GP5 - Analisi dei dati e previsione del coefficiente di perdita

Riccardo Cervero[1], Marco Savino[2], Luca Lazzati[3]

Sommario
L'azienda Bosch, in collaborazione con l'Università degli Studi Di Milano Bicocca, ha messo a disposizione i dati relativi al test della linea di assemblaggio di un particolare tipo di pompe, denominate GP5. Il progetto, dopo un'introduzione tecnica del caso di studio, intende innanzitutto offrire quattro livelli di analisi riguardanti il processo produttivo. Il primo consiste in un'indagine preliminare delle grandezze osservate durante la creazione del pezzo meccanico - categoriche e numeriche -. Verranno mostrate, in particolare, le forti relazioni fra i descrittori relativi alle variabili di pressione e portata. Con lo scopo di estendere lo studio di tali relazioni, verrà dunque presentata un'analisi di correlazione fra tutte le colonne rilevanti presenti all'interno del *database*. Il terzo livello ha implicato un'approfondimento dei complessi legami di correlazione, attraverso la stima di vari modelli di classica regressione lineare, l'utilizzo di tecniche di *shrinkage* - quali *Ridge* e *Lasso* - e modelli misti di tipologia *multilevel*. Infine, l'ultima indagine riguarda i valori anomali, sia univariati che multivariati. Inoltre, il progetto offre una soluzione grafica *real-time* per il monitoraggio delle grandezze del processo produttivo a bassissima latenza e l'individuazione di eventuali *outliers*, e presenta un algoritmo per rilevare eventuali valori scorretti della portata in base al tipo di pompa e alle misurazioni del coefficiente di dispersione e della pressione, oltre a rappresentare la distanza fra le osservazioni e questo limite dinamico nello spazio tridimensionale.

Keywords
Monitoraggio – Previsione

[1] *794126, Dipartimento di Informatica, Sistemistica e Comunicazione*
[2] *793516, Dipartimento di Informatica, Sistemistica e Comunicazione*
[3] *850334, Dipartimento di Informatica, Sistemistica e Comunicazione*

Indice

1 Caso di studio **1**
2 Data Preparation **2**
3 Analisi preliminari **3**
3.1 Variabili categoriche 3
3.2 Variabili numeriche 3
Descrittori della pressione • Descrittori della portata • Altre variabili
3.3 Target: coefficiente di dispersione 5
3.4 *Outliers Detection* 7
4 Sistema di monitoraggio *real-time* **8**
4.1 Demo 9
5 Limite dinamico di portata **9**
6 Modelli di previsione **9**
6.1 Modelli OLS 9
Modello OLS con variabili di portata • Modelli di previsione con variabili di pressione
6.2 Tecniche di *shrinkage* 13
Regressione *Ridge* • Regressione *Lasso*
6.3 Regressione Multilevel 14
Modelli multilivello con variabili di portata • Modelli di previsione con variabili di pressione
6.4 Conlclusioni sui modelli predittivi 16
Riferimenti bibliografici **17**

1. Caso di studio

L'azienda Bosch, in collaborazione con l'Università degli Studi di Milano Bicocca, ha messo a disposizione i dati relativi al test della linea di assemblaggio di un particolare tipo di pompe, denominate GP5. In particolare, le misurazioni sono state raccolte nell'ambito di un circuito pneumatico di prova, che ha implicato due precise fasi: un test a regime, in cui pressione e portata vengono rilevate a seguito dell'impostazione di una velocità di rotazione della pompa pari a 2300 rotazioni per minuto (rpm), e un test in fase di controllo, legato ad una velocità minore, precisamente pari a 140 rotazioni per minuto. La produzione di ogni pezzo dipende da alcune principali grandezze:

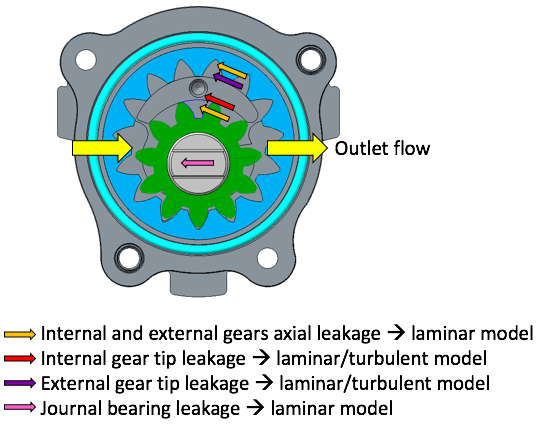
- Q_{GP}, ovvero la portata attuale del prodotto misurata tramite apposito trasduttore in litri all'ora [L/h]

- $Q_{GP,theor}$, la portata teorica del prodotto, nota a priori (108.36 [L/h]) e fornita dall'ufficio *R&D*
- P_{out} è la pressione in uscita misurata tramite apposito trasduttore [bar][1].
- α, ovvero il coefficiente di perdita (*leakage coefficient*), misurato in [(l/h)/bar], che riassume la prestazione delle pompe GP5: essendo direttamente proporzionale agli scostamenti meccanici interni, descrive la sensibilità del prodotto alle variazioni di pressione in uscita; se questo coefficiente di flusso è basso, la portata della pompa è meno influenzata dalla *outlet pression*, e il pezzo è più robusto e affidabile.

Nel dettaglio, gli scostamenti interni causati dalla portata in uscita sono classificati in

- distanza assiale, pari all'altezza della custodia e degli ingranaggi che costituiscono la pompa
- distanze di punta, che comprendono il diametro di punta degli ingranaggi e il diametro esterno ed interno della mezzaluna; quest'ultimo, in particolare, si identifica come la distanza della componente *half moon* dalla componente rotante della pompa
- spostamento degli ingranaggi e dei cuscinetti interni rispetto alla custodia.

Riassumendo, durante il test, il principale ingranaggio interno ruota eccentricamente all'interno della custodia della pompa, generando una bassa pressione in uscita. Maggiore la reazione a questo flusso uscente, in termini di spostamento delle proprie componenti interne, minore l'indice di prestazione del prodotto, indicato con α. La direzione e l'entità di ogni scostamento può essere intuita dalla seguente Figura:



La freccia rosa al centro ("*journal bearing leakage*") indica un movimento del più piccolo ingranaggio centrale rispetto alla posizione originaria. Le frecce rossa e viola disegnano rispettivamente lo sfalsamento interno ed esterno della componente superiore definita "a mezza luna" (nell'ambito delle cosiddette distanze di punta). Infine, le due frecce arancioni rappresentano una variazione delle distanze assiali interna ed esterna.
La produzione GP5 può essere dunque distinta in "classi di coefficiente di perdita", ossia gruppi di pompe che condividono lo stesso *leakage coefficient* α, che viene messo in relazione con la pressione e la portata con la seguente formula[2] verificata sperimentalmente:

$$Q_{GP} = Q_{GP,theor}\alpha \cdot P_{out} \tag{1}$$

L'andamento del coefficiente di perdita, poiché fondamentale per il monitoraggio del prodotto, verrà analizzato non soltanto al variare del programma selezionato o della fase impostata, ma soprattutto in relazione alle altre misurazioni estratte durante il circuito pneumatico di prova. Nel dettaglio, a ciascuna delle due fasi di test sono associate interessanti osservazioni numeriche relative a picco e media della *outlet pression*, picco e media della portata in uscita, picco e media della coppia in fase finale - ovvero in corrispondenza di una velocità di 100 rotazioni al minuto -, e temperatura di prova del liquido con cui viene lubrificata la pompa per la rotazione.
Altre variabili significative sono presenti nel *database* fornito, tra cui il codice dei turni in cui è suddivisa la linea di lavoro, ora e data del *record*, descrizione dell'esito finale della prova, e il *Programma*, corrispondente alla classe di prodotto GP5, che è raggruppabile nelle due macro-categorie *Daimler (DAI* e Standard (STD)[3].

