inoltre osservano che il test per confrontare le medie dei blocchi non può richiamarsi a tale giustificazione a causa della restrizione della casualizzazione; ma se gli errori sono normalmente e indipendentemente distribuiti con media 0 e varianza σ^2, la statistica

$F_0 = MS_{Blocchi} / MS_E$ può essere usata per confrontare le medie dei blocchi. Anderson e McLean (1974) sostengono che la restrizione della casualizzazione impedisce che questa statistica possa essere un valido test per confrontare le medie dei blocchi e che questo rapporto F in realtà è un test per l'eguaglianza delle medie dei blocchi più la restrizione di casualizzazione.
In pratica, poiché l'assunzione di normalità è spesso discutibile, vedere $F_0 = MS_{Blocchi} / MS_E$ come un test F esatto dell'eguaglianza delle medie dei blocchi in generale non è buona pratica. Per questa ragione questo test F viene escluso dalla tabella di analisi della varianza. Tuttavia come procedura approssimata per controllare l'effetto della variabile di blocco, l'esame del rapporto di $MS_{Blocchi}$ con MS_E è certamente ragionevole. Se questo rapporto è grande, ciò implica che il fattore di blocco ha un effetto notevole e che la riduzione della variabilità ottenuta usando i blocchi è stata probabilmente efficace nel migliorare la precisione dei confronti delle medie dei trattamenti.

2.6 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 $x_1,...,x_p$:

$$y = f(x_1, x_2, ..., x_p)$$

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 **X** la matrice dei dati con n righe (le osservazioni) e p colonne (le variabili), **y** il vettore delle n risposte sperimentali, **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:

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

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

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

dove f_i è una funzione lineare di X_i. In generale, dalla matrice originale dei dati **X** si ottiene la matrice del modello $\mathbf{X_M}$. Questa matrice può contenere, oltre alle colonne della matrice **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 ($\mathbf{x_1 \cdot x_2}$). 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.

2.6.1 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:

$$\mathbf{y = X \cdot \beta + e} , \tag{2.22}$$

dove β è il vettore dei coefficienti veri da stimare, **X** è la matrice del modello ed **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:

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

La soluzione consiste nel determinare il vettore dei coefficienti **b**. I seguenti passaggi algebrici portano alla soluzione cercata:

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

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

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

L'analisi dimensionale è la seguente:

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

con $p' = p + 1$.

La matrice $\mathbf{X}^{\mathbf{T}} \cdot \mathbf{X}$ che compare nell'espressione matriciale di **b** introdotta precedntemente 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 varianza-covarianza** (o *matrice di dispersione*). La matrice varianza-covarianza 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 varianza-covarianza 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 varianza-covarianza dipende soltanto 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 varianza-covarianza 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 varianza-covarianza. 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 $\mathbf{(X^T \cdot X)^{-1}}$ 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:

$$\rho_{ij} = \frac{c_{ij}}{\sqrt{c_{ii} \cdot c_{jj}}}, \tag{2.25}$$

dove c_{ii} sono gli elementi della diagonale di $\mathbf{(X^T \cdot X)^{-1}}$ 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 varianza-covarianza è possibile ricavare la **matrice di influenza** (o *matrice dei leverage* o **hat matrix**). Il vettore delle risposte calcolate $\hat{y}$ si determina come:

$$\hat{y} = \mathbf{X \cdot b}. \tag{2.26}$$

Sostituendo al vettore dei coefficienti **b** l'espressione ricavata in precedenza, si ottiene un'importante relazione:

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

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

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

Poichè dipende solo da **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:

$$\frac{1}{n} \le h_{ii} \le \frac{1}{r_i} \ , \tag{2.29}$$

  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 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 σ^2:

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

  Un'importante estensione si questa equazione applicabile ai punti $\mathbf{x_0}$ che non sono necessariamente nel disegno sperimentale è la seguente:

$$\mathrm{var}(\hat{Y}_i | x_0) = \sigma^2 \cdot \ \mathbf{x_0}^{\mathbf{T}} \cdot (\mathbf{X}^{\mathbf{T}} \cdot \mathbf{X})^{-1} \cdot \mathbf{x_0} = \sigma^2 \cdot h_{00} \, . \tag{2.31}$$

  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:

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

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. É 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 *influente* 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 tr(**H**) = p. Ciò significa che tralasciando σ^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.

2.6.2 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):

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

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

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

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):

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

Vale pertanto la seguente relazione:

$$SS_T = SS_R + SS_E \ . \tag{2.36}$$

É 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:

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

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:

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

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:

$$R^2_{adj} = 1 - \frac{(\sum_{i=1}^{n} (\hat{y}_i - y_i)^2)/(n-p)}{(\sum_{i=1}^{n} (y_i - \bar{y})^2)/(n-1)}, \tag{2.39}$$

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^2_{adj} 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^2_{adj} è massimo coinciderà anche con il punto in cui la differenza tra R^2_{adj} 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:

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

dove *PRESS* sta per *PRedictive Error Sum of Squares* 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:

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

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^2_{pred} 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:

$$e_{i,-i} = \frac{e_i}{1 - h_{ii}} , \tag{2.42}$$

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 leverage, gli effetti sui residui possono essere abbastanza impressionanti.

2.6.3 Diagnostici del modello di regressione

2.6.3.1 Residui ridotti

Questo paragrafo riassume alcune procedure diagnostiche (Montgomery e Peck, 1992). É 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 (ε_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 (1.11). 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:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}. \tag{2.43}$$

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

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

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:

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

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*):

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

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 σ^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 σ^2 basata su un insieme di dati rimuovendo l'i-esima osservazione. Indicando la stima di σ^2 così ottenuta con $S_{(i)}^2$ si può mostrare che:

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

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

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

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

2.6.3.2 Diagnostici di influenza

Gli elementi diagonali della matrice **H**, cioè i leverage, 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:

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T \cdot \mathrm{X}^{\mathrm{T}} \cdot \mathrm{X} \cdot (\hat{\beta}_{(i)} - \hat{\beta})}{p \cdot MS_E} \qquad i = 1, 2, \ldots, n \,. \tag{2.49}$$

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:

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

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:

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

mentre la corrispondente misura normalizzata è data dall'espressione:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{ii}}{s_{(i)} \cdot \sqrt{h_{ii}}} = r'_{i/i} \cdot \left(\frac{h_{ii}}{1-h_{ii}} \right)^{1/2} . \tag{2.52}$$

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 $3 \cdot \sqrt{p'/n}$, con $p' = p + 1$.

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:

$$DFBETA_i = \hat{\mathrm{b}} - \hat{\mathrm{b}}_{i/i} = \frac{(\mathrm{X}^{\mathrm{T}} \cdot \mathrm{X})^{-1} \cdot x_i \cdot r_i}{1-h_{ii}} . \tag{2.53}$$

Il corrispondente parametro scalato, relativo a ciascuna variabile, è

$$DFBETAS_{ij} = \frac{\hat{b}_j - \hat{b}_{j(i)}}{s_{(i)} \cdot \sqrt{d^{jj}}} = \frac{\left\{ (\mathrm{X}^{\mathrm{T}} \cdot \mathrm{X})^{-1} \cdot \mathrm{x}_i \right\}_j \cdot r_i}{s_{(i)} \cdot \sqrt{d^{jj}}} = \frac{\left\{ (\mathrm{X}^{\mathrm{T}} \cdot \mathrm{X})^{-1} \cdot \mathrm{x}_i \right\}_j \cdot r'_{i/i}}{\sqrt{d^{jj}} \cdot \sqrt{1-h_{ii}}} , \tag{2.54}$$

dove il termine $\left\{\left(\mathrm{X}^{\mathrm{T}}\cdot\mathrm{X}\right)^{-1}\cdot\mathrm{x}_i\right\}_j$ rappresenta il j-esimo elemento del vettore $\left(\mathrm{X}^{\mathrm{T}}\cdot\mathrm{X}\right)^{-1}\cdot\mathrm{x}_i$.
Valori di controllo di questo parametro sono $2/\sqrt{n}$ e $3/\sqrt{n}$.

2.6.4 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 λ è il parametro di trasformazione che deve essere determinato. Box e Cox (1964) hanno dimostrato come il parametro di trasformazione λ possa essere stimato simultaneamente con altri parametri del modello (media generale ed effetti dei trattamenti). La procedura di calcolo consiste nell'eseguire, per differenti valori di λ, un'analisi della varianza standard su:

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

dove $\tilde{y} = \ln^{-1}\left[(1/n)\sum \log y\right]$ è la media geometrica delle osservazioni. La stima di massima verosimiglianza di λ 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 λ, 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 λ.
Tale intervallo è determinato calcolando:

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

(dove $t_{\alpha/2,\nu}$ è il percentile che stacca alla sua destra un'area pari ad α/2 sulla distribuzione *t* di student, con *ν* numero di gradi di libertà) e tracciando una linea parallela all'asse λ 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 λ, 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 λ, si possono analizzare i dati usando $\hat{y}^{(\lambda)}$ come risposta, a meno che λ non sia zero, nel qual caso si può usare $\ln y$.

2.7 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 $0 \le d_i \le 1$, 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:

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

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:

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

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:

$$y < T$$

$$d = \begin{cases} 1 & \\ \left(\dfrac{U-y}{U-T}\right)^{r} & T \le y \le U \\ 0, & y > U \end{cases}. \tag{2.59}$$

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:

$$d = \begin{cases} 0 & y < L \\ \left(\dfrac{y-L}{T-L}\right)^{r_1} & L \le y \le T \\ \left(\dfrac{U-y}{U-T}\right)^{r_2} & T \le y \le U \\ 0, & y > U \end{cases}. \tag{2.60}$$

L'ulteriore possibilità offerta da Design Expert è data dal porre come obiettivo dell'ottimizzazione il fatto che il valore della risposta cada all'interno di un intervallo stabilito. 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:

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

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 si è svolta attraverso la progettazione di un piano sperimentale da effettuarsi in produzione, l'esecuzione degli esperimenti e la raccolta dati, l'analisi delle risposte e l'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

L'impianto pilota (Figura 3.1) su cui e stata condotta la parte sperimentale, è costituito da un sistema di macinazione (mulino a microsfere), 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 3.1 *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).
Il funzionamento si basa sul 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. La foto dell'apparecchiatura e i relativi dati tecnici sono riportati in Figura 3.1 e Tabella 3.1. In Figura 3.2 si riporta il particolare del sinottico.

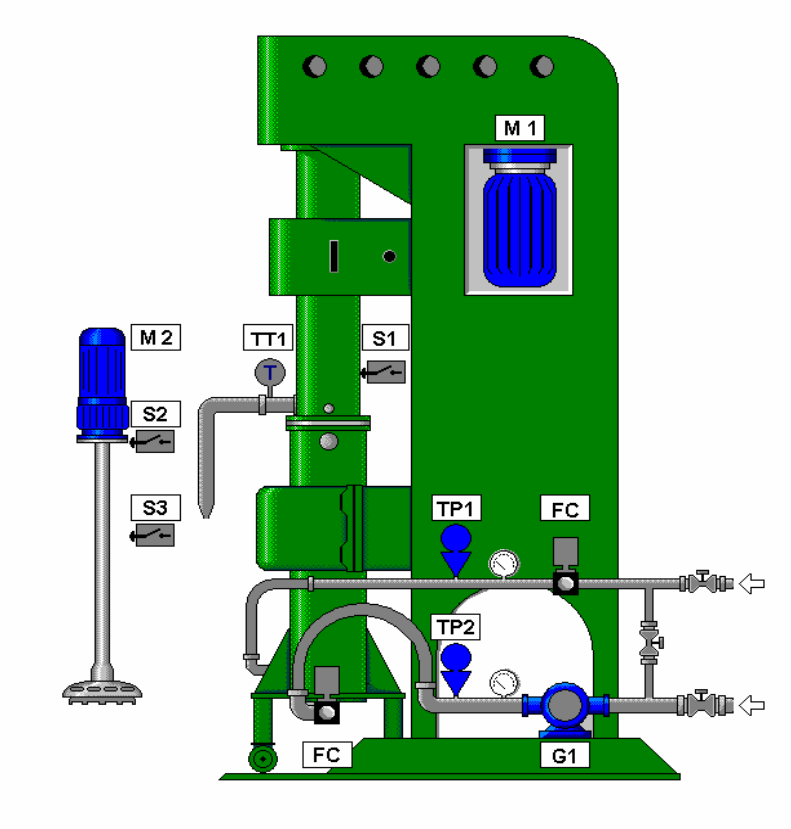

Figura 3.2 *Impianto pilota: particolare del sinottico*

Tabella 3.1 *Dati tecnici dell'impianto pilota*

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

3.2 Progettazione del piano sperimentale

Il lavoro è consistito nella progettazione di un piano sperimentale da effettuarsi in produzione. Si tratta di un piano fattoriale generale, in cui sono state considerate tre variabili:

- Portata della pompa, fattore quantitativo su tre livelli (350, 400 e 450 kg/h), impostabile attraverso il sistema di supervisione e controllo.
- Velocità della girante, fattore quantitativo su due livelli (830 e 1130 giri/minuto), impostabile attraverso il sistema di supervisione e controllo.
- Numero di ricircoli, fattore qualitativo su due livelli (3 e 4)

I livelli delle variabili sono stati definiti in modo da descrivere un'ampia zona sperimentale e le normali condizioni operative dell'apparecchiatura. Si è inoltre tenuto conto nella progettazione del piano della presenza di due differenti lotti di materia prima (pigmento) per l'esecuzione delle prove. Tali fattori sono stati considerati quindi come blocchi, in modo tale da valutarne l'influenza sulle singole risposte. A questo proposito, sulla base della quantità disponibile di materia prima, è stato possibile decidere di effettuare una replicazione del piano sperimentale, ognuna delle quali è avvenuta utilizzando lo stesso lotto di materia, cioè all'interno di ogni singolo blocco.

