Luigi Roggia
data scientist @ Kiwi Data Science www.kiwidatascience.com
partner di Gmsl s.r.l. www.gmsl.it

L'importanza di avere un modello matematico

Tutta la Fisica e le applicazioni che da essa derivano, si basano sul fatto che ogni teoria è descritta con modelli matematici; è la presenza di un modello matematico che contribuisce in modo irrinunciabile a rendere scientifica una teoria, un metodo, un processo.

Avere un modello matematico è utile ed importante perchè permette di prevedere e descrivere formalmente il comportamento di un sistema.

In un contesto industriale questo vuol dire non basarsi *solo* sull'esperienza e il buon senso, bensì avvalersi anche di una descrizione oggettiva, formale e capace di fornire previsioni in termini di valori numerici.

Facciamo un esempio molto semplice, immaginiamo di avere una reazione chimica la cui resa aumenti col tempo; lasciamo perdere qualsiasi altra possibile variabile, giusto solo per semplicità di descrizione.

In azienda tutti sanno che maggiore è il numero di ore in cui si lasciano reagire i componenti e maggiore è la resa percentuale della reazione. Questo è interessante e utile, ma assolutamente incompleto, perchè di fatto nessuno sa quale sia il tempo in corrispondenza del quale si ottenga la massima resa percentuale della reazione; non si sa neppure se l'aumento della resa segua un comportamento lineare, oppure no.

Questo è grave, perchè se il comportamento fosse lineare vorrebbe dire che effettivamente mi basterebbe aspettare un tempo suffcientemente lungo per avere una resa del 100%. Basterebbe quindi avere la formula della retta che descrive il processo per calcolare quando si arriva al 100%. Diversamente se non fosse lineare, allora potremmo anche avere una soglia oltre la quale la resa comincia a diminuire, magari perchè si innescano altre reazioni concorrenti.

Supponiamo ora che l'azienda in questione allestisca un esperimento in cui misura la resa della reazione dopo diversi intervalli di tempo, proprio allo scopo di poter creare un modello del processo attraverso il *metodo di regressione,* che è uno dei metodi più comunemente impiegati in qualsiasi campo per ottenere una formula, ovvero un *modello matematico*, a partire da un insieme di punti sperimentali.

Si decide quindi, sempre nel nostro ipotetico esempio, di misurare la resa percentuale ogni ora per un tempo massimo di dieci ore; si raccolgono quindi 10 punti sperimentali, che vengono riportati nella seguente tabella in due colonne. La colonna *time* riporta il tempo, le ore da uno a dieci e la colonna *yield* riporta la resa percentuale corrispondente:

$time$ (hr)	$yield$
1	0.27
2	0.38
3	0.6
4	0.64
5	0.73
6	0.81
7	0.75
8	0.68
9	0.65
10	0.48

Grazie a questi dati, utilizzando un software come **Minitab** o un linguaggio come **R**, è possibile creare un modello di regressione ed ottenere così il nostro modello matematico del processo.

Ecco quindi che finalmente abbiamo un modello, *una formula*, che descrive il nostro processo industriale:

$$y(t) := -0.02\ t^2 + 0.25\ t$$

La resa del processo, y, dipende dal quadrato del tempo, t. La formula non è lineare quindi. Questo vuol dire che esiste un certo tempo t in corrispondenza del quale il nostro processo rende il massimo possibile, mentre *sia prima che dopo* la resa è senz'altro inferiore. Possiamo trovare il tempo di massima resa e quantificarla; si tratta infatti semplicemente di trovare le coordinate del massimo della parabola!