A partire da tale caso di studio, sono stati proposti tre *task*:

1. analisi preliminare delle grandezze del processo produttivo e implementazione di un sistema di monitoraggio *real-time* del coefficiente di perdita (Sezioni 3 e 4)
2. definizione, dato un intervallo del coefficiente di perdita, di un limite dinamico della portata Q_{GP} rispetto alla pressione in uscita (Sezione 5)
3. stima di un modello efficace per la previsione del coefficiente di perdita α (Sezione 6).

2. Data Preparation

Il *database* si compone di 296605 osservazioni, descritte da 33 colonne. Tuttavia, è stato necessario focalizzare l'analisi su quelle che presentavano un certo grado di rilevanza e utilità per l'oggetto di studio. Ciò ha comportato la rimozione, durante una prima fase di *pre-processing*, delle variabili che

[1] Registrara nel database come `media_pressione_velocita_1`.

[2] Q_{GP} e P_{out} sono note rispettivamente col nome di `media_portata_velocita_1`, e `media_pressione_velocita_1`

[3] La macro-categoria *Daimler* è descritta dal codice programma `18_GP5_910_CW`, mentre la tipologia Standard comprende tutti gli altri codici programma

presentavano le seguenti caratteristiche: si manifestavano come un unico valore costante[4], mostravano una distribuzione eccessivamente sbilanciata verso una sola classe[5], erano ridondanti, poco interessanti o erano già state segnalate tali dai fornitori del *dataset*[6].
Successivamente, è stata rimossa una riga, poiché penalizzata dalla mancanza dei valori di pressione. Infine, poiché la differenza di scala era spesso eccessiva per molte colonne numeriche, si è deciso di normalizzare la matrice originale, mappando tutte le variabili non categoriche fra 0 e 1.

3. Analisi preliminari

Dopo le operazioni di pulizia appena menzionate, si è proceduto ad analizzare in maniera preliminare le colonne rilevanti, per ottenere un primo approfondimento sui descrittori e i rapporti di dipendenza fra gli stessi.

3.1 Variabili categoriche

I dati nominali conservati sono relativi a due precisi aspetti del processo: la segmentazione del lavoro in turni differenti e il programma selezionato per la creazione del pezzo, ovvero la tiplogia di pompa.
Per quanto riguarda la prima, è stato necessario aggregare le medesime classi registrate erroneamente in maniera diversa, uniformandone la denominazione. In questo modo, sono state ottenute 5 classi: ”A”, ”B”, ”C”, ”D”, ”0”. Le ultime due, poiché osservate in minor misura e non specificate inizialmente dai fornitori del *dataset*, sono state etichettate come *missing values*. Come osservabile in Figura 1, i principali blocchi in

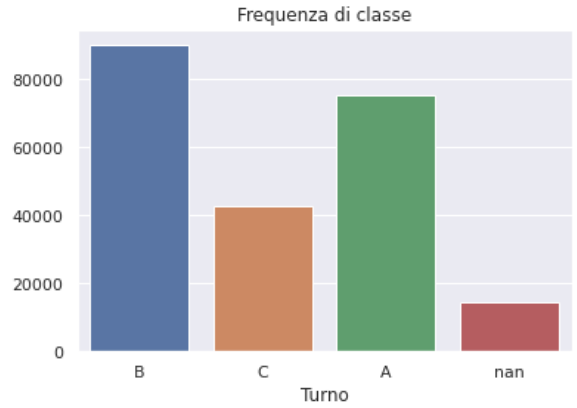


Figura 1. *Bar chart* per la visualizzazione delle rispettive frequenze osservate dei turni.

cui è organizzato il processo presentano una frequenza diversa: il turno ”B” costituisce più del 40% dei *records*, ”A” si presenta nel ∼34% dei casi e ”C” nel ∼19%.
La composizione del ”Programma” mostra un fortissimo sbilanciamento verso la classe GP5 denominata `18_GP5_910_CW` (Figura 2), che corrisponde alla pompa ”*Daimler*”. Le restanti tipologie, raggruppate sotto denominazione ”Standard”, formano in totale meno del 23% dei casi e sono principalmente rappresentate dalle categorie `17_GP5_430_CCW` (10.6%) e `12_GP5_430B_D1` (8.5%). Per una visualizzazione migliore,

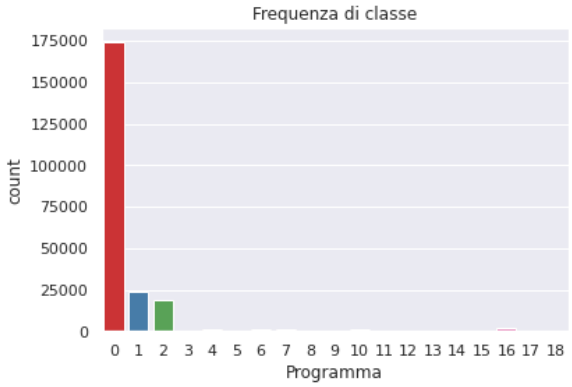


Figura 2. *Bar chart* per la visualizzazione delle rispettive frequenze osservate dei programmi.

in Figura 3 è riassunta la distribuzione delle frequenze logaritmiche delle varie classi GP5.

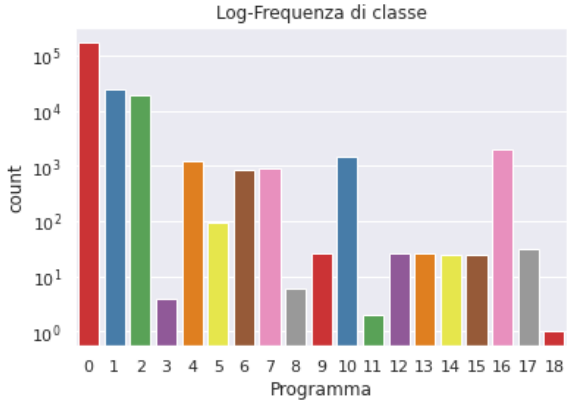


Figura 3. *Bar chart* per la visualizzazione delle rispettive frequenze logaritmiche dei programmi.

Tra le variabili categoriche esaminate, ”Esito” descrive le eventuali - rarissime - imperfezioni del prodotto. A tal proposito, il 99.5% delle volte il risultato è ottimale, mentre l’anomalia più comune è indicata come ”scarto picco coppia max fase pulizia iniziale” (0.1% delle osservazioni).

3.2 Variabili numeriche

I dati numerici registrano, per ogni pompa, due principali set di valori, relativi al picco e alla media di pressione (in bar) e portata (in litri all’ora), misurati in corrispondenza di due velocità diverse:

[4] Le costanti rimosse sono: `"Banco"`, `"Master"`, `"Picco coppia zero"`, `"Picco coppia iniziale"`, `"Media coppia iniziale"`, `"Velocità 1"`, `"Picco pressione velocità 2"`, `"Media pressione velocita 2"`, `"Picco portata velocità 2"`, `"Media portata velocità 2"`.

[5] È questo il caso di `"Velocità 2"`, `"Esito"` e di conseguenza `"N. Esito"`, `"Coppia max ciclo"`, `"Velocità a regime"`.

[6] Le colonne inutili sono: `"Codice da Linea"` - identificativo del pezzo -, `"Media coppia zero"`, `'Data'` e `'Ora'`, poiché i dati sono stati raccolti in un arco temporale non uniforme.

- a regime: 2300 *r.p.m*
- 140 *r.p.m*

Si vedrà che, a prescindere dalla velocità, le due grandezze condividono un'elevatissima dipendenza lineare.