Sono state analizzate pertanto tutte le possibili combinazioni dei livelli dei tre fattori, ciascuna delle quali ripetuta due volte, per un totale di ventiquattro esperimenti, svolti in una sequenza

completamente casualizzata generata dal software. I punti sperimentali, in ordine di esecuzione randomizzato, sono riportati in Tabella 3.2.

Tabella 3.2 *Riassunto del piano sperimentale*

			Fattore 1	Fattore 2	Fattore 3
Ordine standard	Ordine di esecuzione	Blocco	A:portata pompa	B:passaggi	C:velocità girante
			kg/h	n	giri/min
1	**1**	lotto A	350	3	830
19	**2**	lotto A	350	4	1130
11	**3**	lotto A	450	4	830
7	**4**	lotto A	350	4	830
21	**5**	lotto A	400	4	1130
15	**6**	lotto A	400	3	1130
13	**7**	lotto A	350	3	1130
17	**8**	lotto A	450	3	1130
5	**9**	lotto A	450	3	830
9	**10**	lotto A	400	4	830
3	**11**	lotto A	400	3	830
23	**12**	lotto A	450	4	1130
12	**13**	lotto B	450	4	830
2	**14**	lotto B	350	3	830
24	**15**	lotto B	450	4	1130
20	**16**	lotto B	350	4	1130
6	**17**	lotto B	450	3	830
10	**18**	lotto B	400	4	830
14	**19**	lotto B	350	3	1130
16	**20**	lotto B	400	3	1130
22	**21**	lotto B	400	4	1130
4	**22**	lotto B	400	3	830
8	**23**	lotto B	350	4	830
18	**24**	lotto B	450	3	1130

La colonna relativa all'ordine standard è indicativa della sequenza randomizzata con la quale sono state eseguite le prove.

3.3 Esecuzione delle prove: raccolta dati e risposte

Ciascuna prova è consistita nel far passare attraverso il mulino a microsfere 300 kg di prodotto da raffinare, seguendo quanto riportato in Tabella 3.2. Il prodotto, pompato da una bacinella tramite la pompa volumetrica (G1) all'interno del cilindro contenente le microsfere, è raccolto in una seconda bacinella all'uscita. Le variabili di risposta analizzate sono state la resa coloristica, la viscosità misurata in cps, il consumo di energia elettrica del motore (definito nel prosieguo con il termine lavoro) misurato in KWh e il tempo di lavorazione misurato in minuti (entrambe lette a quadro). La resa coloristica (nel prosieguo della trattazione denominata semplicemente resa) è stata determinata

applicando il metodo interno SAMIA s.a.s. IL02 (CV% = 1,1), 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 s.a.s. 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°C, la velocità del sensore è stata impostata a 45.3 rpm.
Per il tempo di lavorazione, sono stati annotati i diversi orari di inizio e fine per ciascun ricircolo, assumendo come tempo di lavorazione totale la somma dei tempi parziali necessari per ogni passaggio. Il riassunto delle prove con le rispettive risposte è riportato in Tabella 3.3:

Tabella 3.3 *Riassunto delle prove sperimentali e delle risposte ottenute.*

	Std	Run	Block	Factor 1 A:portata pompa kg/h	Factor 2 B:passaggi n	Factor 3 C:velocità girante giri/min	Response 1 RESA	Response 2 viscosità (a resa) cPs	Response 3 POTENZA ass KW	Response 4 tempo min
	1	1	Block 1	350.00	3	830.00	110.8	560	26.6	238
	19	2	Block 1	350.00	4	1130.00	115	665	35.4	260
	11	3	Block 1	450.00	4	830.00	108	510	16.7	225
	7	4	Block 1	350.00	4	830.00	111.9	640	21.8	247
	21	5	Block 1	400.00	4	1130.00	115	720	30.4	260
	15	6	Block 1	400.00	3	1130.00	111.8	760	22.9	182
	13	7	Block 1	350.00	3	1130.00	113.2	670	27.2	261
	17	8	Block 1	450.00	3	1130.00	110.5	610	21.9	155
	5	9	Block 1	450.00	3	830.00	107.5	540	13.4	166
	9	10	Block 1	400.00	4	830.00	110	600	20.6	243
	3	11	Block 1	400.00	3	830.00	108.3	640	16.3	198
	23	12	Block 1	450.00	4	1130.00	113.8	670	30.5	231
	12	13	Block 2	450.00	4	830.00	112.2	600	20.2	207
	2	14	Block 2	350.00	3	830.00	115.3	715	21.6	245
	24	15	Block 2	450.00	4	1130.00	121	730	34.4	217
	20	16	Block 2	350.00	4	1130.00	120	1000	44.3	255
	6	17	Block 2	450.00	3	830.00	112	690	18.2	188
	10	18	Block 2	400.00	4	830.00	114	985	34.1	277
	14	19	Block 2	350.00	3	1130.00	122.2	2100	39.5	230
	16	20	Block 2	400.00	3	1130.00	119	1400	35.2	209
	22	21	Block 2	400.00	4	1130.00	119	5000	45.2	262
	4	22	Block 2	400.00	3	830.00	115.1	3800	27.7	204
	8	23	Block 2	350.00	4	830.00	115.6	5200	42.7	251
	18	24	Block 2	450.00	3	1130.00	115.3	3650	31.3	181

Si è successivamente passati all'analisi delle risposte, effettuata con il supporto del software. La procedura, dopo una prima valutazione del piano sperimentale, si articola attraverso la scelta del

modello da accostare ai dati, la successiva analisi della varianza, il controllo dell'adeguatezza del modello e il calcolo della superficie di risposta. Dopo aver ripetuto il procedimento per le quattro risposte considerate, si è passati alla fase di ottimizzazione delle condizioni operative.

3.4 Valutazione del piano sperimentale

Prima di procedere con l'analisi delle risposte, il software permette di valutare il disegno sperimentale. Ciò viene fatto generando una risposta casuale normalmente distribuita con varianza unitaria e accostando ad essa un modello standard selezionato a priori. In Tabella 3.4 è mostrata la ripartizione dei gradi di libertà.

Tabella 3.4 *Gradi di libertà associati al piano sperimentale*

3 Factors: A, B, C

Design Matrix Evaluation for Response Surface 2FI Model

No aliases found for 2FI Model

Aliases are calculated based on your response selection,
taking into account missing datapoints, if necessary.
Watch for aliases among terms you need to estimate.

Degrees of Freedom for Evaluation

Blocks	1
Model	6
Residuals	16
Lack Of Fit	*16*
Pure Error	*0*
Corr Total	23

A recommendation is a minimum of 3 lack of fit df and 4 df for pure error.
This ensures a valid lack of fit test.
Fewer df will lead to a test that may not detect lack of fit.

"No aliases found" si riferisce al fatto che per il modello selezionato (di secondo ordine), non sono presenti effetti sovrapposti che non sono stimabili. La Tabella 3.5 riporta i risultati della valutazione del modello:

Tabella 3.5 *Valutazione del disegno sperimentale*

Term	StdErr**	VIF	Ri-Squared	Power at 5 % alpha level for effect of 0.5 Std. Dev.	1 Std. Dev.	2 Std. Dev.
Block 1	0,20					
Block 2						
A	0,25	1	0	15.6 %	46.8 %	96.3 %
B	0,20	1	0	21.1 %	63.3 %	99.6 %
C	0,20	1	0	21.1 %	63.3 %	99.6 %
AB	0,25	1	0	15.6 %	46.8 %	96.3 %
AC	0,25	1	0	15.6 %	46.8 %	96.3 %
BC	0,20	1	0	21.1 %	63.3 %	99.6 %

****Basis Std. Dev. = 1.0**

For Categorical Terms, The minimum Power for each group of terms is reported.

Standard errors should be similar within type of coefficient. Smaller is better.

Ideal VIF is 1.0. VIF's above 10 are cause for alarm,
indicating coefficients are poorly estimated due to multicollinearity.

Ideal Ri-squared is 0.0. High Ri-squared means terms are correlated with each other,
possibly leading to poor models.

Power should be approximately 80% for the effect you want to detect.
Be sure to set the Model (on previous screen) to be an estimate of the terms you expect to be significant.

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 nel disegno sperimentale. 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 1, cosa che si osserva nel disegno considerato. VIF pari a uno o superiore indica che è presente una certa correlazione tra i coefficienti. VIF è legato a *R-squared* dalla formula:

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

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 3.5 si intende la capacità del disegno sperimentale di individuare che i termini specifici siano statisticamente significativi. L'output afferma che questo disegno ha il 96,3% 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 scoprire la significatività si riduce al 63,3 o al 46,8%. 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.

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.

In Tabella 3.6 sono riportati i leverage dei diversi punti sperimentali. Si assume come valore di controllo per i leverage l'unità. É possibile osservare quindi che nessun punto ha un'influenza elevata relativamente agli altri sul disegno. I punti che presentano più alto leverage sono quelli ai confini del dominio sperimentale, dove il "rumore di fondo" (Figura 3.3) è maggiore (si ricordi che leverage e rumore di fondo sono legati dalla formula (2.44)):

Tabella 3.6 *Valori di leverage per i punti sperimentali*

Measures Derived From the $(X'X)^{-1}$ Matrix

Std	Leverage	Point Type
1	0.3958	Fact
2	0.3958	Fact
3	0.2083	??
4	0.2083	??
5	0.3958	Fact
6	0.3958	Fact
7	0.3958	Fact
8	0.3958	Fact
9	0.2083	??
10	0.2083	??
11	0.3958	Fact
12	0.3958	Fact
13	0.3958	Fact
14	0.3958	Fact
15	0.2083	??
16	0.2083	??
17	0.3958	Fact
18	0.3958	Fact
19	0.3958	Fact
20	0.3958	Fact
21	0.2083	??
22	0.2083	??
23	0.3958	Fact
24	0.3958	Fact
Average =	**0.3333**	

Watch for leverages close to 1.0. Consider replicating these points or make sure they are run very carefully.

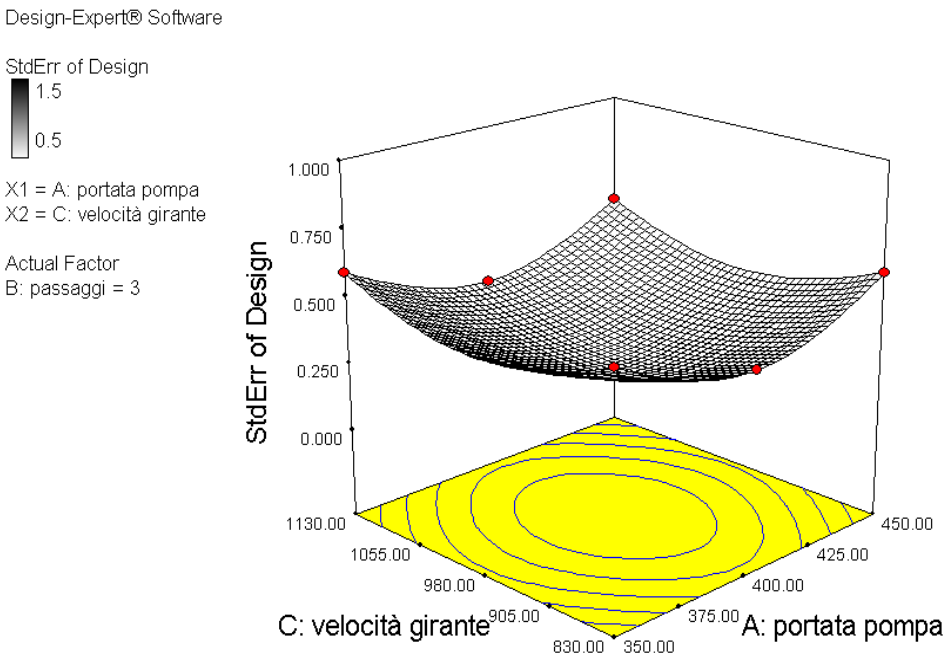

Figura 3.3 *Errore standard nel dominio sperimentale*

La forma della superficie in Figura 3.3 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 3.3 è stata considerata una deviazione standard pari a uno. La forma ideale è una superficie simmetrica che presenta contorni circolari sul piano delle variabili, come si osserva in figura. Un'altra caratteristica desiderabile è il relativamente basso errore negli intorni del centro.

3.5 Analisi delle risposte

3.5.1 Resa

La prima risposta presa in considerazione è la resa coloristica. Sulla base dei dati raccolti il software suggerisce di utilizzare un modello lineare, come indicato in Tabella 3.7

Tabella 3.7 *Suggerimento del modello per la risposta "resa"*

Response 1 **RESA** **Transform:** **None**

****** WARNING: The Quadratic Model is Aliased! ******

****** 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	312018,01	1	312018,0104			
Block vs Mean	175,50	1	175,5004167			
Linear vs Block	170,3664583	3	56,78881944	33,81059226	< 0.0001	Suggested
2FI vs Linear	9,231666667	3	3,077222222	2,170780173	0.1313	
Quadratic vs 2FI	0,000208333	1	0,000208333	0,000137782	0.9908	Aliased
Cubic vs Quadratic	6,751041667	3	2,250347222	1,695198985	0.2209	Aliased
Residual	15,92979167	12	1,327482639			
Total	312395,79	24	13016,49125			

"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	
Linear	1,296000088	0,84223433	0,817323961	0,741273009	52,3350802	Suggested
2FI	1,190615431	0,887872577	0,845824793	0,720137093	56,61043567	
Quadratic	1,229656682	0,887873607	0,835547957	0,693069524	62,08564084	Aliased
Cubic	1,152164328	0,921248481	0,855622215	0,678452211	65,04241876	Aliased

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

Da notare che sia il modello quadratico che quello cubico sono "aliased", cioè non è possibile stimare tutti gli effetti perché sovrapposti. Per farlo, sarebbe necessario aumentare il numero di prove sperimentali. Il suggerimento del miglior modello viene fatto sulla base del test *F*, confrontando tra loro modelli a cui in successione vengono aggiunti termini, come indicato in §1.6.2. L'attenzione è in seguito focalizzata sul modello che massimizza gli indici statistici di fitting e di predizione (§1.6.2). In particolare sono considerati migliori i modelli che danno il miglior risultato per quanto riguarda R^2_{adj} e R^2_{pred}. In questo caso viene consigliato un modello lineare;

tuttavia si è deciso di scegliere per la successiva analisi un modello di secondo ordine, in modo tale da indagare le interazioni tra fattori.