Vedasi box di seguito per i dettagli della formula ed il calcolo...
N.B.: ***questo documento è stato realizzato con Mathcad Prime 3.1*** *www.gmsl.it/mathcad per cui il seguente box può essere collassato per far sì che nella presentazione finale la formula non sia visibile*

Data l'equazione generale della parabola e i corrispondenti valori della nostra equazione:

$ax^2 + bx + c$ $\quad a := -0.02$ $\quad b := 0.25$ $\quad c := 0$

nel nostro esempio, le coordinate del suo vertice sono date dal tempo in cui la resa è massima e dal corrispondente valore della resa:

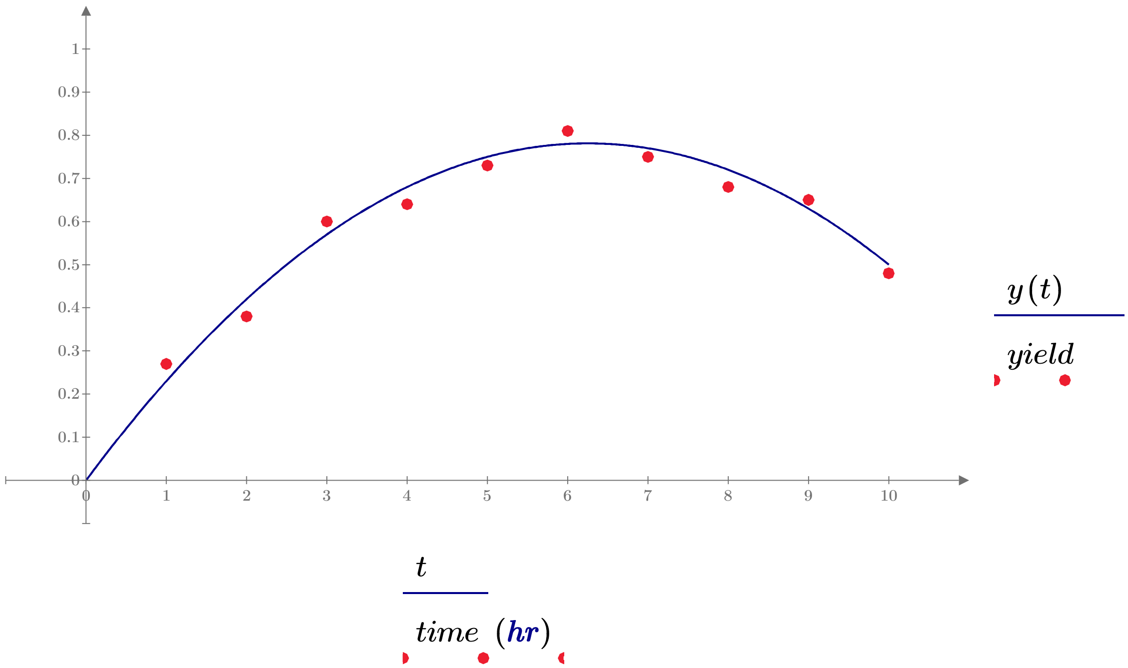
$t_{max\_resa} := \dfrac{-b}{2 \cdot a}\ hr$ $\quad y_{max} := \dfrac{4 \cdot a \cdot c - b^2}{4 \cdot a}$

e quindi eseguendo il calcolo:

$t_{max\_resa} = 6.25\ hr$ $\quad y_{max} = 0.781$

Qundi dopo 6.25 ore si ha la massima resa del nsotro processo, che è una resa del 78%

Il seguente grafico rappresenta il rosso i punti sperimentali raccolti e con la linea blu il modello ottenuto con il metodo di regressione, che è una parabola, come abbiamo visto:



Avere il modello del sistema permette quindi non solo di ottenere valutazioni quantitative, indicazioni sul comportamento e previsioni... Permette anche di capire meglio il processo; in questo esempio magari non era noto il fatto che dopo un certo tempo la resa cominciasse a dimuire. Scoprire questo effetto e quantificarlo può permettere di indivduare aspetti del sistema che prima non erano noti.

Luigi Roggia
data scientist @ Kiwi Data Science www.kiwidatascience.com
partner di Gmsl s.r.l. www.gmsl.it

Il ruolo del design of experiment

Un processo industriale, sia esso in produzione o in ricerca e sviluppo, è molto complesso e non è possibile avere un'equazione esatta che lo descriva. In altre parole nessuno svilupperà mai un modello fisico del processo come si è invece fatto per le leggi di gravitazione, la termodinamica, la meccanica quantistica, ecc...

Ma abbiamo visto che avere uno studio matematico della funzione che descrive il processo è molto importante e utile... Quindi come fare?

Ecco che qui compare il design of experiment (DoE) che fornisce una soluzione!