3.2.1 Descrittori della pressione

Innanzitutto, le grandezze relative al picco e alla media hanno una distribuzione multimodale nell'ambito di entrambe le misurazioni della velocità (Figura 4), rivelando picchi di diversa altezza. Il range osservato è molto superiore per quanto concerne la velocità a regime (come riassunto dalla Tabella 1). Inoltre, le distribuzioni dei valori appaiono quasi identiche per classe di velocità, rivelando una fortissima similarità fra l'andamento del picco e e della pressione media (Figura 5).

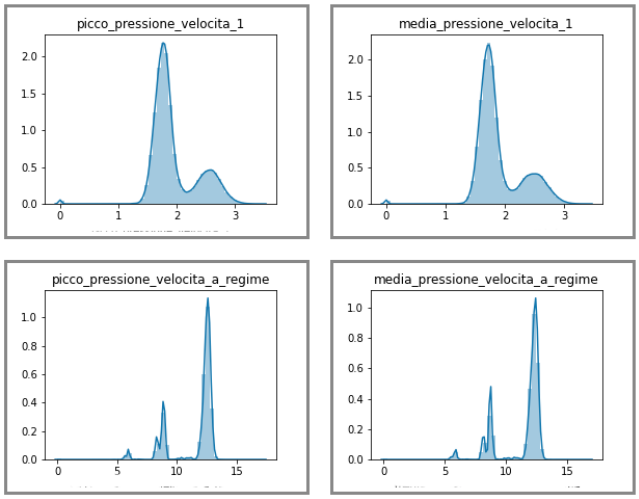


Figura 4. Distribuzioni delle grandezze relative alla pressione.

Pressione	Minimo	Media	Massimo
Picco (140)	0.19	1.94	3.45
Picco (2300)	5.38	11.61	14.21
Media (140)	0.1	1.88	3.38
Media (2300)	5.33	11.41	14.01

Tabella 1. Rassunto relativo alla media e al range dei valori di pressione.

3.2.2 Descrittori della portata

Le rispettive distribuzioni di portata (Figura 6) rivelano andamenti molto meno regolari. In particolare, i *records* a regime sono caratterizzati da una coda molto estesa verso destra, lasciando presupporre la presenza di una quota rilevante di valori anomali e permettendo di intravedere una differenza ancor più ampia rispetto a quanto visto con una velocità di 140 *r.p.m.* I comportamenti mostrati dalle osservazioni relative alla pressione della pompa si ritrovano in forma identica

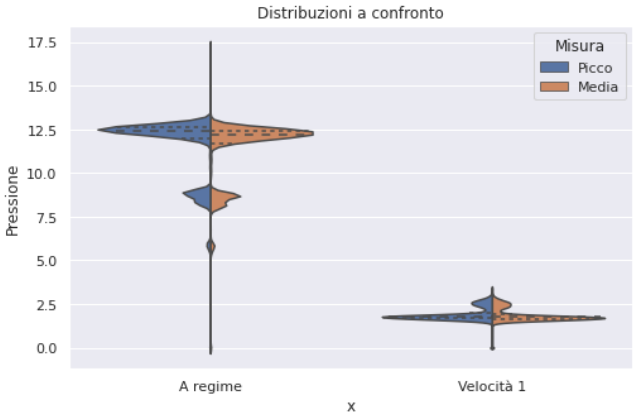


Figura 5. Distribuzioni delle grandezze relative alla pressione, in corrispondenza della velocità "1" e a regime, appaiate nell'ambito della stessa misurazione. È possibile notare sia la superiorità della pressione a regime, che la fortissima vicinanza dell'andamento del picco rispetto alla pressione media.

Portata	Minimo	Media	Massimo
Picco (140)	30.30	58.96	93.6
Picco (2300)	930.5	1309.6	1422.9
Media (140)	28.07	58.5	90
Media (2300)	927.63	1304.8	1416.2

Tabella 2. Rassunto relativo alla media e al range dei valori di portata.

esaminando le misurazioni di portata: anche qui, esiste una grandissima somiglianza fra le distribuzioni di picco e media, e - come anticipato - una differenza di scala ancor più ampia fra i valori estratti per le due velocità (Figura 7).
Infine, è interessante notare come le colonne relative alla pressione e alla portata condividano dipendenze lineari con grado molto elevato, come visibile in Figura 8: la correlazione assoluta media raggiunge addirittura un livello di $\sim$0.95, con un minimo di 0.79 - tra i picchi di portata misurati fra le due velocità - e 8 collinearità perfette. In particolare, oltre alla quasi identicità delle distribuzioni nell'ambito della stessa grandezza - menzionata in precedenza -, appaiono interessanti le seguenti dipendenze lineari:

- picco pressione e picco portata, entrambi alla velocità di 140 *r.p.m.* ($\sim$0.91)
- media pressione e picco portata, sempre per quanto riguarda la fase di controllo ($\sim$0.91)
- media portata e picco pressione, a 140 *r.p.m.* ($\sim$0.92)
- media portata e media pressione, per la fase di controllo ($\sim$0.92)
- media pressione e media portata a velocità di regime, fra cui s'individua una correlazione perfetta

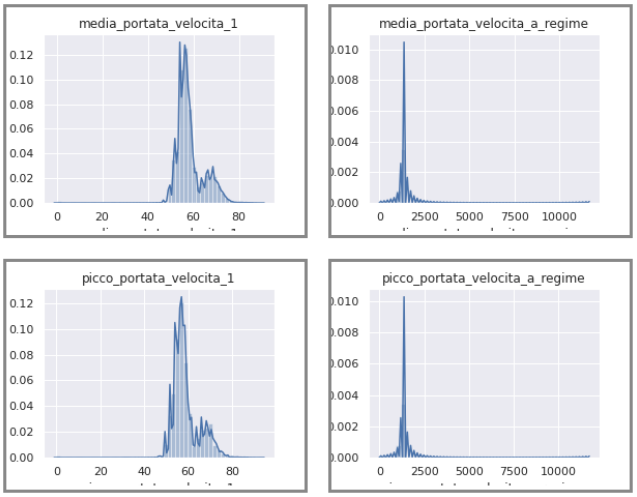


Figura 6. Distribuzioni delle grandezze relative alla portata.

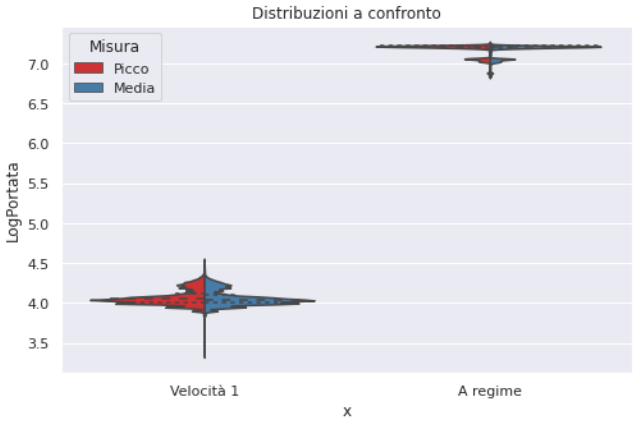


Figura 7. Distribuzioni delle grandezze relative alla portata, in corrispondenza della velocità "1" e a regime, appaiate nell'ambito della stessa misurazione. Poichè la differenza era troppo grande per permettere una visualizzazione comprensibile, è stato necessario adottare una scala logaritmica.

- picco pressione e picco portata a velocità di regime, sempre una correlazione perfetta
- picco portata e media pressione a velocità di regime: perfetta collinearità.