Si è passati pertanto all'analisi della varianza (riportata in Tabella 3.8), per verificare la significatività dei fattori. Si ottiene che tutti i fattori considerati sono significativi, in particolar modo la portata della pompa e la velocità della girante. Come era prevedibile nessuna interazione risulta essere significativa. Nella parte inferiore dell'output sono riportate le prestazioni del modello fittato, introdotte in precedenza. La grandezza C.V.% è il coefficiente di variazione definito come:

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

Esso esprime la variabilità non spiegata o residua dei dati come percentuale della risposta media. Si osserva 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 sono da preferire 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 ottimale.

Tabella 3.8 *ANOVA per la risposta "resa"*

Response 1 RESA

ANOVA for Response Surface 2FI Model

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Block	175,50	1	175,5004167			
Model	179,60	6	29,93302083	21,11579972	< 0.0001	significant
A-portata pompa	35,11	1	35,105625	24,7647356	0.0001	
B-passaggi	8,76	1	8,760416667	6,179904289	0.0244	
C-velocità girante	126,50	1	126,5004167	89,23781793	< 0.0001	
AB	4,73	1	4,730625	3,337148316	0.0864	
AC	1,05	1	1,050625	0,741147618	0.4020	
BC	3,450416667	1	3,450416667	2,434044586	0.1383	
Residual	22,68104167	16	1,417565104			
Cor Total	377,7795833	23				

The Model F-value of 21.12 implies the model is significant. There is only a 0.01% chance that a "Model F-Value" this large could occur due to noise.

Values of "Prob > F" less than 0.0500 indicate model terms are significant.
In this case A, B, C are significant model terms.
Values greater than 0.1000 indicate the model terms are not significant.
If there are many insignificant model terms (not counting those required to support hierarchy), model reduction may improve your model.

Std. Dev.	1,190615431	R-Squared	0,887872577
Mean	114,0208333	Adj R-Squared	0,845824793
C.V. %	1,044208673	Pred R-Squared	0,720137093
PRESS	56,61043567	Adeq Precision	20,61505561

The "Pred R-Squared" of 0.7201 is in reasonable agreement with the "Adj R-Squared" of 0.8458.

"Adeq Precision" measures the signal to noise ratio. A ratio greater than 4 is desirable. Your ratio of 20.615 indicates an adequate signal. This model can be used to navigate the design space.

Vengono quindi calcolati i coefficienti di regressione per il modello (Tabella 3.9):

Tabella 3.9 *Calcolo dei coefficienti di regressione*

Coefficient Factor	Estimate	Standard df	95% CI Error	95% CI Low	High	VIF
Intercept	114.02	1	0.24	113.51	114.54	
Block 1	-2.70	1				
Block 2	2.70					
A-portata pompa	-1.48	1	0.30	-2.11	-0.85	1.00
B-passaggi	0.60	1	0.24	0.089	1.12	1.00
C-velocità girante	2.30	1	0.24	1.78	2.81	1.00
AB0.54	1	0.30	-0.087	1.17	1.00	
AC0.26	1	0.30	-0.37	0.89	1.00	
BC0.38	1	0.24	-0.14	0.89	1.00	

Final Equation in Terms of Coded Factors:

RESA	=
+114.02	
-1.48	* A
+0.60	* B
+2.30	* C
+0.54	* A * B
+0.26	* A * C
+0.38	* B * C

Per ogni coefficiente è 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 $\sqrt{MS_E / n}$, con *n* numero di prove. Da notare che gli intervalli di confidenza per i coefficienti delle interazioni presentano 0 al loro interno, segno della loro bassa influenza sulla risposta. É inoltre riportata l'equazione finale in termini di fattori adimensionali. In Tabella 3.10 sono riportate le equazioni ottenute per il modello:

Tabella 3.10 *Equazioni del modello di regressione per la risposta" resa"*

Final Equation in Terms of Actual Factors:

passaggi	3
RESA	=
+130.48778	
-0.073983	* portata pompa
-8.88889E-004	* velocità girante
+3.41667E-005	* portata pompa * velocità girante
passaggi	4
RESA	=
+118.04167	
-0.052233	* portata pompa
+4.16667E-003	* velocità girante
+3.41667E-005	* portata pompa * velocità girante

The Diagnostics Case Statistics Report has been moved to the Diagnostics Node.
In the Diagnostics Node, Select Case Statistics from the View Menu.

Proceed to Diagnostic Plots (the next icon in progression). Be sure to look at the:
1) Normal probability plot of the studentized residuals to check for normality of residuals.
2) Studentized residuals versus predicted values to check for constant error.
3) Externally Studentized Residuals to look for outliers, i.e., influential values.
4) Box-Cox plot for power transformations.

If all the model statistics and diagnostic plots are OK, finish up with the Model Graphs icon.

La fase successiva prevede il controllo diagnostico sul modello: vengono verificate le assunzioni fatte sul modello: l'assunzione di normalità e indipendenza degli errori e la costanza della varianza. La diagnostica prevede l'esame del comportamento dei residui, osservabile dai rispettivi grafici. Per quanto riguarda l'assunzione di normalità viene riportato il grafico di probabilità normale in Figura 3.4 (§1.3.2.1). L'idealità è data dalla retta segnata in rosso, indicativa della distribuzione normale. Dal grafico si nota che non ci sono motivi per dubitare dell'assenza di normalità. Inoltre non si nota la presenza di *outlier*, elementi che si discosterebbero dall'andamento rettilineo e riscontrabili agli estremi della distribuzione. La ragione per cui in ascissa sono riportati i residui internamente studentizzati sta nel fatto che in tal maniera si tiene conto anche della posizione del punto nel dominio sperimentale, cosa che non avviene per i residui ordinari (si veda §1.6.3.1).

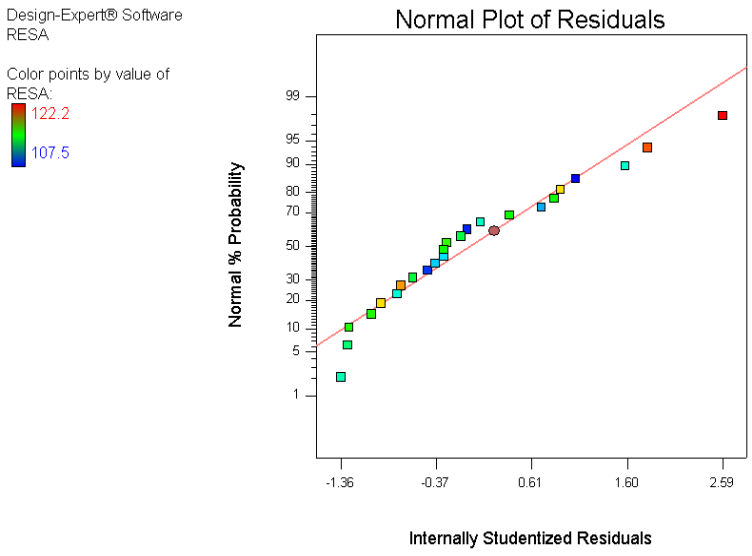

Figura 3.4 *Grafico di probabilità normale dei residui per la risposta " resa"*

Viene in seguito riportato il grafico dei residui rispetto ai valori predetti. L'importanza di questo grafico risiede nel fatto che tramite esso è possibile verificare l'eventuale disomogeneità della varianza. Qualsiasi andamento regolare nella distribuzione dei residui è indice di una varianza non costante. In particolare si cercano di evitare situazioni in cui il grafico assume una forma ad "imbuto", cosa che non si verifica nel caso in esame (Figura 3.5).

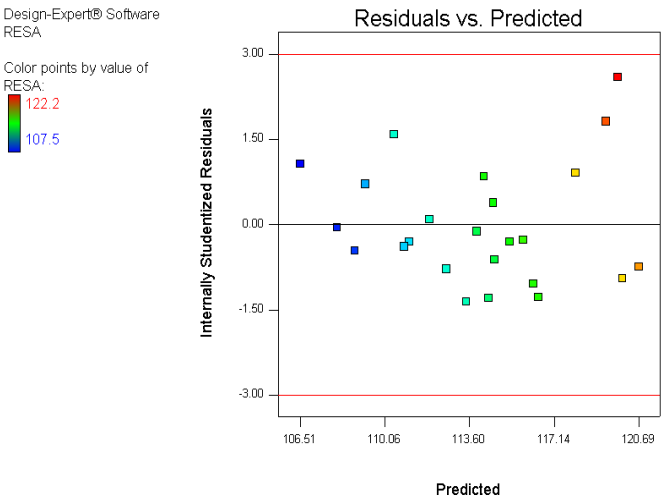

Figura 3.5 *Grafico dei residui rispetto ai valori previsti per la risposta "resa"*

Altro grafico importante è quello dei residui in funzione dalle sequenza sperimentale (Figura 3.6). Si nota che non sono presenti andamenti significativi o sistematici, in particolare non si osservano sequenze di residui positivi alternate a sequenze di residui negativi, ma si nota piuttosto un andamento altalenante, segno dell'indipendenza e della mancanza di correlazione tra i residui.

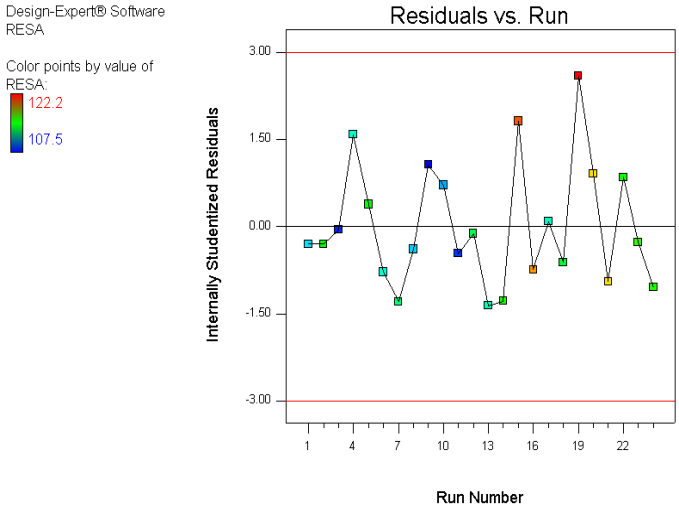

Figura 3.6 *Grafico dei residui rispetto la sequenza sperimentale*

In seguito è riportato il grafico dei residui all'interno di ogni singolo blocco. Dall'analisi appare evidente che nel secondo blocco è osservabile una maggiore variabilità nella risposta. Effettivamente se si va a considerare il rapporto tra i quadrati medi relativi ai blocchi (175,50 da

Tabella 3.8) e quelli relativi all'errore (1,42), si ottiene un valore relativamente elevato, segno che i blocchi hanno una certa influenza sulla risposta e che la riduzione della variabilità ottenuta con il loro utilizzo in questo caso è stata efficace. Probabilmente questa differenza è anche dovuta a possibili differenze nelle fasi di lavorazione precedenti alla raffinazione, che hanno determinato una componente di variabilità aggiuntiva a quella data dall'utilizzo di due differenti lotti di materia prima.

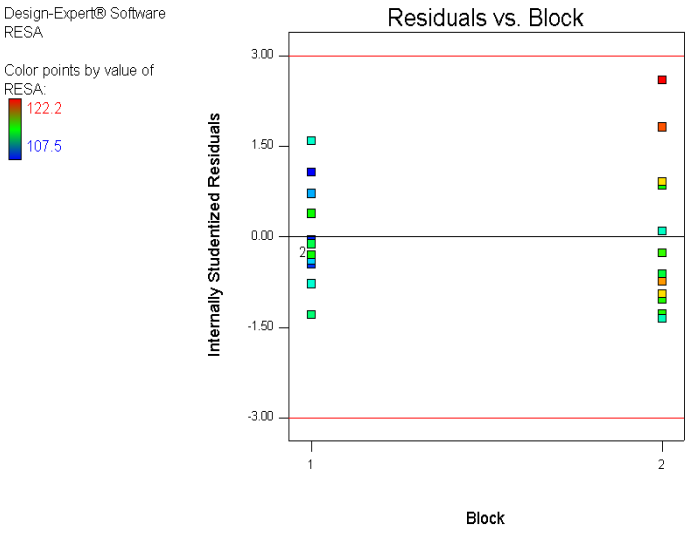

Figura 3.7 *Grafico dei residui rispetto ai blocchi per la risposta "resa"*

Si riporta quindi il grafico dei valori predetti contro i valori osservati: esso mostra la presenza o meno di valori che non sono facilmente predetti dal modello. L'andamento rettilineo costituisce l'idealità (Figura 3.8).