Per capire come questo avvenga, ricordiamo cosa ci importa fare e sapere relativamente ad un processo industriale:

- sapere quali variabili sono davvero importanti al fine della resa del processo
- trovare quale combinazione dei valori delle diverse variabili ci fornisca la migliore resa
- minimizzare i costi

In sostanza vogliamo ottimizzare il processo.

Ottimizzare il processo vuol sostanzialmente massimizzare, o minimizzare, o portare ad un valore target il risultato; il tutto cercando di non spendere soldi inutilmente.

Per fare questo, tutto quel che servirebbe sarebbe la formula dell'equazione che descrive il processo e a quel punto basterebbe trovare i massimi o i minimi o le zone dove la funzione assume il valore desiderato...!

Non disponendo di tale equazione generale, il DoE non è altro che un insieme di metodi per esplorare in modo ottimale e scientifico la curva che descrive il processo, senza averne la formula esatta, procedendo per piccole campionature attraverso una procedura che ci permette di esplorare il processo fino a giungere alla condizione ricercata, sia essa un massimo, un minimo o un dato valore.

In altre parole col DoE si ottengono modelli *locali* della funzione generale, che è ignota e resterà sempre ignota. Un modello locale è comunque tutto quel che serve, perchè un processo industriale non coprirà mai tutti i valori possibili di una certa variabile.

Inoltre lo studio della pendenza di un modello locale permette di capire in che direzione muoversi per ricercare un estremo (massimo o minimo, ammesso che esistano) o un certo valore desiderato.

Il DoE è sempre un processo iterativo, perchè consiste in una vera e propria esplorazione di una curva ignota.

Luigi Roggia
data scientist @ Kiwi Data Science www.kiwidatascience.com
partner di Gmsl s.r.l. www.gmsl.it

L'idea generale del DoE

Il DoE ha un approccio generale pensato per essere sia efficace che sensato per una realtà industriale.

L'idea è che non sarebbe nè economico, nè sensato campionare valori a caso per raccogliere misure su tutto il range possibile. Allora le variabili vengono *fattorizzate*, ovvero per esempio nel caso del tempo, si decide che si lavora solo tipicamente con due valori del tempo e si acquisiscono i corrispondenti valori della resa in prossimità dei due estremi.

I due valori di tempo utilizzati sono detti *livelli* del fattore tempo.

L'approccio più comune del DoE prevede di utilizzare appunto due livelli per ogni fattore. In questo modo, se si hanno k fattori, tutte le combinazioni possibili di livelli sono 2^k

Con 2^k misure riesco quindi a valutare tutti i possibili casi relativi alla porzione di *spazio sperimentale* in cui mi sono posto.

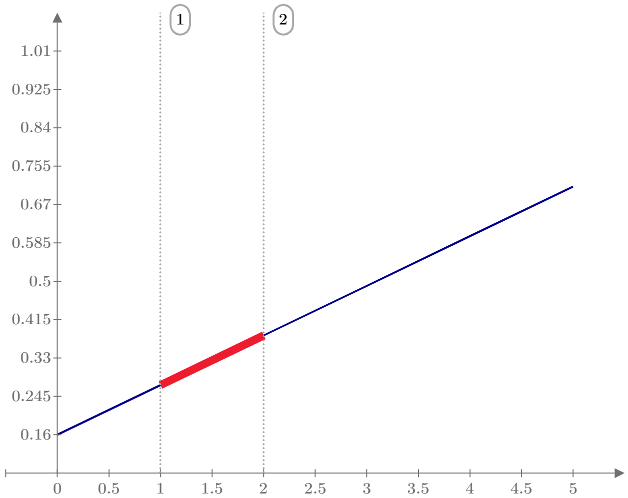
Nel nostro esempio il fattore è uno solo, il tempo; mi bastano quindi $2^1 = 2$ misure per caratterizzare una data porzione di spazio sperimentale.

Supponiamo per esempio di non conoscere la forma generale della curva, la parabola ottenuta precedentemente e di lavorare solo a una e a due ore. Così facendo, otterremmo una equazione di regressione dalla forma lineare, di cui avremmo esplorato una porzione limitata al range tra t=1 e t=2.