Appare quindi evidente la sistematica dipendenza fra pressione e portata, che, osservando i dati forniti, paiono spesso essere state generate dalla stessa distribuzione casuale, a prescindere dalla misura scelta e dalla fase impostata. Per questa ragione, in fase di generazione dei modelli di regressione, la matrice del disegno potrebbe essere affetta da un eccessivo problema di multicollinearità, rischiando di divenire quasi-singolare o addirittura non invertibile, penalizzando una stima corretta sia dei coefficienti di regressione che degli indici di bontà di adattamento della funzione di regressione. Sarà pertanto necessario rimuovere tali variabili più correlate, oppure adottare strategie utili alla selezione di un modello a più bassa dimensionalità, come, ad esempio, le tipologie *Ridge* e *Lasso*.

3.2.3 Altre variabili

Le altre colonne numeriche, filtrate dalla fase di *pre-processing* e ritenute rilevanti ai fini dello studio, sono: media coppia in fase finale (100 *rpm*), picco coppia in fase finale e temperatura di prova del liquido.
Per quanto riguarda la prima, la distribuzione è fortemente asimmetrica verso valori elevati, con una discreta porzione di *outliers*. Il picco della stessa grandezza presenta una coda ancor più lunga (in scala logaritmica in Figura 9). Medesima condizione di elevata asimmetria vale per la temperatura, compresa fra un range di 37.61 e 49.1. È importante citare che anche fra la media e il picco della coppia in fase finale (100 *rpm*) esista una forte dipendenza lineare positiva ($\sim$0.83). Bisognerà pertanto valutare questa ulteriore collinearità in fase di stima del modello per la previsione del coefficiente.
Al contrario, queste due variabili non presentano un'elevata dipendenza rispetto i dati di pressione e portata: la correlazione rimane sempre ben al di sotto del 20%. Lo stesso vale per i valori di temperatura.

3.3 Target: coefficiente di dispersione

Il coefficiente di dispersione è calcolato con la formula

$$\alpha = \frac{108.36 - \text{Media portata (@140)}}{\text{Media pressione (@140)}} \tag{2}$$

e i valori sono compresi in un range molto ampio, fra un minimo di $\sim$5.7 e un massimo di $\sim$817, con media e mediana rispettivamente pari a $\sim$27.8 e $\sim$29.43. La distribuzione è sbilanciata verso alti valori, come visibile in Figura 10. In Figura, è possibile notare una certa concentrazione dei *records* in un intervallo di ampiezza ridotta attorno alla mediana. Questa considerazione è appurata dal basso coefficiente di variazione: $\sim$0.26.
Poiché il coefficiente di *leakage* è indice della performance della pompa prodotta, è ragionevole supporre che ad ogni classe GP5, come conseguenza della differenza fra le proprie caratteristiche di progettazione e quelle delle altre categorie, si associ un diverso livello di prestazione - quindi un diverso *range* di α. Da un'analisi grafica superficiale (Figura 11), questa assunzione appare corretta: le distribuzioni del coefficiente, raggruppate per "Programma", appaiono spesso distanti, facendo ipotizzare l'esistenza di una notevole variabilità fra i gruppi, potenzialmente interessante per la costruzione di un modello predittivo. Inoltre, è possibile che tale varianza fra le classi GP5 si manifesti non solo fra i livelli di performance, ma anche fra i coefficienti di regressione delle variabili esplicative scelte. In altre parole, è probabile che la selezione di un determinato Programma comporti una relazione significativamente diversa - rispetto agli altri tipi di pompe - fra α e le altre grandezze. Se questa ipotesi venisse verificata, allora

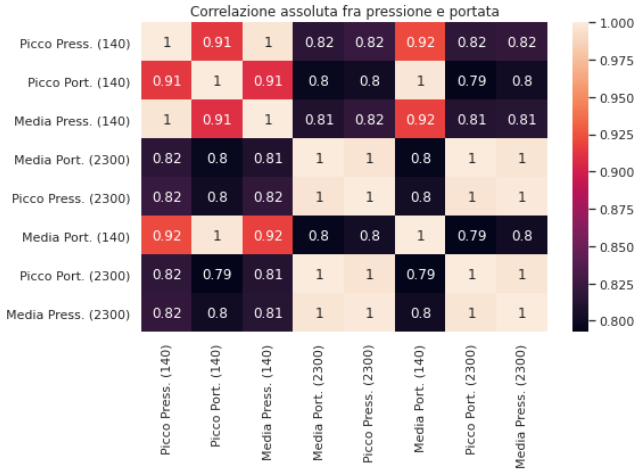


Figura 8. Matrice di correlazione, con valori assoluti, fra tutte le variabili di pressione e portata.

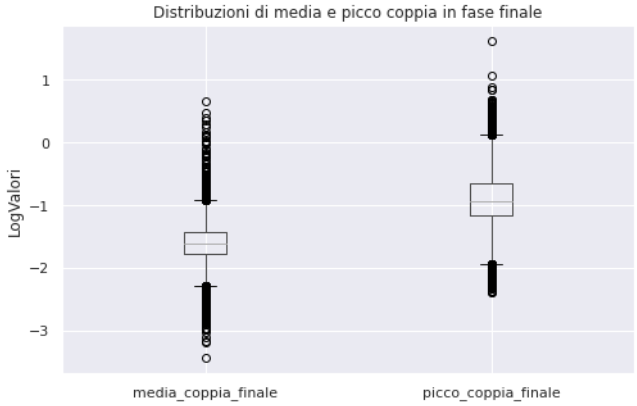


Figura 9. *Boxplot* per la visualizzazione della distribuzione delle variabili ”Media coppia finale” e ”Picco coppia finale”.

si potrebbe dimostrare quanto il contributo delle varie parti coinvolte nel processo - monitorate attraverso la misurazione dei valori registrati nelle colonne nel database - si modifichi, influenzato dagli specifici fattori contestuali relativi alla diversa progettazione di ogni classe di pompa. In questo modo, sarebbe anche possibile ottimizzare il valore del coefficiente α, monitorando questi fattori in maniera differente a seconda del Programma. Pertanto, con lo scopo di analizzare in maniera più approfondita la variabilità della funzione di regressione lineare di α in base al tipo di pompa in produzione, si è scelto di stimare una tipologia di modello misto adatta ai dati influenzati da effetti contestuali: il cosiddetto *multilevel model*. Tale approfondimento è trattato nella Sezione 6.
Oltre alla relazione con ”Programma”, il coefficiente di dispersione mostra un collinearità molto elevata con molte altre colonne. Tralasciando le distribuzioni da cui α è stato generato (equazione 4), le variabili con cui è più fortemente correlato sono relative alla pressione a regime (”Picco pressione velocita a regime” al 75% e ”Media pressione velocita a regime”

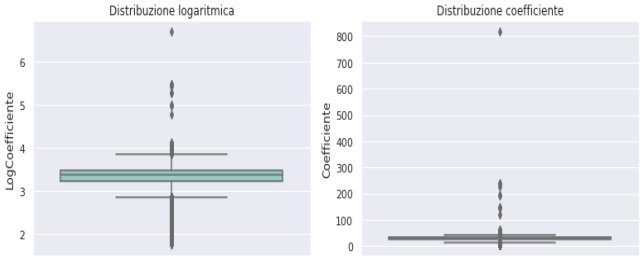


Figura 10. Distribuzione del coefficiente di *leakage* in scala logaritmica e originale.