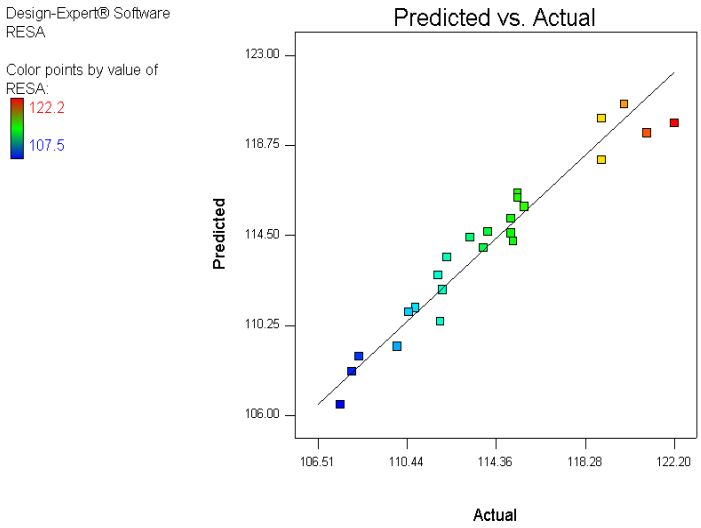

Figura 3.8 *Grafico dei valori predetti rispetto ai valori osservati*

Il grafico di Box-Cox fornisce informazioni nel caso sia necessario effettuare una trasformazione della variabile. In questo caso non viene raccomandata alcuna trasformazione. Il grafico è riportato in Figura 3.9.

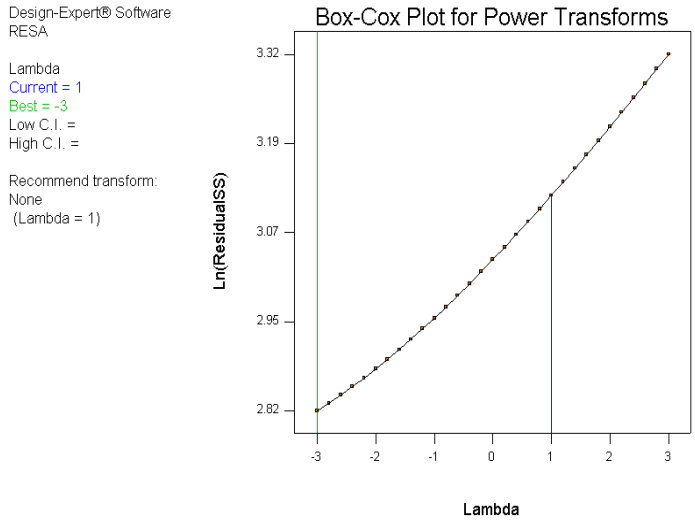

Figura 3.9 *Grafico di Box-Cox per la trasformazione della variabile di risposta*

Il software fornisce a questo punto i grafici relativi alle misure di influenza (residui esternamente studentizzati, leverage, *DFFITS* e distanza di Cook), riportate rispetto alla sequenza sperimentale.
I residui 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 riportati anche i limiti oltre i quali il valore è da considerare anomalo. Non si osservano anomalie sebbene il residuo relativo alla prova 19 si scosti leggermente rispetto agli altri, restando tuttavia all'interno dei limiti.
Di seguito è riportato il grafico dei leverage (Figura 3.11): come precedentemente osservato non ci sono punti che presentano leverage elevato; i punti a leverage inferiori sono quelli centrali sui confini del dominio sperimentale (per i quali la portata della pompa è pari a 400 kg/h). Non risultano dunque punti che presentano grande influenza sulla scelta del modello.

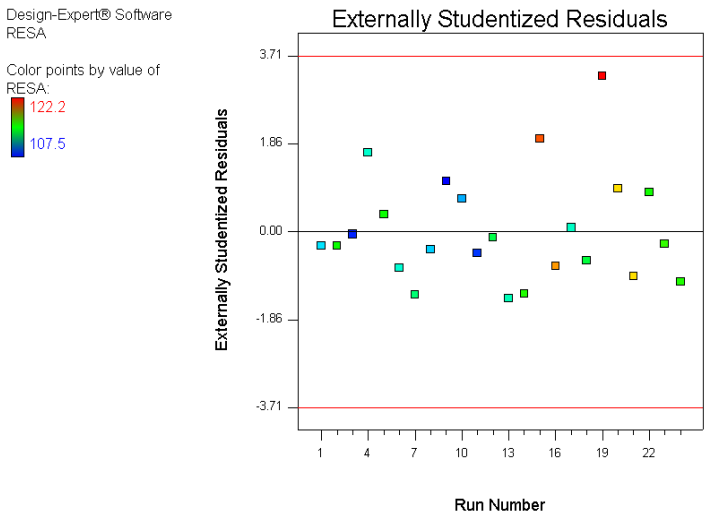

Figura 3.10 *Grafico dei residui esternamente studentizzati rispetto la sequenza temporale*

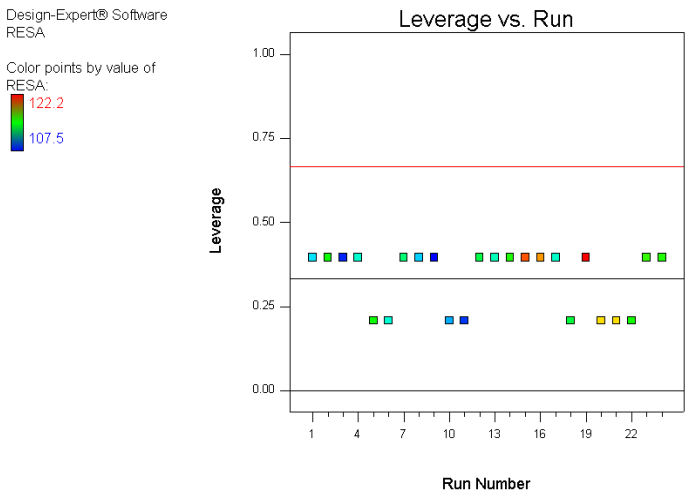

Figura 3.11 *Grafico dei leverage rispetto la sequenza sperimentale.*

Dall'esame del grafico di *DFFITS* rispetto la sequenza sperimentale, si osserva che è presente un valore che oltrepassa i limiti segnati (Figura 3.12). Tale valore corrisponde alla prova 19, caratterizzata dai seguenti livelli dei fattori:

- Portata = 350 kg/h
- Velocità della girante = 1130 rpm
- Numero passaggi = 3

e da un valore di resa pari a 122,2, il massimo osservato. Dal momento che i valori di leverage risultano essere contenuti e data la dipendenza di *DFFITS* sia dal leverage che dal valore predetto nel punto considerato, si può concludere che tale anomalia sia dovuta ad un limite nella predizione del punto, cioè ad un residuo elevato. Effettivamente dal grafico dei residui esternamente standardizzati, si è notato che il punto corrispondente alla prova 19 si avvicina molto più degli altri ai limiti; d'altra parte si è deciso di accostare ai dati un modello di secondo ordine, con un valore di R^2_{pred} inferiore a quello di un modello lineare consigliato invece dal software. Il punto considerato ha quindi un'influenza superiore agli altri nei confronti del modello di secondo ordine.

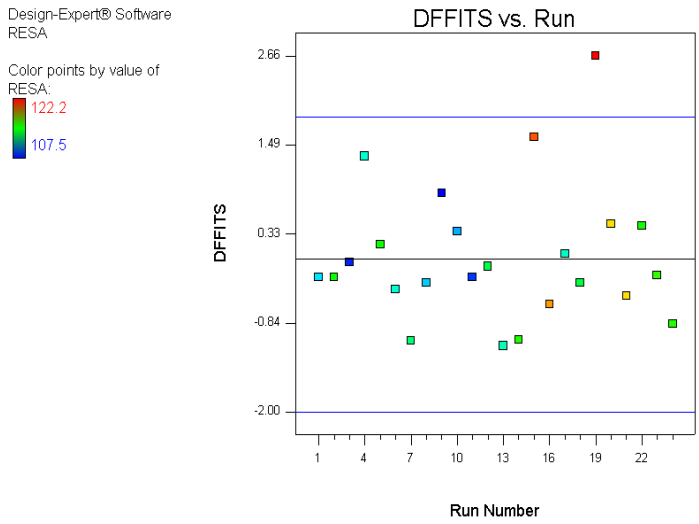

Figura 3.12 *Grafico dei DFFITS rispetto la sequenza sperimentale*

Lo scostamento del punto 19 dal resto dei dati è indicato anche dal grafico di Figura 3.13, che riporta la distanza di Cook in funzione della sequenza delle prove. Dato che essa misura quanto la regressione cambia non considerando il dato nell'analisi, si spiega perché tale dato risulta più spostato verso i limiti, anche se non si ravvisano condizioni di anomalia o tali per cui ripetere la prova.

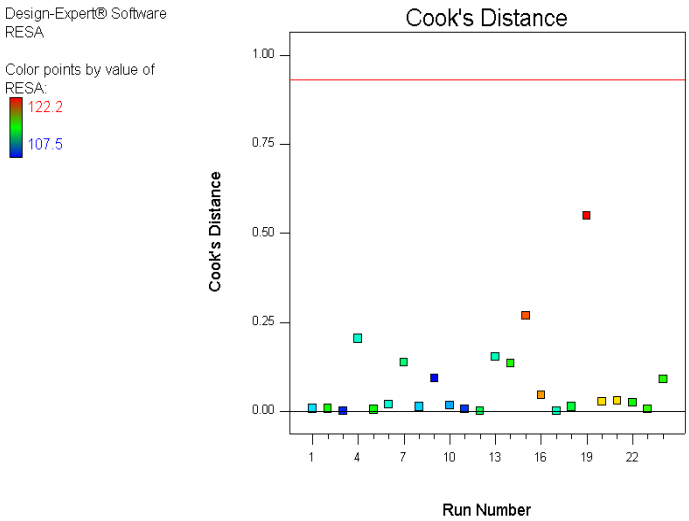

Figura 3.13 *Grafico della distanza di Cook rispetto alla sequenza sperimentale*

Dalla diagnostica risulta che il modello può essere utilizzato per descrivere la risposta. Si è ora in grado di descrivere l'andamento della risposta in funzione dei fattori considerati. Ciò viene fatto generando le superfici di risposta riportata in Figura 3.14 e Figura 3.15 per 3 e 4 passaggi.

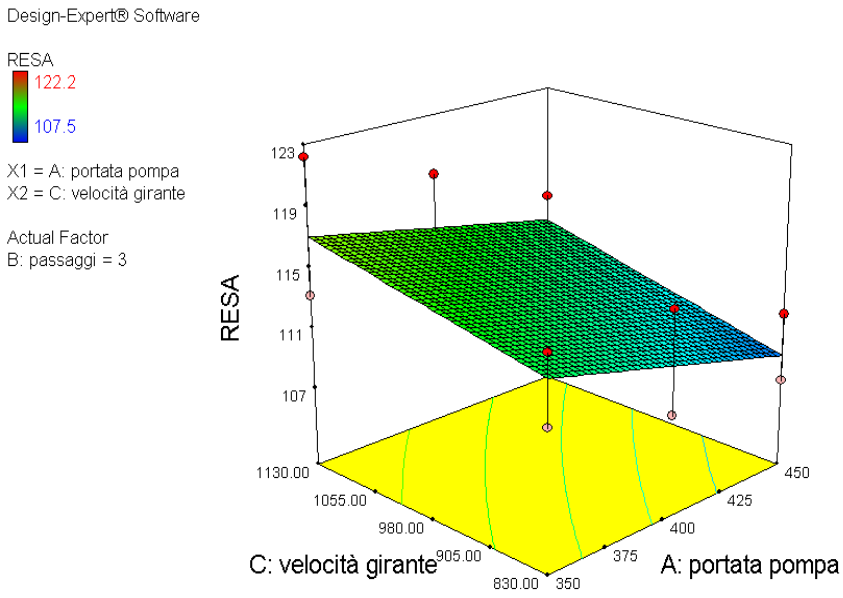

Figura 3.14 *Superficie di risposta per la risposta " resa" a 3 passaggi*

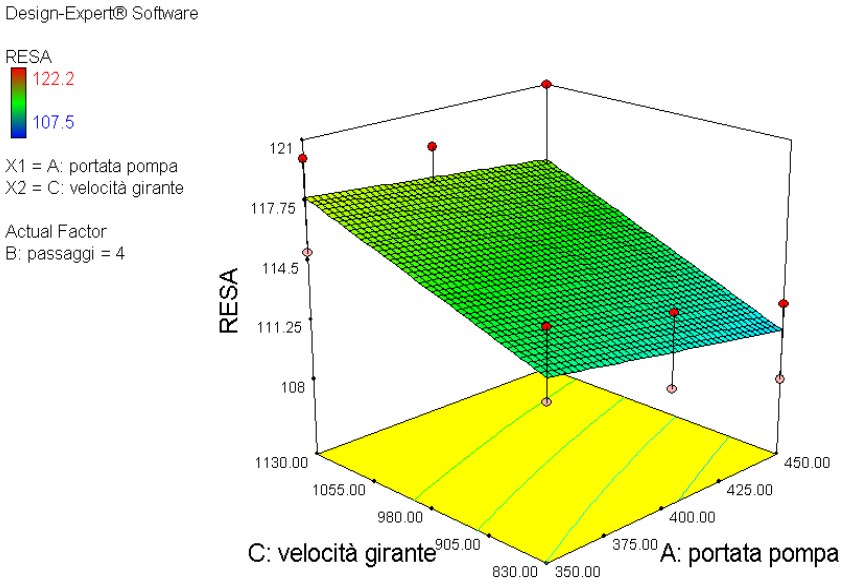

Figura 3.15 *Superficie di risposta per la risposta "resa" a 4 passaggi*

Dalle superfici si può osservare che, nonostante l'impiego di un modello di secondo ordine, non sono presenti curvature, a indicare il fatto peraltro già noto dall'analisi della varianza, che le interazioni tra i fattori non sono significative. I migliori risultati di resa si ottengono in entrambi i casi per basse portate della pompa e alte velocità della girante. Il risultato era aspettato: a basse portate infatti aumenta il tempo di permanenza del prodotto in macchina e quindi la raffinazione del prodotto. Allo stesso modo un'alta velocità della girante aumenta lo sforzo di taglio e gli impatti tra le microsfere e la massa da disperdere, determinando anche in questo caso una migliore raffinazione. Il numero di passaggi sembra influire meno rispetto agli altri fattori; in effetti dall'analisi della varianza di Tabella 3.8 risultava un *p*-value leggermente superiore rispetto gli altri fattori seppure al di sotto di un livello di significatività del 5%. Per quattro passaggi si ottengono tuttavia risultati leggermente superiori. Dalle superfici si può apprezzare anche la bontà del fitting: i punti in rosso e rosa rappresentano infatti i punti sperimentali, il cui valor medio è in entrambi i casi ben approssimato dalle superfici.