Il grafico seguente mostra in blu la retta di regressione individuata e in rosso il range che abbiamo esplorato.

$$y(t) := 0.11 \cdot t + 0.16$$

Questa è la formula del modello matematico ottenuto esplorando tra 1 e 2, con appunto solo 2 misure in corrispondenza di tali valori di tempo



Dal grafico si vede chiaramente che evidentemente al crescere del tempo la resa aumenta.

Si decide quindi di eseguire un secondo esperimento nel verso della crescita, aumentando il tempo.

La prossima campionatura prevede l'esperimento a 4 e a 5 ore...

Tra 4 e 5 ore si ottiene un coefficiente angolare positivo, quindi la resa sta ancora crescendo.

Continuando l'esplorazione, si decide di eseguire un terzo esperimento a 7 e a 8 ore, si scopre in questo caso che il coeffieinte angolare della retta di regressione è diventato negativo...

Questo vuol dire che fra il secondo e il terzo esperimento si è superato un massimo in corrspondenza del quale il verso della pendenza è cambiato.

Evidentemente esiste un massimo di resa fra 5 e 7 ore.

Si esegue pertanto una misura centrale fra 5 e 7, ovvero a 6 ore, perchè se si vuole studiare la curvatura fra due punti, occorre misurare anche il valore della risposta fra i due punti, diversamente l'unica cosa visibile sarebbe per forza solo una retta.

Stabilito che esiste una curvatura fra 5 e 7, con i 7 punti ottenuti fin'ora è possibile infine creare un modello quadratico del sistema.

Il risultato è che abbiamo esplorato una curva sconosciuta, seguendo un metodo ben preciso e abbiamo ottenuto un modello del processo che ci ha permesso di individuare le condizioni per ottimizzare il processo stesso.

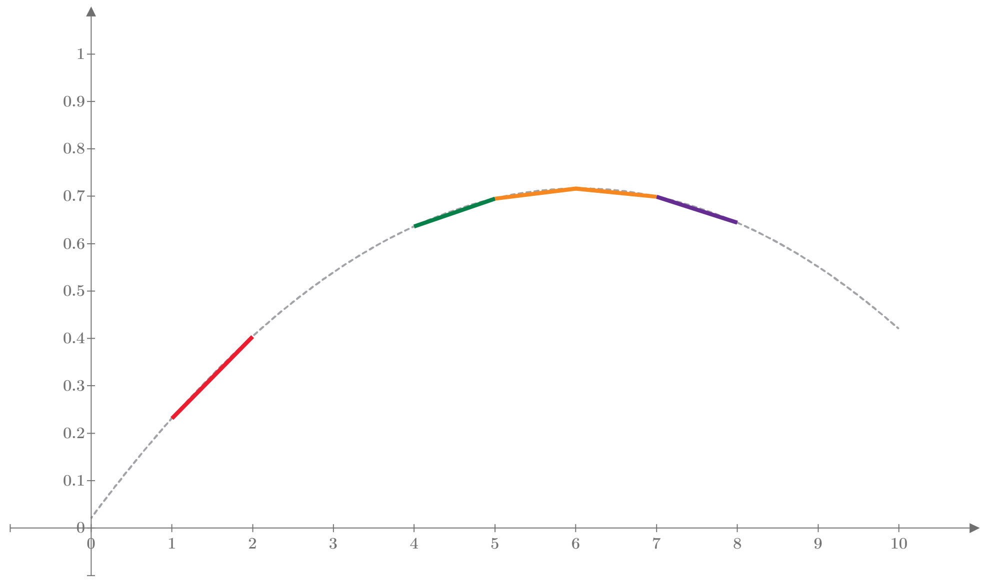
Luigi Roggia
data scientist @ Kiwi Data Science www.kiwidatascience.com
partner di Gmsl s.r.l. www.gmsl.it

$$y(t) := -0.019 \cdot t^2 + 0.23 \cdot t + 0.02$$

Questa è l'equazione di regressione ottenuta tramite l'approccio del DoE.

Il seguente grafico mostra i quattro esperimenti che abbiamo eseguito per arrivare infine a creare un modello del processo.