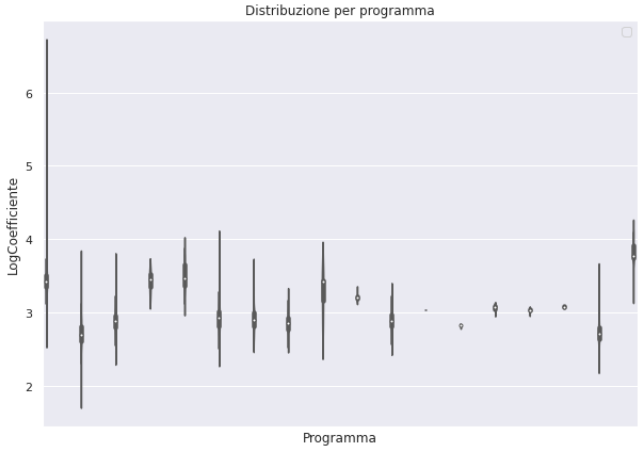


Figura 11. Distribuzioni del coefficiente α raggruppate per classe GP5 (”Programma”).

al 74.6%) e al picco di portata durante la stessa fase (74%). Normalmente, queste dovrebbero essere quindi incluse nel *subset* dei candidati esplicativi per la regressione lineare di α, poiché potrebbero contribuire ad una stima accurata. Tuttavia, come si vedrà nella Sezione 6, la *feature selection* sarà un'operazione difficile, a causa della multicollinearità rilevata fra le misurazioni di portata e pressione - in entrambe le fasi - e del conseguente rischio di ottenere una matrice del disegno quasi-singolare o invertibile. Ad esempio, se si volessero includere entrambe le variabili di pressione in fase di regime, la penalizzazione alla correttezza del modello non deriverebbe soltanto dalla loro reciproca collinearità, bensì anche dalla loro dipendenza lineare con le colonne utilizzate per calcolare α.
La correlazione mostrata da α rispetto a ”Media coppia finale” appare moderata (16.9%), e ”Picco coppia finale” ha una dipendenza lineare pari a circa il 7%. È necessario ricordare che anche fra queste due grandezze esiste una collinearità rischiosa, la quale, nel caso in cui esse venissero incluse assieme nel modello, finirebbe per distorcere i coefficienti di regressione. Infine, la correlazione rispetto alla temperatura è ridotta: -6.9%.
In seguito, poiché le osservazioni erano corredate da dati temporali relativi alla data - giorno, mese e anno - e all'orario,

si è proceduto a verificare la possibilità di estrarre eventuali trend temporali. Tuttavia, le misurazioni appaiono cronologicamente disomogenee su due livelli. Innanzitutto, i *record* nono sono stati forniti con coerenza temporale per quanto riguarda l'anno (Figura 12). In ogni caso, non è comunque

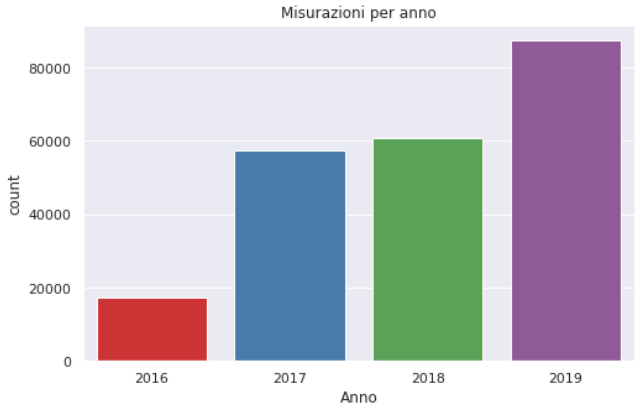


Figura 12. Frequenza delle misurazioni per anno. I dati non sono stati forniti in maniera omogenea.

visibile alcuna differenza significativa fra le distribuzioni della performance nei vari anni (Figura 13).

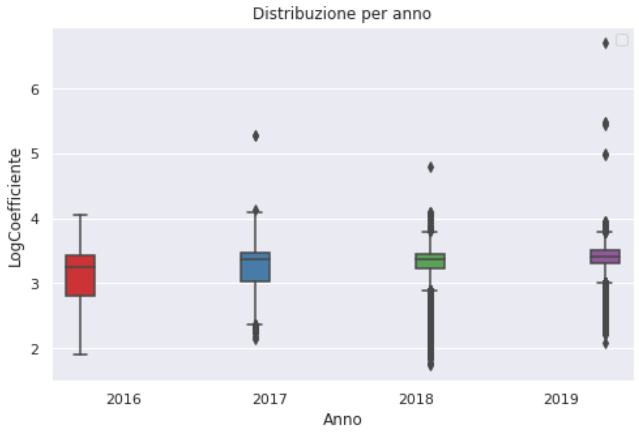


Figura 13. Distribuzioni del coefficiente α raggruppate per anno.

Neppure per quanto riguarda il mese, le registrazioni sono omogenee, come visibile in Figura 14.
Eppure, nonostante dalla collezione dei dati possa sembrare impossibile estrarre una certa periodicità per i motivi appena spiegati, l'andamento del coefficiente di dispersione è inaspettatamente armonico (Figura 15). Tuttavia, è sufficiente eseguire un clustering grafico dei punti in base al programma di appartenenza per dedurre la motivazione di tale apparente periodicità: essa non deriva da un'autocorrelazione temporale fra i valori del coefficiente, bensì dall'ordine con cui è stato scelto di effettuare le misurazioni sui diversi gruppi GP5, che, come si è visto in precedenza, hanno distribuzioni differenti.

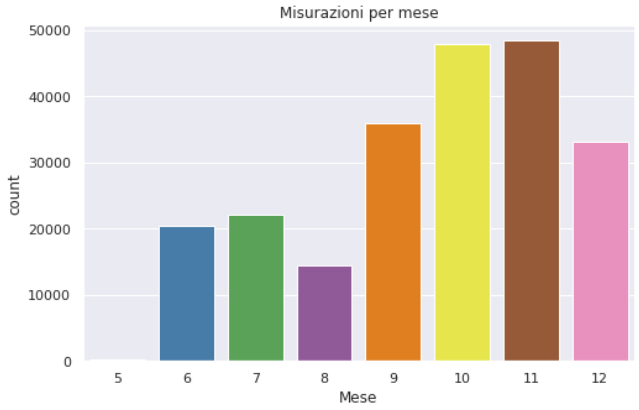


Figura 14. Frequenza delle misurazioni per mese. I dati non solo non sono stati forniti in maniera omogenea, ma i primi 4 mesi sono addirittura mancanti, e maggio non riporta quasi osservazioni.

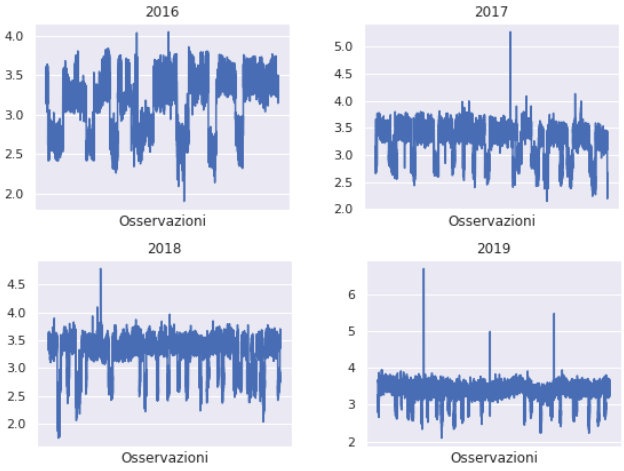


Figura 15. Andamento del coefficiente di dispersione α per anno.

L'andamento, perciò, appare generato da una sinusoide solamente perché le osservazioni alternano classi di pompa con performance elevate a categorie con un α minore. Quanto detto è evidente in Figura 16.