3.5.2 Consumo di energia

La procedura vista per la resa si ripropone per le successive risposte. Per quanto riguarda il consumo di energia della girante, il software suggerisce di applicare un modello lineare (Tabella 3.11).

Tabella 3.11 *Suggerimento del modello per la risposta "lavoro"*

Response 3 lavoro Transform: None

*** **WARNING: The Quadratic Model is Aliased!** ***

*** **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	19159,15	1	19159,15042			
Block vs Mean	510,60	1	510,60375			
Linear vs Block	1142,90	3	380,9654861	24,35681896	< 0.0001	Suggested
2FI vs Linear	29,34	3	9,78	0,584230754	0.6340	
Quadratic vs 2FI	7,60	1	7,600208333	0,438070589	0.5181	Aliased
Cubic vs Quadratic	11,69104167	3	3,897013889	0,188149344	0.9024	Aliased
Residual	248,548125	12	20,71234375			
Total	21109,83	24	879,57625			

"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	
Linear	3,95487291	0,793636302	0,76105256	0,660538384	488,850469	Suggested
2FI	4,091449735	0,814010229	0,744264064	0,524021919	685,444532	
Quadratic	4,165246425	0,819287873	0,734955547	0,493341131	729,627193	Aliased
Cubic	4,551081602	0,827406225	0,683578078	0,270189483	1050,982488	Aliased

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

La successiva analisi della varianza, le prestazioni del modello e il calcolo delle equazioni sono riportate in Tabella 3.12, 3.13 e 3.14.

Tabella 3.12 *ANOVA per la risposta "lavoro"*

Response 3 lavoro
ANOVA for Response Surface Linear Model
Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Block	510,60	1	510,60375			
Model	1142,90	3	380,9654861	24,35681896	< 0.0001	significant
A-portata pompa	328,52	1	328,515625	21,00346592	0.0002	
B-passaggi	231,26	1	231,2604167	14,78550763	0.0011	
C-velocità girante	583,12	1	583,1204167	37,28148334	< 0.0001	
Residual	297,18	19	15,64101974			
Cor Total	1950,68	23				

The Model F-value of 24.36 implies the model is significant. There is only a 0.01% chance that a "Model F-Value" this large could occur due to noise.

Values of "Prob > F" less than 0.0500 indicate model terms are significant.
In this case A, B, C are significant model terms.
Values greater than 0.1000 indicate the model terms are not significant.
If there are many insignificant model terms (not counting those required to support hierarchy), model reduction may improve your model.

Tabella 3.13 *Determinazione dei coefficienti di regressione per la risposta "lavoro"*

Std. Dev.	3.95	R-Squared	0.7936
Mean	28.25	Adj R-Squared	0.7611
C.V. %	14.00	Pred R-Squared	0.6605
PRESS	488.85	Adeq Precision	19.031

The "Pred R-Squared" of 0.6605 is in reasonable agreement with the "Adj R-Squared" of 0.7611.

"Adeq Precision" measures the signal to noise ratio. A ratio greater than 4 is desirable. Your ratio of 19.031 indicates an adequate signal. This model can be used to navigate the design space.

Factor	Coefficient Estimate	df	Standard Error	95% CI Low	95% CI High	VIF
Intercept	28.25	1	0.81	26.56	29.94	
Block 1	-4.61	1				
Block 2	4.61					
A-portata pompa	-4.53	1	0.99	-6.60	-2.46	1.00
B-passaggi	3.10	1	0.81	1.41	4.79	1.00
C-velocità girante	4.93	1	0.81	3.24	6.62	1.00

Final Equation in Terms of Coded Factors:

lavoro	=
+28.25	
-4.53	* A
+3.10	* B
+4.93	* C

Tabella 3.14 *Equazioni del modello per la risposta "lavoro"*

Final Equation in Terms of Actual Factors:

passaggi 3
lavoro =
+29.19611
-0.090625 * portata pompa
+0.032861 * velocità girante

passaggi 4
lavoro =
+35.40444
-0.090625 * portata pompa
+0.032861 * velocità girante

The Diagnostics Case Statistics Report has been moved to the Diagnostics Node.
In the Diagnostics Node, Select Case Statistics from the View Menu.

Proceed to Diagnostic Plots (the next icon in progression). Be sure to look at the:
1) Normal probability plot of the studentized residuals to check for normality of residuals.
2) Studentized residuals versus predicted values to check for constant error.
3) Externally Studentized Residuals to look for outliers, i.e., influential values.
4) Box-Cox plot for power transformations.

If all the model statistics and diagnostic plots are OK, finish up with the Model Graphs icon.

In questo caso tutti i fattori considerati sono significativi per quanto riguarda l'effetto sulla risposta. Si osserva dagli indici statistici in Tabella 3.13 che per il lavoro il modello è meno predittivo e fitta meno i dati rispetto al caso della resa.

Dalla diagnostica del modello riportata in Figura 3.16 e Figura 3.17, non risultano anomalie. Il grafico di probabilità normale mostra un andamento rettilineo, e non sono riscontrabili particolari andamenti nei grafici relativi ai residui. Come ci si può aspettare quindi, il diagramma di Box-Cox non consiglia alcuna trasformazione stabilizzatrice della varianza. Il valore corrente di λ è all'interno dell'intervallo di confidenza al 95% in cui si trova il minimo, quindi è accettabile.

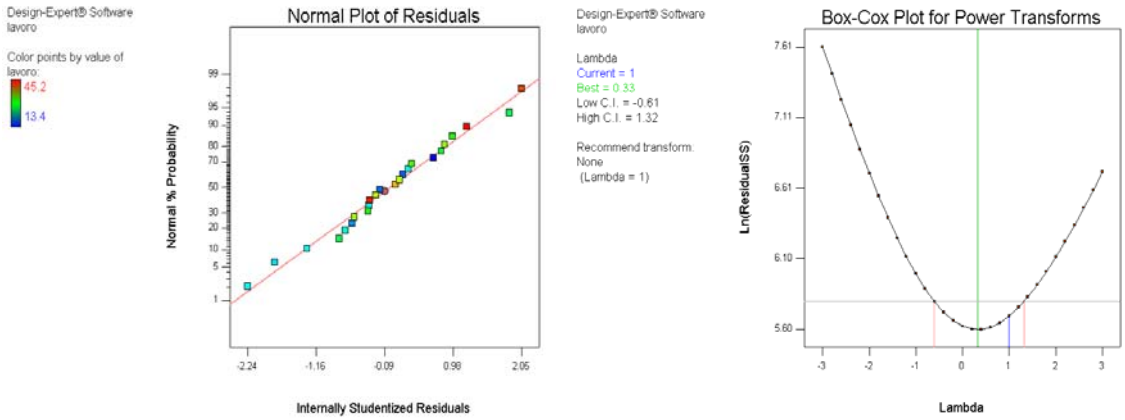

Figura 3.16 *Grafici di probabilità normale e di Box-Cox per la risposta "lavoro"*

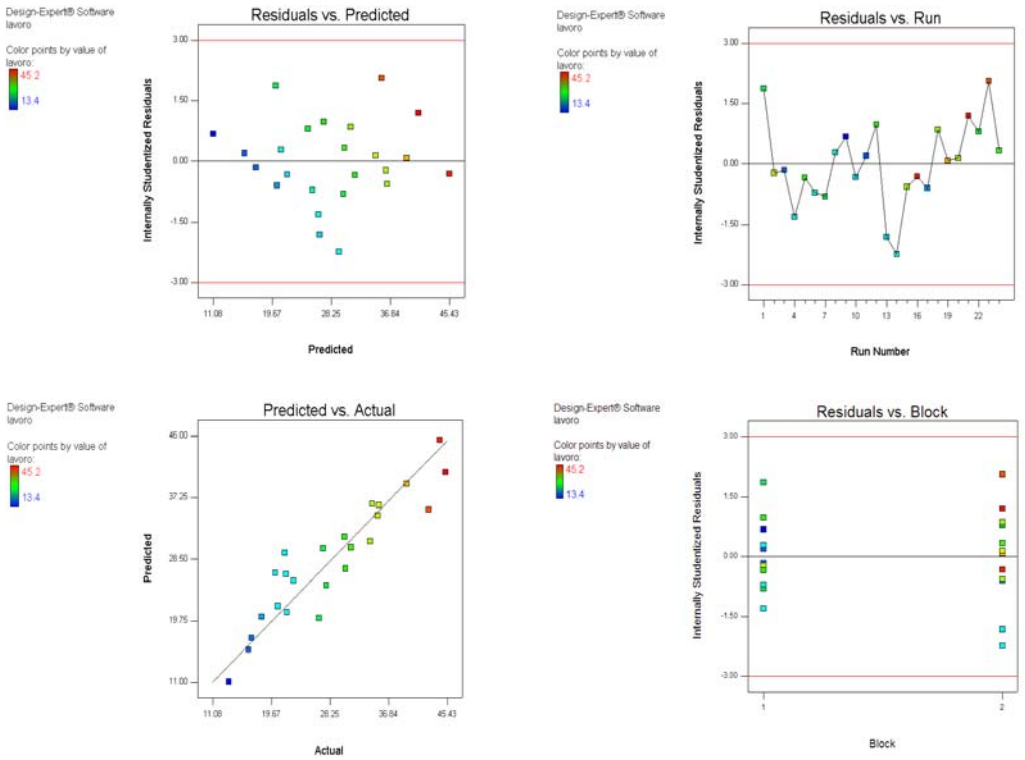

Figura 3.17 *Diagnostica del modello per la risposta "lavoro"*

Si può inoltre notare dal grafico relativo ai residui all'interno dei due blocchi (Figura 3.17, in basso a destra), che per la risposta considerata i blocchi hanno un'importanza inferiore rispetto al caso della resa.
Dalle misure di influenza riportate in Figura 3.18 non risultano punti particolarmente influenti per l'accostamento del modello e il relativo calcolo dei coefficienti.

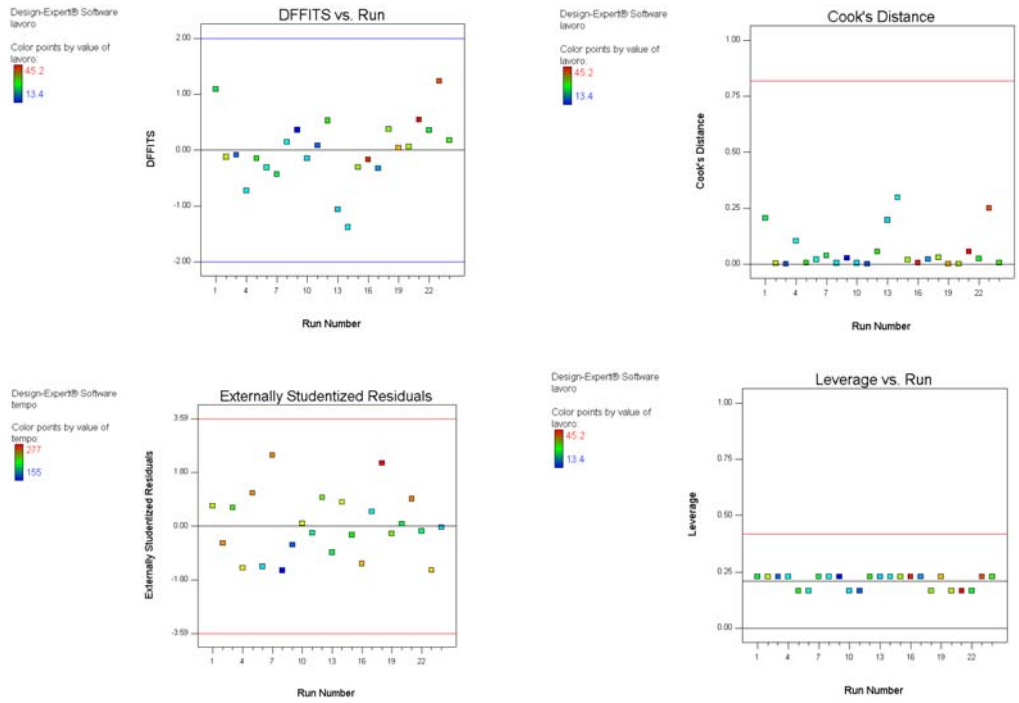

Figura 3.18 *Grafici delle misure di influenza per la risposta "lavoro"*

Dopo aver validato il modello e dopo averne verificata l'adeguatezza a descrivere i dati sperimentali, si passa alla costruzione delle superfici di risposta (Figura 3.19 e 3.20). Dal momento che si è scelto un modello lineare si ottengono due piani. Come si può notare i risultati più alti si hanno per 4 passaggi, per basse portate della pompa e alte velocità della girante. Anche questo risultato era aspettato: con più passaggi e basse portate si aumenta il tempo di lavorazione, quindi aumenta il consumo. Giustamente alte velocità della girante richiedono maggiore energia.