Sono effettivamente quattro esplorazioni, guidate l'una dalla precedente, svolte per esplorare una curva ignota al fine di crearne una "mappa", ovvero in termini matematici un'equazione che rappresenti il nostro modello.



Luigi Roggia
data scientist @ Kiwi Data Science www.kiwidatascience.com
partner di Gmsl s.r.l. www.gmsl.it

In rosso il primo esperimento, in verde il secondo, in viola il terzo e in arancione il quarto ed utlimo. La linea grigia tratteggiata rappresenta il modello finale. Con una sola variabile, o meglio un solo fattore, il discorso è molto semplice.

Ma tornando alla realtà, i fattori sono sempre più numerosi e senza il metodo del DoE dovremmo testare a caso un numero molto grande di combinazioni di valori, senza un metodo razionale e *statisticamente valido* che ci guidi.

Il bello della matematica è che ci permette di svolgere le stesse analisi in qualsiasi dimensione! Pertanto il DoE permette di analizzare anche per esempio superfici 3D nel caso in cui si abbiano una risposta e due fattori, così come ipersuperfici, ovvero superfici in dimensioni superiori a 3, che si hanno ogniqualvolta che i fattori siano più di due.

Chiaramente la rappresentazione grafica è possibile solo fino 3 dimensioni, ma nulla ci vieta di esplorare le ipersuperfici e trarre conclusioni ugualmete sperimentalmente valide e matematicamente corrette!

Luigi Roggia
data scientist @ Kiwi Data Science www.kiwidatascience.com
partner di Gmsl s.r.l. www.gmsl.it

Un esempio reale con due fattori

Questo esempio, basato su dati reali, è tratto dal corso ufficiale di Minitab riguardo il Response Surface Design, che è uno dei metodi di DoE per studiare la curvatura (ovvero per studiare e modellare problemi di ottimizzazione).

In particolare i progetti Minitab di riferimento sono Yield1, Yield2 e Yield3.

Il problema è il seguente:

si ha un processo chimico industriale di cui si vuole incrementare la resa, sempre espressa come percentuale. Dopo una fase di screening (anch'essa eseguita con il DoE) si è arrivati a determinare che fra i vari possibili fattori che influenzano la resa (yield), i due importanti sono il tempo (espresso in minuti) e la temperatura (espressa in °C).

L'azienda sta attualmente lavorando a 130 °C per 75 minuti.

Per cercare di massimizzare la resa occorre individuare un massimo della funzione che descrive il processo; tale funzione è ovviamente ignota. Quindi col DoE si procede esplorando una piccola zona alla volta.

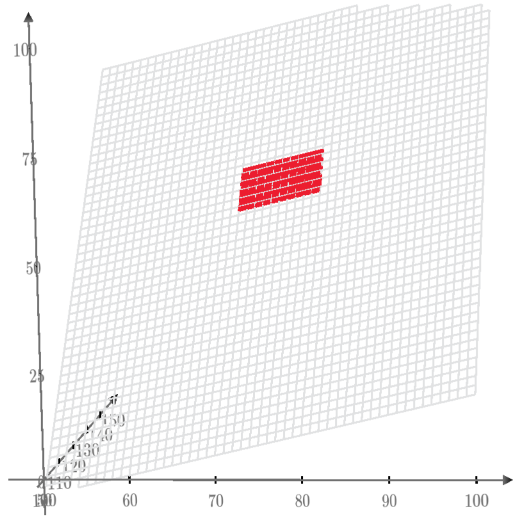
Per iniziare si centra l'esperimento attorno alle condizioni attuali, in modo da capire in che direzione proseguire per trovare un massimo (ovvero, da che parte si sale?)

Per centrarsi attorno alle condizioni attuali si scelgono i livelli 70 e 80 per il tempo e 127.5 e 132.5 per la temperatura. Si eseguono misurazioni anche nel punto centrale (75 min, 130 °C) per capire se questa porzione di superficie presenta curvatura.

I risultati evidenziano che la curvatura non è rilevante in questa porzione della superficie e tramite l'analisi dei dati raccolti nell'esperimento, otteniamo un'equazione di regressione che fornisce la resa y in funzione di tempo (t) e temperatura (T):

$$y(t,T) := -207.2 + 0.470 \cdot t + 1.8 \cdot T$$

Luigi Roggia
data scientist @ Kiwi Data Science www.kiwidatascience.com
partner di Gmsl s.r.l. www.gmsl.it



La porzione in rosso è lo spazio sperimentale che abbiamo esplorato con questo primo DoE.