3.4 *Outliers Detection*

L'ultima fase di analisi preliminare consiste nell'identificazione degli *outliers*. Grazie ad essa, è possibile, in una fase di ottimizzazione della produzione, ricostruire il trend del processo, individuare le anomalie non segnalate dalla variabile "Esito" ed inferire eventuali fattori in grado di causarle. Innanzitutto, si è proceduto ad estrarre gli *outliers* univariati all'interno della distribuzione del coefficiente di dispersione α. Per farlo, poichè è stato verificato che tale distribuzione non fosse di tipo Normale, è stato applicato il metodo non parametrico della distanza interquartile, segnalando pertanto i punti inferiori a k volte il primo quartile e superiori a k volte

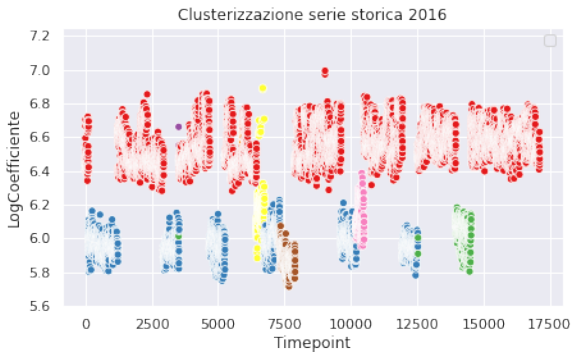


Figura 16. Andamento del coefficiente di dispersione α durante l'anno 2016, clusterizzato in base alla variabile "Programma". Il colore rosso indica la tiplogia di pompa *Daimler* - la più popolosa -, mentre gli altri colori sono associati alle sottocategorie Standard.

il terzo quartile, con $k = 1.5$. È stato appurato che queste anomalie non dipendono da un particolare programma[7], né si verificano in particolari anni o mesi.

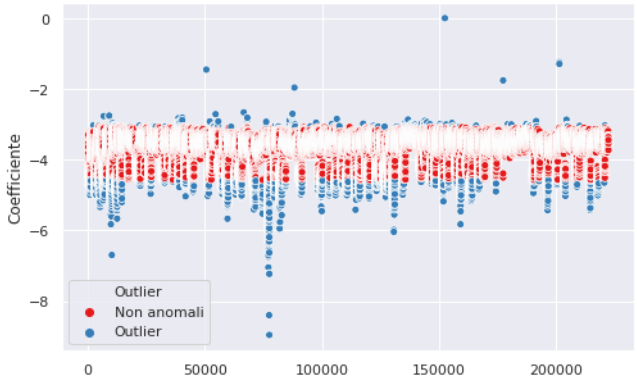


Figura 17. Identificazione degli outlier univariati all'interno della distribuzione del coefficiente di dispersione.

Si è successivamente proceduto ad estrarre *outlier* multivariati grazie all'algoritmo *Local Outlier Factor* (LOF), basato sul calcolo delle deviazioni locali di ogni punto dal proprio vicinato. In questo caso, i valori anomali individuati rappresentano un porzione molto grande del database (75%) e prescindono, come in precedenza, da particolari condizioni o programmi. Questa elevata quantità può dipendere dall'incidenza che gli specifici fattori contestuali di progettazione hanno sulle varie distribuzioni, elevando il grado di variabilità nell'andamento globale, così come dall'alternanza di fasi diverse nel processo. Sarebbe pertanto più utile applicare l'algoritmo di *anomaly detection* separatamente in corrispondenza della cosiddetta velocità 1 (140rpm) e della fase di regime, e su ciascuna classe GP5.

[7] Le anomalie sono presenti sia in corrispondenza della categoria *Daimler* sia Standard.

4. Sistema di monitoraggio *real-time*

Il processo industriale può essere ottimizzato grazie all'implementazione di sistemi di monitoraggio che individuino in tempo reale le eventuali anomalie, registrino i dati e li rappresentino graficamente su schermo per ottenere una visualizzazione *user-friendly* dell'andamento delle variabili rilevanti. Nell'ambito di questo progetto, viene proposta una soluzione in grado di eseguire automaticamente le attività elencate. Tale sistema è stato interamente scritto e testato all'interno dell'ambiente di programmazione `Python`[8].
Il programma, dunque, consta di tre elementi principali. Innanzitutto, si ha una componente di *data ingestion*, la quale permette di ricevere il flusso di dati osservati dai sensori durante il processo di creazione di ogni pompa, ed elaborarlo per due scopi separati:

- rendere i dati adatti alla proiezione grafica sullo schermo;
- registrare metadati relativi alle misurazioni ricevute, *timestamp* e valore decodificato, e segnalare se tale dato consiste in un *outlier* univariato, aggiornando ad ogni *step* il calcolo del metodo della distanza interquartile.

Questa operazione è stata simulata attraverso il client del *software* `Apache Kafka`[9], che si presta perfettamente al *task* definito: si configura, infatti, come una piattaforma *open source* a bassissima latenza per la gestione di *feed* in tempo reale. Grazie al suo utilizzo, il sistema è in grado di ricevere grandi moli di osservazioni in intervalli di tempo molto ridotti, effettuare le elaborazioni prima citate ad altissima velocità e visualizzare il risultato grafico senza dover ritardare o ricaricare l'applicazione, e senza che l'utente debba controllare le operazioni in *background*. Nel dettaglio, il sensore invia le misurazioni attraverso un `Producer` di `Apache Kafka`, e il sistema, connettendosi a un `Consumer`, effettua una lettura continua di questi dati all'interno del *topic* in cui sono stati memorizzati. La lettura di un nuovo valore comporta, poi, l'estrazione dei metadati. Ad ogni sensore verrà applicato un `Producer` diverso, e per ciascuno verrà aperta una connessione verso il sistema tramite un `Consumer` separato, in modo da mantenere le letture indipendenti ed evitare un sovraccarico.
Il secondo elemento consiste in una componente grafica, curata mediante i *tool* di visualizzazione forniti dall'interfaccia `Dash` di `Plotly`[10]. Essa riceve i valori letti dalla *topic* di `Apache Kafka` e aggiorna in tempo reale due tipologie di visualizzazione per i dati inviati dai sensori: un *BoxPlot* per monitorare la distribuzione totale, e un grafico *LineChart* per l'andamento.

[8] Il codice è disponibile al seguente link: `https://github.com/RCrvro/Industry-Lab---Progetto/tree/master/Monitoraggio%20realtime`.

[9] Documentazione ufficiale del client di `Apache Kafka` in `Python`: `https://pypi.org/project/kafka-python/`.

[10] Documentazione ufficiale di `Dash`: `https://plotly.com/dash/`.

Infine, la terza componente - nominata "`writer`"[11] - esegue un programma indipendente, che registra i metadati - *timestamp*, dato numerico e segnalazione di eventuali valori anomali - all'interno di un database locale. Quest'ultimo, dunque, svolgerà la funzione di *logfile* del processo.

4.1 Demo

È disponibile una demo del sistema di monitoraggio al link **`youtu.be/R7G`**.
In questo caso, si è scelto di simulare la visualizzazione di ipotetiche osservazioni relative al coefficiente di dispersione α (in verde) e alla media di portata a 140rpm (in arancione), perché, essendo variabile soggetta a limiti imposti in fase di produzione, appare ragionevole monitorarla in tempo reale.
Nel video, sul lato destro dello schermo, sono visibili quattro pagine del terminale. La prima in alto esegue un `Producer` per simulare l'invio delle misurazioni del coefficiente α da parte del proprio sensore, tramite scrittura manuale di alcuni valori. La seconda attiva un secondo `Producer`, per simulare il sensore della variabile di portata. La terza implementa il programma `writer`: ogni dato ricevuto dal sensore - in questo caso di α - viene scritto, assieme ai metadati, nel *logfile* `"Coefficiente.csv"` apparso sul bordo destro del monitor dopo l'invio del primo valore. Alla fine del video, il file verrà aperto, per mostrare il risultato delle registrazioni. Inoltre, sempre per quanto concerne il `writer`, la pagina mostrerà un conteggio dei messaggi ricevuti. Infine, l'ultima finestra in basso inizializza l'applicazione grafica in corrispondenza della porta `8050` e mantiene connesso il sistema di monitoraggio.
L'applicazione è programmata per la segnalazione di eventuali errori, che possono essere esaminati cliccando l'icona blu nell'angolo inferiore destro della pagina *Web*.