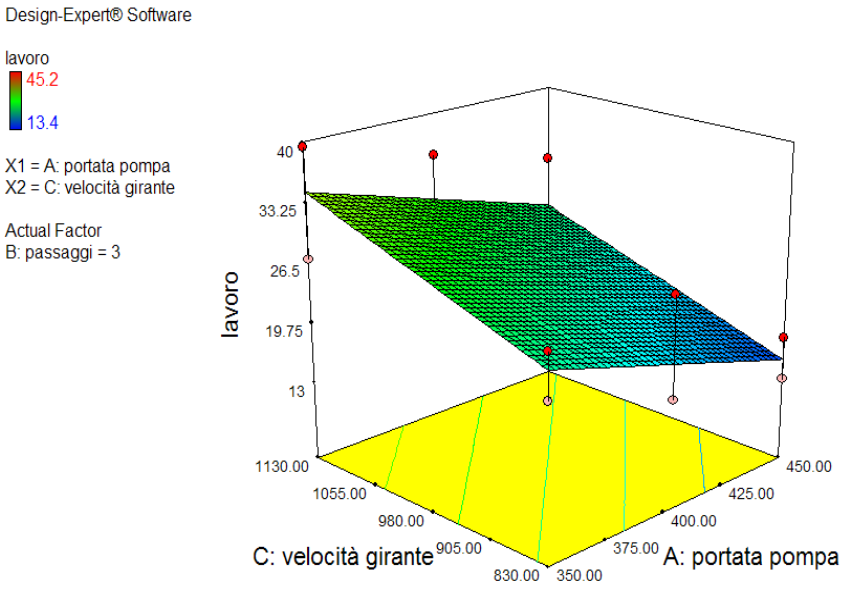

Figura 3.19 *Superficie di risposta per la risposta "lavoro" a 3 passaggi*

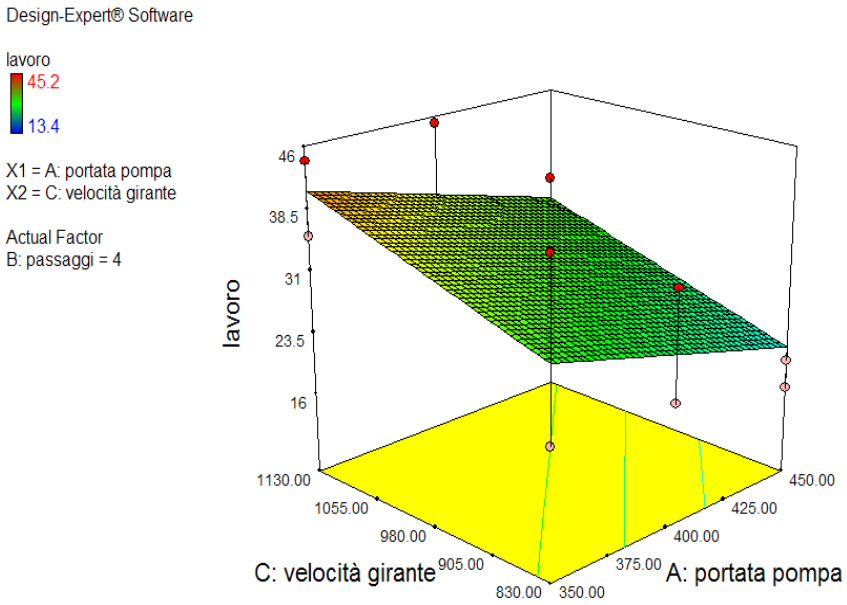

Figura 3.20 *Superficie di risposta per la risposta "lavoro" a 4 passaggi*

3.5.3 Tempo di lavorazione

Anche per il tempo di lavorazione viene consigliato un modello lineare (Tabella 3.15).

Tabella 3.15 *Suggerimento del modello per la risposta " tempo"*

***** 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	1211402,67	1	1211402,667			
Block vs Mean	150,00	1	150			
Linear vs Block	20396,40	3	6798,798611	22,9324706	< 0.0001	Suggested
2FI vs Linear	1601,29	3	533,7638889	2,118296739	0.1381	
Quadratic vs 2FI	266,02	1	266,0208333	1,05966805	0.3196	Aliased
Cubic vs Quadratic	1581,60	3	527,2013889	2,896683296	0.0790	Aliased
Residual	2184,02	12	182,0017361			
Total	1237582,00	24	51565,91667			

"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	
Linear	17,21831568	0,78359271	0,749423137	0,653474142	9019,837078	Suggested
2FI	15,87381065	0,845111445	0,787028237	0,658174681	8897,48516	
Quadratic	15,84429445	0,855331485	0,787819511	0,636611451	9458,761669	Aliased
Cubic	13,49080191	0,916093862	0,84617208	0,658989963	8876,263934	Aliased

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

L'analisi della varianza (Tabella 3.16) mostra che solo portata della pompa e numero di passaggi sono significativi. Il modello ha buone prestazioni sia in termini di fitting che di previsione.

Tabella 3.16 *ANOVA per la risposta "tempo"*

Response 4 tempo

ANOVA for Response Surface Linear Model
Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Block	150,00	1	150			
Model	20396,40	3	6798,798611	22,9324706	< 0.0001	significant
A-portata pompa	10868,06	1	10868,0625	36,65817125	< 0.0001	
B-passaggi	9520,17	1	9520,166667	32,11169424	< 0.0001	
C-velocità girante	8,17	1	8,166666667	0,027546314	0.8699	
Residual	5632,94	19	296,4703947			
Cor Total	26179,33	23				

The Model F-value of 22.93 implies the model is significant. There is only a 0.01% chance that a "Model F-Value" this large could occur due to noise.

Values of "Prob > F" less than 0.0500 indicate model terms are significant.
In this case A, B are significant model terms.
Values greater than 0.1000 indicate the model terms are not significant.
If there are many insignificant model terms (not counting those required to support hierarchy), model reduction may improve your model.

Std. Dev.	17,21831568	R-Squared	0,78359271
Mean	224,6666667	Adj R-Squared	0,749423137
C.V. %	7,663938728	Pred R-Squared	0,653474142
PRESS	9019,837078	Adeq Precision	12,4856059

The "Pred R-Squared" of 0.6535 is in reasonable agreement with the "Adj R-Squared" of 0.7494.

"Adeq Precision" measures the signal to noise ratio. A ratio greater than 4 is desirable. Your ratio of 12.486 indicates an adequate signal. This model can be used to navigate the design space.

Successivamente vengono determinati i coefficienti del modello di regressione (Tabella 3.17). Da notare che l'intervallo di confidenza al 95% per il coefficiente relativo alla velocità della girante include 0, a conferma del fatto che tale fattore ha un'influenza limitata sulla risposta in esame.

Tabella 3.17 *Calcolo delle equazioni del modello per la risposta "tempo"*

Coefficient Factor	Estimate	Standard df	95% CI Error	95% CI Low	High	VIF
Intercept	224.67	1	3.51	217.31	232.02	
Block 1	-2.50	1				
Block 2	2.50					
A-portata pompa	-26.06	1	4.30	-35.07	-17.05	1.00
B-passaggi	19.92	1	3.51	12.56	27.27	1.00
C-velocità girante	0.58	1	3.51	-6.77	7.94	1.00

Final Equation in Terms of Coded Factors:

tempo =
+224.67
-26.06 * A
+19.92 * B
+0.58 * C

Final Equation in Terms of Actual Factors:

passaggi 3
tempo =
+409.43889
-0.52125 * portata pompa
+3.88889E-003 * velocità girante

passaggi 4
tempo =
+449.27222
-0.52125 * portata pompa
+3.88889E-003 * velocità girante

The Diagnostics Case Statistics Report has been moved to the Diagnostics Node.
In the Diagnostics Node, Select Case Statistics from the View Menu.

Proceed to Diagnostic Plots (the next icon in progression). Be sure to look at the:
1) Normal probability plot of the studentized residuals to check for normality of residuals.
2) Studentized residuals versus predicted values to check for constant error.
3) Externally Studentized Residuals to look for outliers, i.e., influential values.
4) Box-Cox plot for power transformations.

If all the model statistics and diagnostic plots are OK, finish up with the Model Graphs icon.

Si passa quindi alla fase di diagnostica per verificare le assunzioni fatte per il modello. Dai grafici di Figura 3.21 si osserva che i residui seguono la distribuzione normale, mentre non viene consigliata nessuna trasformazione della risposta, come si osserva dal grafico di Box-Cox (Figura 3.21), dato che il valore corrente di λ cade all'interno dell'intervallo di confidenza del valore minimo. Effettivamente, dai successivi grafici dei residui (Figura 3.22), non si osservano andamenti sistematici o anomalie, segno di indipendenza dei residui e costanza della varianza. In questo caso inoltre si può notare come non ci sia grande variabilità nelle osservazioni all'interno dei singoli blocchi (Figura 3.22, in basso a destra), pertanto è inferiore l'effetto del blocco sulla riduzione della variabilità della risposta, fatto osservabile anche dai "*Mean Square*" dell'analisi della varianza.

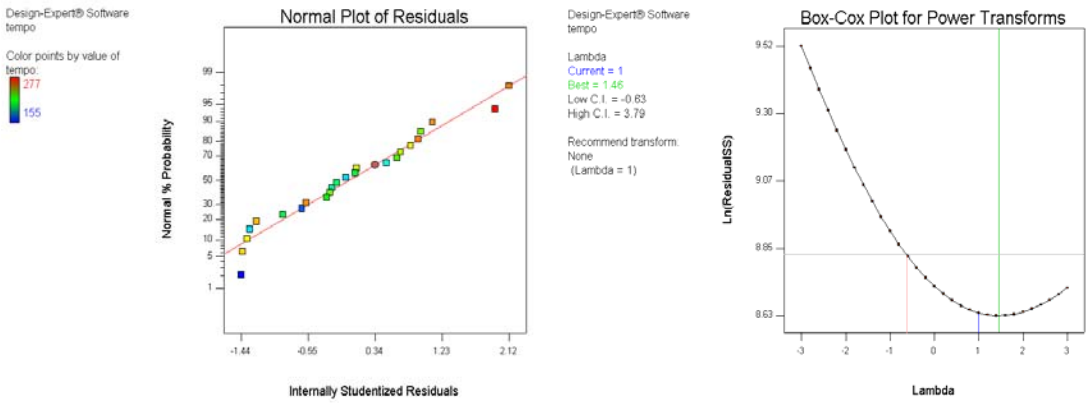

Figura 3.21 *Grafici di probabilità normale e Box-Cox per la risposta "tempo"*

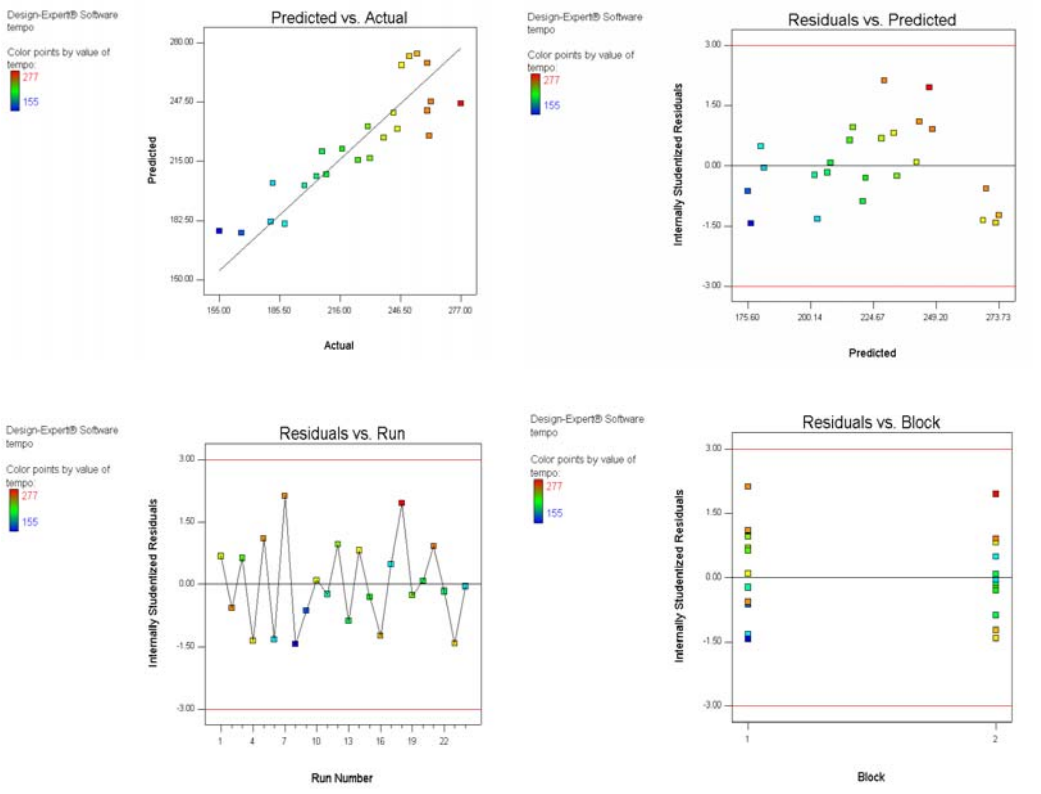

Figura 3.22 *Diagnostica del modello per la risposta "tempo"*

Anche le misure di influenza (Figura 3.23), non mettono in evidenza alcun punto come influente sul modello. Non si osservano situazioni particolari.