In grigio chiaro è rappresentata la funzione individuata con il metodo di regressione.

La direzione in cui procedere è quella in cui sia tempo che temperatura hanno valori crescenti; si vede dall'inclinazione del piano.

In realtà la vera direzione viene stimata con algoritmi specifici (steepest ascent/descent), calcolati con Minitab o R

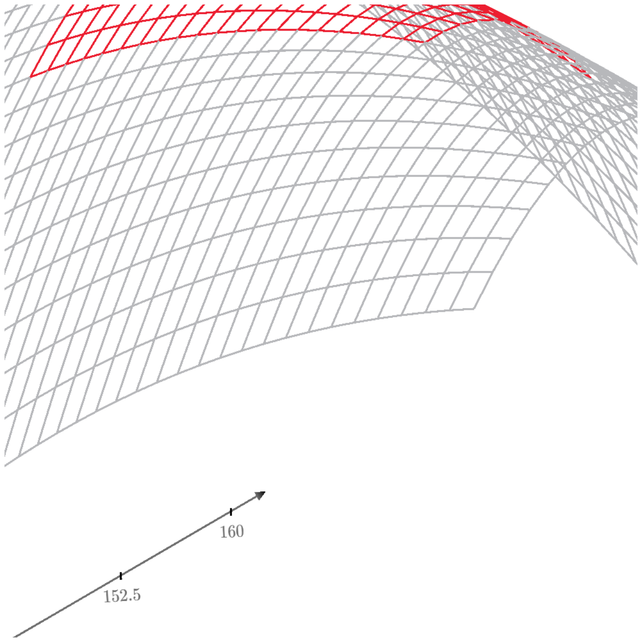
Grazie a questo primo risultato si decide di fare un secondo esperimento centrato attorno a 90 minuti e 145 °C, scegliendo come livelli 80 e 100 per il tempo e 140 e 150 per la temperatura.

Ovviamente si eseguono misure anche nel punto centrale (90,145) per valutare se la curvatura sia importante in questa nuova porzione della superficie (o in questo nuovo spazio sperimentale).

Svolgendo l'analisi con Minitab si trova che la curvatura in questo secondo spazio sperimentale è presente in modo importante.

L'equazione di regressione risultante dall'analisi contiene infatti termini quadratici:

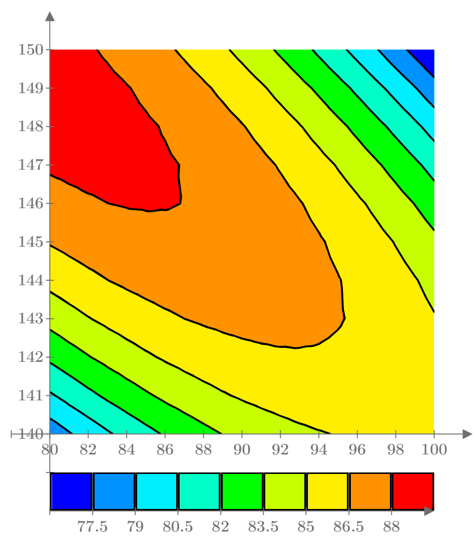
$$y(t,T) := -4139 + 18.21 \cdot t + 47.01 \cdot T - 0.0234 \cdot t^2 - 0.1316 \cdot T^2 - 0.0975 \cdot t \cdot T$$



In rosso il nuovo spazio sperimentale, in grigio la superficie descritta dal modello che abbiamo ricavato.

Questo secondo esperimento ha indubbiamente condotto all'indiduazione di una situazione di ottimo.

Il contour plot qui sotto rappresenta lo spazio sperimentale impiegato.



Il contour plot rappresenta molto bene i valori dello spazio sperimentale rappresentato in rosso sulla superficie 3D.

L'insieme dei due grafici aiuta a capire l'andamento della superficie della funzione.

Il contour plot è il grafico tecnico a cui affidarsi