5. Limite dinamico di portata

Poiché la *performance* della pompa GP5 dipende strettamente dal proprio valore di portata in uscita, è necessario controllare che tale misurazione non scenda al di sotto di una determinata soglia, variabile a seconda della macro-categoria selezionata: *Daimler* o Standard[12]. Oltre al tipo di "programma", però, tale grandezza condivide una relazione matematica con le variabili del coefficiente di *leakage* α e della pressione (equazione 4). Pertanto, il processo produttivo necessita di un monitoraggio dei valori di portata in uscita rispetto alle colonne menzionate. A tal proposito, oltre ad un algoritmo che calcoli, a partire da α e pressione, la portata in uscita e valuti se tale misurazione rientra nel *range* definito dalla miglior[13] e peggior osservazione in corrispondenza della categoria GP5, è possibile rappresentare graficamente tale limite dinamico come una regione ammissibile all'interno di uno spazio tridimensionale, e osservare la differenza fra il piano e una data misurazione. In questo modo, è possibile monitorare già graficamente la presenza di eventuali anomalie, e si è in grado di valutare visivamente la distanza dai comportamenti tipici del processo produttivo.
È disponibile una demo dell'algoritmo di rappresentazione grafica del limite dinamico di portata al link **`youtu.be/cxH`**: date in input le coordinate di α e pressione, viene generata la regione ammissibile prima descritta (in blu) e indicato il punto relativo all'osservazione. La regione critica si modifica a seconda della categoria di pompa, e il colore del punto varia in base alla correttezza dell'osservazione fornita in input: se essa rientra nel limite indicato per la data classe GP5, il punto è rappresentato in verde (Figura 18), altrimenti in rosso (Figura 19).

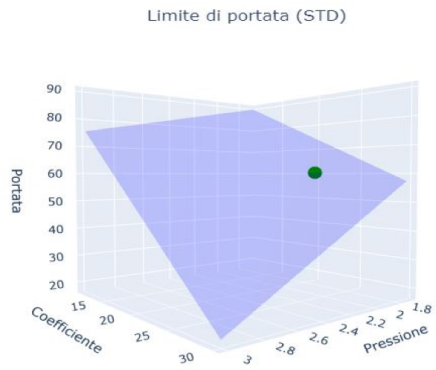


Figura 18. Il punto osservato è compreso nella regione ammissibile. Ciò significa che i valori di coefficiente di dispersione, pressione e soprattutto portata - in funzione di essi - sono da considerarsi regolari. Il punto è pertanto colorato in verde.

6. Modelli di previsione

La previsione del coefficiente di perdita α (*leakage coefficient*) è stata effettuata utilizzando tre diverse tecniche di regressione:

1. Modello OLS per regressione lineare multipla
2. Tecniche di *shrinkage*: regressione *Ridge* e *Lasso*
3. Modello misto: regressione *multilevel*

6.1 Modelli OLS

La regressione lineare rappresenta un metodo di stima del valore atteso condizionato di una variabile - dipendente (o endogena) -, dati i valori di altre variabili indipendenti (o esogene), assumendo pertanto l'esistenza di una relazione

[11] Il codice del "`writer`" è disponibile al link: `https://github.com/RCrvro/Industry-Lab---Progetto/blob/master/Monitoraggio%20realtime/writer.py`.

[12] I valori corretti di portata in uscita per la pompa *Daimler* sono compresi fra 45.86 e 63.36. Per quanto riguarda la categoria Standard, gli estremi considerati sono 57.36 e 75.36.

[13] La regione ammissibile viene limitata superiormente perché eventuali valori oltre il migliore registrato potrebbero derivare da anomalie del pezzo prodotto.

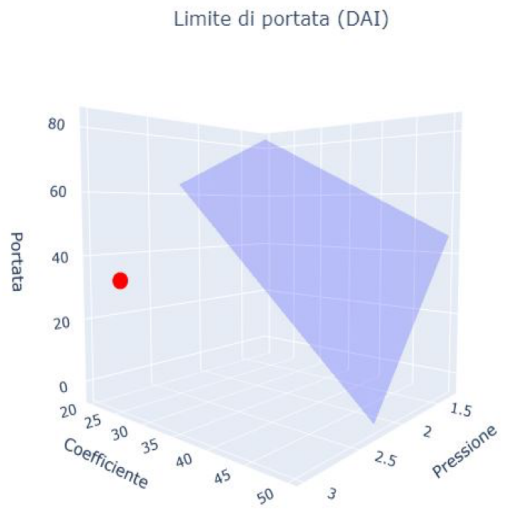


Figura 19. Il punto osservato non è compreso nella regione ammissibile. Ciò significa che l'osservazione è anomala, ed è pertanto segnalata in rosso.

lineare tra le X e la variabile target Y. La formulazione è rappresentata dalla seguente equazione:

$$Y = \beta_0 + \beta_1 x_1 + \beta_n x_n + u_i \tag{3}$$

dove:

- β_0 è l'intercetta della retta di regressione, ovvero il valore atteso di Y quando le altre variabili indipendenti sono pari a zero
- β_1, β_n sono i coefficienti angolari della retta di regressione, interpretabili come pesi di rilevanza assegnati a ciascuna *feature*
- u_i è l'errore statistico.

Come discusso in precedenza, le variabili di portata e pressione si distribuiscono quasi identicamente, sia per quanto riguarda le diverse fasi - di controllo e a regime -, sia per quanto concerne le due diverse misurazioni di media e picco. Pertanto, si è proceduto inizialmente a stimare alcun modelli per la previsione del coefficiente di *leakage* α mantenendo separate le variabili di portata e di pressione. Inoltre, dato che la collinearità tra tali colonne avrebbe potuto causare una sottostima della significatività delle X e una sovrastima dell'indice di bontà di adattamento R^2, si è deciso di includere soltanto le grandezze di pressione e portata che presentavano una correlazione pari o inferiore all'82%.
La significatività statistica delle variabili verrà estratta dal confronto fra il *p-value*, ($Pr(>|t|)$), e il livello di significatività statistica, prefissato 0.05. Il risultato della regressione è quindi basato sull'ipotesi che i coefficienti di regressione associati alle variabili siano nulli, contrapposta all'ipotesi alternativa, per cui essi non sono uguali a zero ed esiste cioè una relazione tra la Y e il regressore selezionato. Più precisamente, soltanto quando *p-value* < 0.05 i β sono significativamente diversi da zero, pertanto rilevanti. Nelle tabelle con cui verranno esposti i risultati di ogni modello OLS, il grado di significatività - che dipende dall'ordine della differenza fra il *p-value* e il limite prefissato - sarà indicato da alcuni asterischi: maggiore il numero degli stessi, da un minimo di zero a un massimo di 3, maggiore la rilevanza della variabile in questione per la previsione della Y.
Per migliorare la stima della funzione di regressione e dell'indice R^2 per ciascun modello *ordinary least squares*, è stata applicata la tecnica *k-folds cross validation* (con $k = 5$): i dati sono suddivisi iterativamente in *training* e *test set*, ristimando e valutando k volte il modello ottenuto, in modo da evitare problemi di sovradattamento e campionamento asimmetrico. Le prossime sezioni riassumono i risultati ottenuti da ciascun modello.