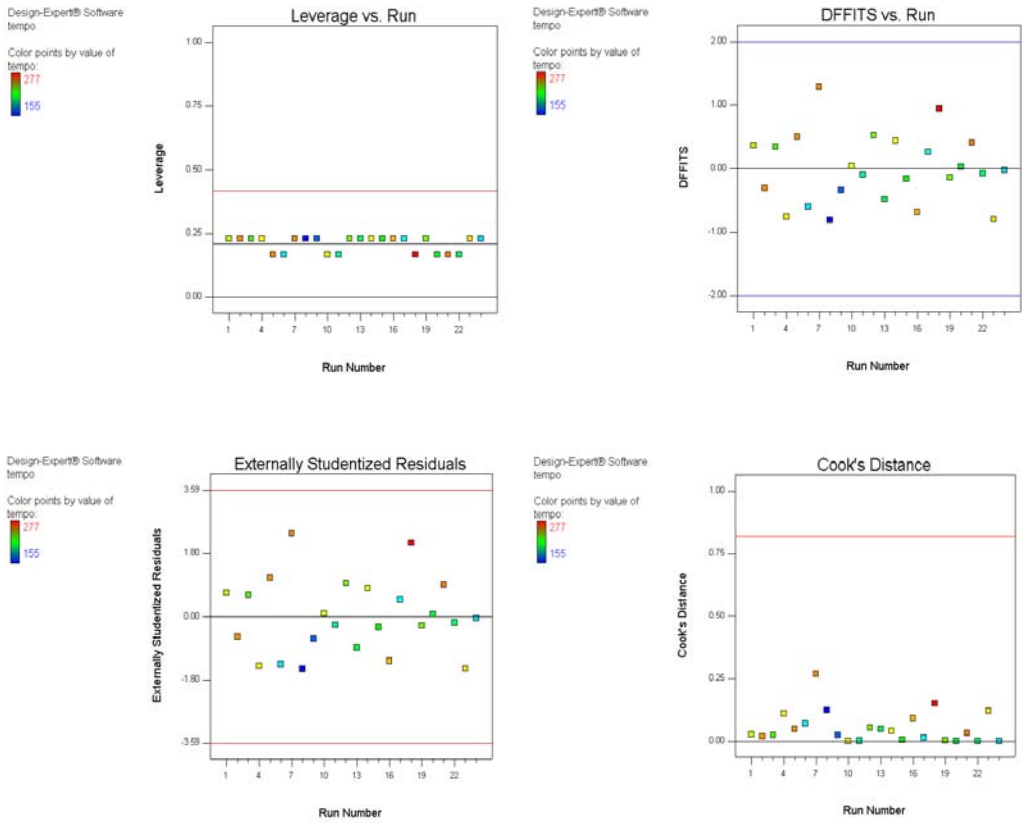

Figura 3.23 *Grafici delle misure di influenza per la risposta "tempo"*

Non avendo riscontrato problemi riguardo l'adeguatezza del modello e l'influenza dei punti, si passa alla costruzione delle superfici di risposta (Figura 3.24 e 3.25). Anche in questo caso si ottengono due piani (modello lineare). Si osserva che tempi di lavorazione più lunghi si osservano come aspettato per 4 passaggi e basse portate della pompa. Si può notare come la velocità della girante sia ininfluente sulla risposta.

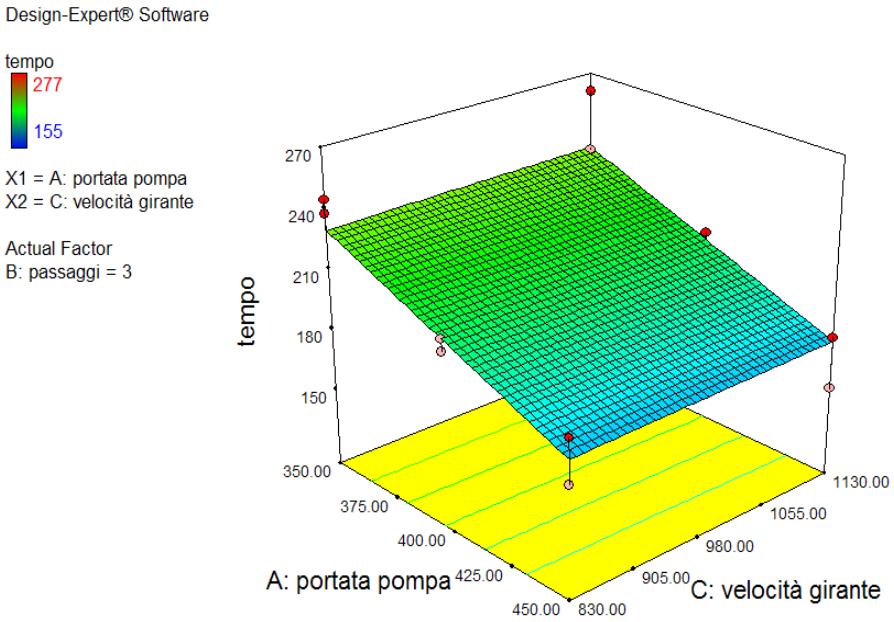

Figura 3.24 *Superficie di risposta per la risposta "tempo" a 3 passaggi*

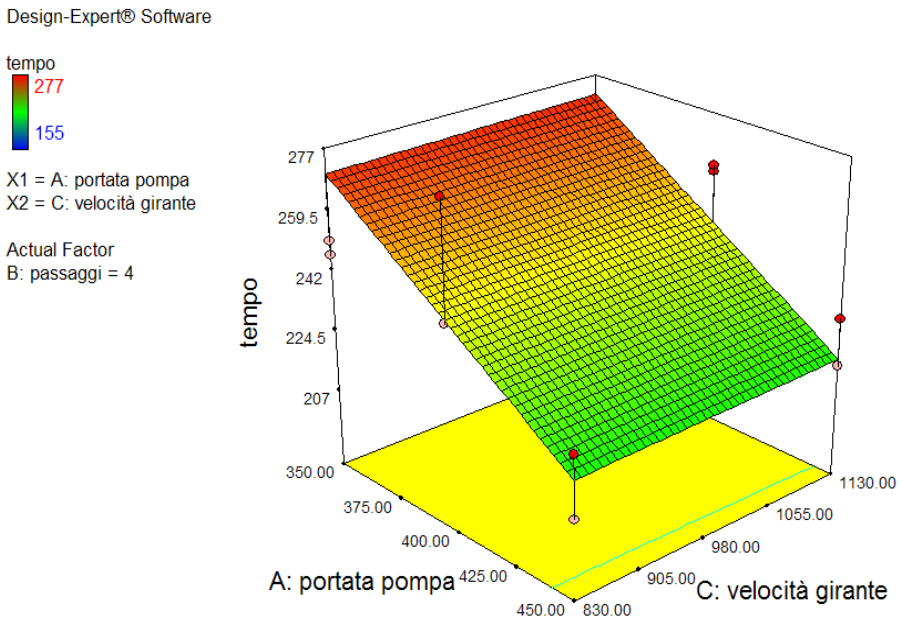

Figura 3.25 *Superficie di risposta per la risposta "tempo" a 4 passaggi*

3.5.4 Viscosità

Nel caso della viscosità il software consiglia di accostare ai dati un modello delle medie (Tabella 3.18)

Tabella 3.18 *Suggerimento del modello per la risposta "viscosità"*

Response 2 **viscosità (a resa)** **Transform:** **None**

***** WARNING: The Quadratic Model is Aliased! *****

***** 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	55495209,38	1	55495209,38			Suggested
Block vs Mean	9683751,04	1	9683751,042			
Linear vs Block	2870008,33	3	956669,4444	0,460484801	0.7131	
2FI vs Linear	5875601,04	3	1958533,681	0,932707115	0.4477	
Quadratic vs 2FI	567675,00	1	567675	0,257801833	0.6190	Aliased
Cubic vs Quadratic	1106860,42	3	368953,4722	0,138691844	0.9349	Aliased
Residual	31922869,79	12	2660239,149			
Total	107521975,00	24	4480082,292			

"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	
Linear	1441,362773	0,067779972	-0,079412664	-0,493364263	63233544,75	
2FI	1449,081718	0,206541964	-0,0910048	-0,832754442	77604348,05	
Quadratic	1483,907684	0,219948543	-0,14407547	-1,018965381	85489080,56	Aliased
Cubic	1631,023957	0,246088874	-0,382170397	-2,026231189	128139751,3	Aliased

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

Ciò significa che la media dei valori osservati è più predittiva di qualsiasi altro modello accostabile ai dati. Dai valori degli indici di fitting e di predizione riportati nella parte inferiore di Tabella 3.18, è possibile vedere che tutti i modelli considerati non riescono a fittare i dati e non hanno capacità predittiva. Appare quindi improbabile proseguire nell'analisi. I valori elevati di deviazione standard che si leggono, indicano una consistente variabilità nei dati. D'altra parte osservando i dati raccolti è facile riscontrare una repentina deriva delle osservazioni nel corso delle prove. Indicativo a questo proposito è il grafico di Figura 3.26 dei valori di viscosità osservati contro la sequenza sperimentale. Dal grafico si vede come in corrispondenza delle ultime prove relative al secondo blocco, siano stati ottenuti i valori di viscosità più elevati che si scostano dagli altri. Questa deriva non è quindi causata dal blocco, ma si pensa che sia dovuta a inconvenienti legati alle lavorazioni che il prodotto ha subito precedentemente alla raffinazione, sulla quale sono state condotte la prove. In particolare è ipotizzabile che per il secondo lotto di materia prima il prodotto sia stato lasciato a riposo in cisterna per un tempo troppo elevato, senza essere adeguatamente agitato prima dei prelievi delle 12 bacinelle da 300 kg con cui sono state effettuate le prove. Si stanno tuttora effettuando ulteriori test in laboratorio per verificare tale ipotesi.

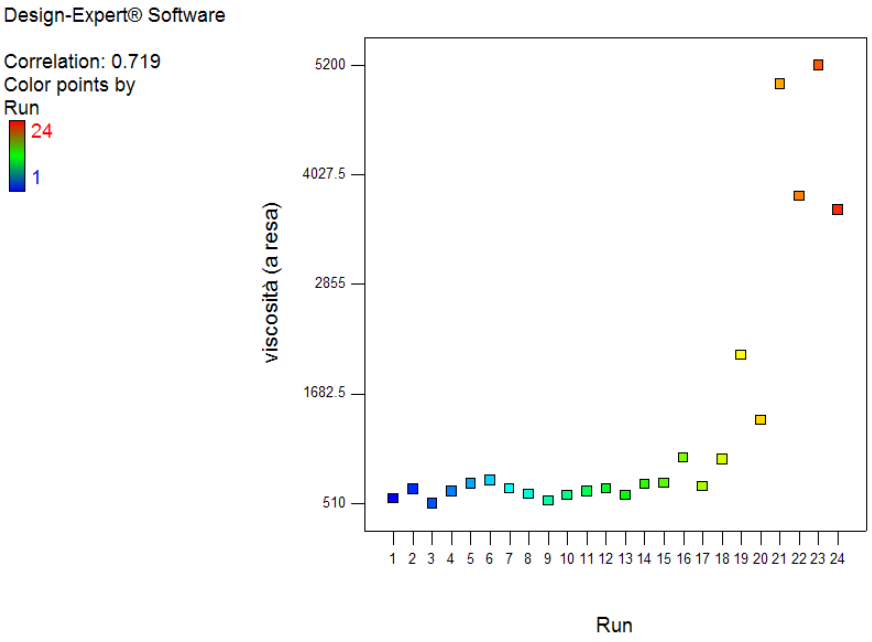

Figura 3.26 *Grafico dei valori osservati di viscosità rispetto alla sequenza sperimentale. E'facilmente riscontrabile la deriva nelle prove del secondo lotto.*

3.6 Ottimizzazione

L'ottimizzazione delle condizioni operative si basa sull'ottimizzazione di una funzione di desiderabilità che tiene conto dei singoli obiettivi che si vogliono ottenere per le risposte (2.57). In linea con gli obiettivi preposti di abbattere i costi migliorando le prestazioni qualitative del processo, sono stati posti i seguenti vincoli per le risposte:

- Resa ad un preciso target pari a 116, considerato per esperienza un buon risultato.
- Consumo di energia minimo.
- Tempo di lavorazione minimo.

É possibile attribuire alle varie risposte un differente peso a seconda di ciò che si reputa più importante per l'ottimizzazione: un peso elevato significa dare importanza al punto in sè, un basso valore significa privilegiare l'intervallo attorno al punto. In particolare nel caso in esame è stato attribuito il peso maggiore (pari a 5) al consumo di energia in quanto è desiderabile abbattere i consumi, mentre alle altre variabili è stato dato un peso comune pari a 1. É inoltre possibile attribuire a ciascuna di esse una diversa "importanza": tale parametro si sceglie a seconda di quanto si voglia privilegiare il raggiungimento dell'ottimo di una variabile rispetto ad un'altra (si veda (2.61)). Esso è un ulteriore grado di libertà che il software permette di scegliere. In questo caso è stato privilegiato il tempo di lavorazione rispetto alle altre variabili. I risultati ottenuti sono riportati in Tabella 3.19. Il software trova le migliori soluzioni per 3 e 4 passaggi. Si noti che le migliori condizioni operative si ottengono per 3 passaggi con la pompa alla portata massima e la velocità

della girante al livello minimo. Di seguito si riportano i grafici di desiderabilità a 3 e 4 passaggi (Figura 3.27 e 3.28). Si noti ancora come in genere sia più desiderabile operare su 3 passaggi. L'andamento nei due casi è comunque lo stesso.

Tabella 3.19 *Ottimizzazione delle condizioni operative*

Constraints

Name	Goal	Lower Limit	Upper Limit	Lower Weight	Upper Weight	Importance
portata pompa	is in range	350	450	1	1	3
passaggi	is in range	3	4	1	1	3
velocità girante	is in range	830	1130	1	1	3
RESA	is target = 116	107,5	122,2	1	1	2
lavoro	minimize	13,4	45,2	1	5	3
tempo	minimize	155	277	1	1	4

Solutions for 2 combinations of categoric factor levels

Number	portata pompa	passaggi	velocità girante	RESA	lavoro	tempo	Desirability
1	450	3	830	109,21875	15,68958333	178,1041667	0,563793928
2	450	3	835,76	109,3022616	15,87902539	178,1266021	0,563629805
3	450	3	856,34	109,6003888	16,55530959	178,206838	0,56068708
4	450	4	830	110,75625	21,89791667	217,9375	0,348597648

4 Solutions found

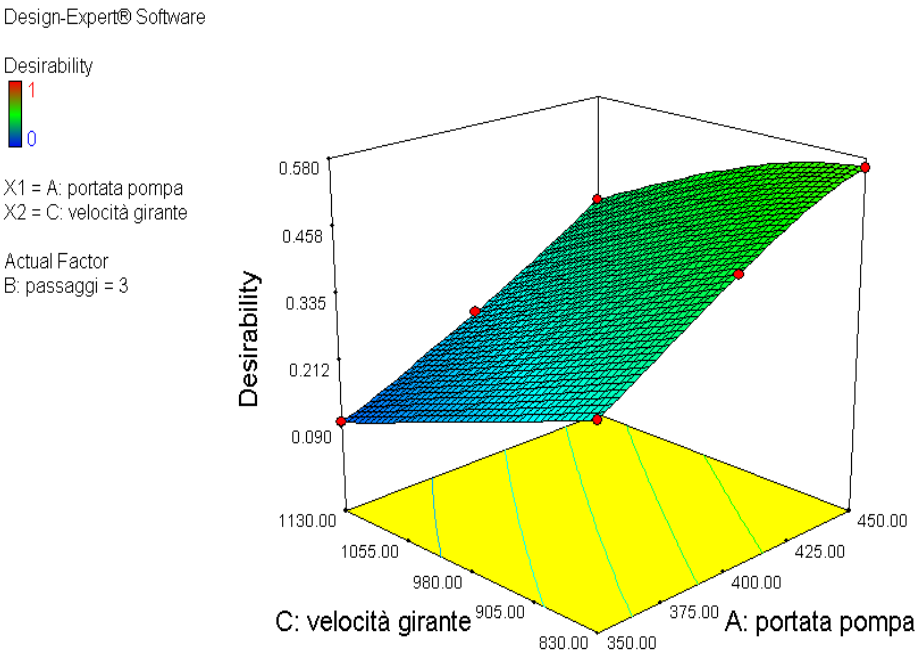

Figura 3.27 *Superficie di risposta per la funzione desiderabilità a 3 passaggi*

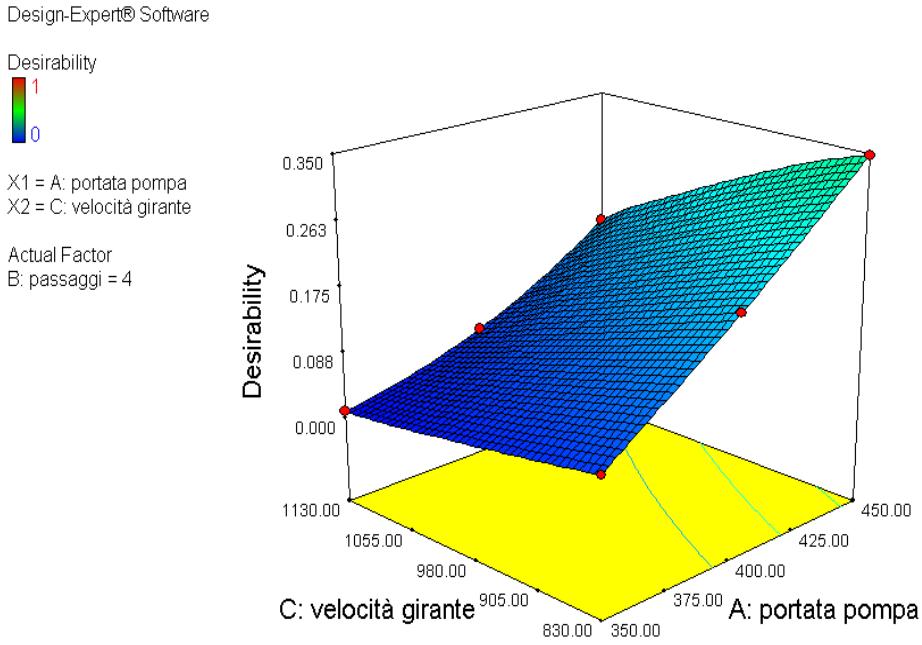

Figura 3.28 *Superficie di risposta per la funzione desiderabilità a 4 passaggi*

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 delle prove 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 in cui l'errore standard legato al piano, cioè il rumore di fondo, è superiore e quindi zone di confine nel dominio sperimentale (come si può ravvisare in Figura 3.3). Sono stati quindi scelti i seguenti livelli per i fattori considerati:

- 1° check point: portata della pompa = 450 kg/h
  velocità della girante = 1130 rpm
  n° passaggi = 4
- 2° check point: portata della pompa = 425 kg/h
  velocità della girante = 1130 rpm
  n° passaggi = 3
- 3° check point: portata della pompa = 375 kg/h
  velocità della girante = 1005 rpm
  n° passaggi = 3

I risultati ottenuti sono riportati in Tabella 3.20.

Tabella 3.20 *Risultati ottenuti per i singoli check points*

Check point 1 :

resa: 118 **lavoro: 30** **tempo: 208**

Factor	Name	Level	Low Level	High Level	Std. Dev.	Coding
A	portata pompa	450	350	450	0	Actual
B	passaggi	4	3	4	N/A	Actual
C	velocità girante	1130	830	1130	0	Actual

Response	Prediction	SE Mean	95% CI low	95% CI high	SE Pred	95% PI low	95% PI high
RESA	116,61875	0,708557907	115,1166743	118,1208257	1,385503306	113,6816142	119,5558858
viscosità (a resa)	1,93644E-05	3,08942E-06	1,28981E-05	2,58306E-05	7,77487E-06	3,09138E-06	3,56374E-05
lavoro	31,75625	1,712510205	28,17192495	35,34057505	4,309722838	22,73589643	40,77660357
tempo	219,1041667	7,455749393	203,4991038	234,7092295	18,7632245	179,8322864	258,3760469

Check point 2 :

resa: 115,5 **lavoro: 29,8** **tempo: 202**

Factor	Name	Level	Low Level	High Level	Std. Dev.	Coding
A	portata pompa	425	350	450	0	Actual
B	passaggi	3	3	4	N/A	Actual
C	velocità girante	1130	830	1130	0	Actual

Response	Prediction	SE Mean	95% CI low	95% CI high	SE Pred	95% PI low	95% PI high
RESA	114,4489583	0,550190163	113,2826073	115,6153094	1,311592284	111,6685069	117,2294098
viscosità (a resa)	1,37918E-05	2,67552E-06	8,19183E-06	1,93917E-05	7,61987E-06	-2,15683E-06	2,97403E-05
lavoro	27,81354167	1,483077341	24,70942512	30,91765822	4,22380612	18,97301386	36,65406948
tempo	192,3020833	6,456868379	178,7877025	205,8164642	18,3891692	153,8131099	230,7910568

Check point 3 :

resa: 113,9 **lavoro: 29.3** **tempo: 221**

Factor	Name	Level	Low Level	High Level	Std. Dev.	Coding
A	portata pompa	375	350	450	0	Actual
B	passaggi	3	3	4	N/A	Actual
C	velocità girante	1005	830	1130	0	Actual

Response	Prediction	SE Mean	95% CI low	95% CI high	SE Pred	95% PI low	95% PI high
RESA	114,7272569	0,407830845	113,8626942	115,5918197	1,258527355	112,0592981	117,3952158
viscosità (a resa)	1,09523E-05	2,2575E-06	6,22732E-06	1,56773E-05	7,48334E-06	-4,71049E-06	2,66151E-05
lavoro	28,23715278	1,251364166	25,61801748	30,85628808	4,148123915	19,55502964	36,91927591
tempo	217,8784722	5,448059579	206,4755525	229,281392	18,05967187	180,0791446	255,6777998

Dall'esame dei dati in Tabella 3.20 si può concludere che è confermata la capacità predittiva del modello. In ogni caso infatti i risultati sperimentali delle prove di conferma risultano all'interno del

range di valori attesi con intervallo di confidenza del 95%. Da notare a questo proposito che l'output del software presenta 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'intevallo che conterrà il valore vero di una singola osservazione il 95% delle volte.

Conclusioni

Da un'analisi di quanto ottenuto dalle prove e dai check point, è possibile affermare che, per questo tipo di lavorazione, le superfici di risposta calcolate possono essere utilizzate per l'ottimizzazione delle variabili operative di processo in funzione delle risposte desiderate. L'utilizzo di blocchi è stato utile nell'eliminare dai dati componenti di variabilità aggiuntive, soprattutto per quanto riguarda la risposta resa. Si ritiene tuttavia che la variabilità riscontrabile all'interno dei due blocchi non sia dovuta solamente all'utilizzo di due lotti di materia prima diversa, ma a differenze nelle fasi di lavorazione precedenti la raffinazione. Non possono essere condotte conclusioni sulla viscosità, visto che la deriva osservata nei dati raccolti non ne permette la modellazione. L'andamento delle superfici di risposta è così riassumibile:

1. Alti valori di resa coloristica si ottengono per :
    - Bassa portata pompa
    - Elevata velocità di agitazione
    - 4 passaggi di raffinazione.
2. Il lavoro del motore tende al minimo (ottimale) per:
    - alta portata pompa
    - bassa velocità di agitazione
    - 3 passaggi di raffinazione
3. Il tempo di lavorazione tende al minimo (ottimale) per:
    - alta portata pompa
    - 3 passaggi di raffinazione
    - velocità di agitazione ininfluente.

Questi risultati appaiono prevedibili a chi tradizionalmente conosce il processo in esame. Tuttavia non è così per la desiderabilità ottenuta dopo aver posto le condizioni di ottimizzazione. Fissato un valore target per la resa coloristica, la superficie di risposta per la desiderabilità generale risulta tendenzialmente massima per 3 passaggi di raffinazione, massima portata della pompa e velocità della girante tendenzialmente vicina a valori minimi.

In conclusione si può affermare che la conoscenza delle condizioni operative (dopo ottimizzazione) è fondamentale per ridurre tempi di lavorazione e consumo energetico, salvaguardando i parametri qualitativi del prodotto che non risultano mai essere al di fuori dei limiti di specifica. Quanto sopra, rappresenta una prova dell'applicabilità di tecniche DOE su scala industriale relativamente al processo SAMIA s.a.s. in esame. Il passo successivo sarà quindi quello di estendere le condizioni operative ottimali determinate sull'impianto pilota, al processo su larga scala (produzione a batch 14 tons), organizzando piani di sperimentazione evolutiva in produzione (EVOP).

Ringraziamenti

Il lavoro descritto in questa Tesi è stato svolto durante un tirocinio di 300 ore svolto presso la sede di Arzignano (VI) della ditta SAMIA s.a.s.. Intendo pertanto ringraziare l'azienda SAMIA s.a.s. per l'opportunità concessa, nonché tutto il personale per la disponibilità mostrata. Un ringraziamento particolare va a tutto il personale del laboratorio, in particolar modo al responsabile, nonché mio tutor in azienda, dott. Daniele Foletto per il tempo dedicatomi. Si ringraziano anche il responsabile della produzione Adriano Battilana, Stefano Povolo e Nico Orsato.
Un ringraziamento speciale va infine al mio relatore ing. Fabrizio Bezzo per la disponibilità e la collaborazione fornita durante la stesura della Tesi.

Riferimenti bibliografici

Anderson, V. L. e R. A. McLean (1974). *Design of Experiments: A Realistic Approach*. Dekker, New York (U.S.A.).

Autori vari, (1998). *Introduzione alla chemiometria: le componenti principali, la classificazione, la regressione, le reti neurali*. UNICHIM, Manuale N. 186.

Box, G. E. P. e D. R. Cox (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society B*, **26**, 211-243.

Box, G. E. P., W. G. Hunter e J. S. Hunter (1978). *Statistics for Experimenters*. Wiley, New York (U.S.A).

Cook, D. R. (1977). Detection of Influential Observations in Linear Regression. *Technometrics*, **19**, 15-18.

Cook, D. R. (1979). Influential Observations in Linear Regression. *Journal of the American Statistical Association*, **74**, 169-174.

Derringer, G. e R. Suich (1980). Simultaneous optimization of several response functions. *J. Qual. Technol*., **12**, 214-219.

Montgomery, D. C. (2005). *Progettazione e analisi degli esperimenti*. McGraw-Hill, Milano.

Montgomery, D. C., E. A. Peck e G. G. Vining (2001). *Introduction to Linear Regression Analysis* (3rd edition). John Wiley & Sons, New York (U.S.A.).

Montgomery, D. C. e S. R. Voth (1994). Multicollinearity and leverage in mixture experiments. *J. Qual. Technol*., **29**, 96-108.

Smith, W. F. (2005). *Experimental Design for Formulation*. SIAM, Philadelphia (U.S.A).

Todeschini, R. (1998). *Introduzione alla chemiometria*, EdiSES, Napoli.