6.1.1 Modello OLS con variabili di portata

Modello 1: Coefficiente ∼ Picco Port. (140) + Media Port.(2300) + Picco Port. (2300) + Temperatura + Media Coppia Finale

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	0.04440
Picco Port. (140)	<2e-16 ***	-0.07284
Media Port. (2300)	<2e-16 ***	-0.18002
Picco Port. (2300)	<2e-16 ***	0.20332
Temperatura	<2e-16 ***	-0.00152
Media Coppia Finale	<2e-16 ***	0.01893

Per la previsione del coefficiente di *leakage* sono state utilizzate, oltre alle variabili di portata che presentano una reciproca correlazione inferiore al 82%, le variabili "Temperatura" e "Media Coppia Finale"[14]. Tutte le variabili risultano statisticamente significative con grado molto elevato. La variabile più rilevante per la previsione di α si rivela essere *Picco Port. (2300)*, con un coefficiente di regressione pari a ∼0.2. Inoltre, è possibile notare una relazione negativa tra la variabile target e le seguenti covariate: picco di portata in fase di controllo (140 rpm), media di portata in fase a regime (2300 rpm) e temperatura. Il coefficiente di determinazione (R^2) associato al modello è abbastanza buono, pari a 0.8737.

Modello 2: Coefficiente ∼ Picco Port. (140) + Media Port.(2300) + Picco Port. (2300) + Temperatura + Picco Coppia Finale
A differenza del precedente caso, la generazione di un modello di previsione del coefficiente ha implicato la sostituzione di "Picco Coppia Finale" a "Media Coppia Finale". Anche qui, tutte le variabili risultano statisticamente significative con grado identicamente elevato. La variabile più rilevante è ancora il picco di portata in fase a regime, caratterizzato da un leggero aumento del proprio coefficiente di regressione.

[14] Nel modello non vengono mai incluse entrambe le variabili relative alla coppia finale, perchè presentano una correlazione eccessiva e rischiano pertanto di compromettere la correttezza delle stime.

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	0.045667
Picco Port. (140)	<2e-16 ***	-0.073420
Media Port. (2300)	<2e-16 ***	-0.202027
Picco Port. (2300)	<2e-16 ***	0.225690
Temperatura	<2e-16 ***	-0.001174
Picco Coppia Finale	<2e-16 ***	0.007755

Il coefficiente di determinazione (R^2) associato al modello è leggermente inferiore al precedente: 0.872.

In seguito, si è proceduto a combinare le colonne relative alla coppia finale - considerate separatamente[15] - e alla temperatura con le singole grandezze di portata che mostravano una correlazione $\leq 82\%$.

Modello 3: Coefficiente $\sim$ Picco Port. (140) + Temperatura + Media Coppia Finale

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	0.076790
Picco Port. (140)	<2e-16 ***	-0.106135
Temperatura	<2e-16 ***	-0.005642
Media Coppia Finale	<2e-16 ***	0.033331

Tutte le variabili mostrano un grado di significatività statistica molto elevato. La variabile più rilevante è il picco di portata in fase di controllo, che presenta una relazione negativa con il *coefficiente*. La motivazione, qui, è ovvia: il picco di portata in fase di controllo è identicamente distribuito alla media di portata a 140 rpm e correlato al 91% con la media di pressione in fase di controllo, che sono le due variabili con cui α è stato generato (equazione 4). Il coefficiente di determinazione (R^2) si abbassa ancora: $\sim$0.81.

Modello 4: Coefficiente $\sim$ Picco Port. (140) + Temperatura + Picco Coppia Finale

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	0.07968
Picco Port. (140)	<2e-16 ***	-0.10728
Temperatura	<2e-16 ***	-0.00495
Picco Coppia Finale	<2e-16 ***	0.01119

Tutte le variabili risultano statisticamente molto significative. La variabile più rilevante è ancora, per le ragione ovvie prima menzionate, il picco di portata in fase di controllo. Il coefficiente di determinazione (R^2) è pari a 0.803.

Modello 5: Coefficiente $\sim$ Media Port.(2300) + Temperatura + Media Coppia Finale

[15] Nota precedente.

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	-0.0103293
Media Port. (2300)	<2e-16 ***	0.0479580
Temperatura	<0.0128 *	-0.0003287
Media Coppia Finale	<2e-16 ***	0.0368901

In questo caso, la variabile "Temperatura" rivela livello di significatività nettamente minore rispetto alle altre variabili. Nonostante le altre variabili risultano rilevanti per determinare il valore di α, la capacità del modello nel catturare la variabilità dei dati subisce comunque una grave diminuzione, producendo un R^2 pari a 0.7443.

Modello 6: Coefficiente $\sim$ Media Port.(2300) + Temperatura + Picco Coppia Finale

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	-0.00876
Media Port. (2300)	<2e-16 ***	0.04856
Temperatura	0.333	
Picco Coppia Finale	<2e-16 ***	0.02097

La variabile "Temperatura" non risulta essere statisticamente significativa: il *p-value* ad essa associato è molto superiore al limite di 0.05, per cui l'ipotesi nulla che il suo coefficiente di regressione sia nullo non viene rifiutata. Si è dunque proceduto a ristimare la funzione di regressione senza tale colonna, ottenendo un R^2 di 0.7393.

Modello 7: Coefficiente $\sim$ Picco Port. (2300) + Temperatura + Media Coppia Finale

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	-0.0105056
Picco Port. (2300)	<2e-16 ***	0.0486005
Temperatura	0.342	
Media Coppia Finale	<2e-16 ***	0.0342434

Anche qui, la temperatura non risulta rilevante ai fini della previsione e viene nuovamente scartata. Le altre *feature* contribuiscono in misura ridotta alla determinazione del coefficiente. L'indice di bontà di adattamento (R^2) associato al modello è pari a 0.7450919.

Modello 8: Coefficiente $\sim$ Picco Port. (2300) + Temperatura + Picco Coppia Finale
La "Temperatura" non risulta nuovamente significativa. Come in precedenza, le altre variabili sono poco rilevanti e producono un R^2 di $\sim$0.74.

6.1.2 Modelli di previsione con variabili di pressione

Modello 1: Coefficiente $\sim$ Picco Press. (140) + Media Press.(2300) + Picco Press. (2300) + Temperatura + Media

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	-0.00904
Picco Port. (2300)	<2e-16 ***	0.04919
Temperatura	0.728	
Picco Coppia Finale	<2e-16 ***	0.01982

Coppia Finale

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	0.065700
Picco Press. (140)	<2e-16 ***	-0.080396
Media Press. (2300)	<2e-16 ***	-0.025513
Picco Press. (2300)	<2e-16 ***	0.036890
Temperatura	0.241	
Media Coppia Finale	<2e-16 ***	-0.001809

La variabile "Temperatura" non risulta statisticamente significativa e viene dunque scartata per la previsione del coefficiente di perdita. Al contrario, i valori dei coefficienti di regressione associati alle altre variabili risultano molto rilevanti nella previsione del *leakage coefficient*. Il modello offre un'ottima misura di performance (R^2), pari a 0.9507.

Modello 2: Coefficiente $\sim$ Picco Press. (140) + Media Press.(2300) + Picco Press. (2300) + Temperatura + Picco Coppia Finale

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	0.065649
Picco Press. (140)	<2e-16 ***	-0.080372
Media Press. (2300)	<2e-16 ***	-0.024678
Picco Press. (2300)	<2e-16 ***	0.036036
Temperatura	0.0614	
Picco Coppia Finale	<2e-16 ***	-0.001603

La temperatura continua ad essere eliminata in quanto non significativa e viene eliminata dal modello. Le altre variabili, invece, contribuiscono efficacemente alla descrizione dell'andamento di α: il coefficiente di determinazione (R^2) è pari a 0.9508.
Tuttavia, è possibile che il fattore penalizzante della multicollinearità, da cui è affetta la matrice del disegno, distorca i rispettivi risultati della regressione. Per tale ragione, è stato necessario valutare successivi modelli, che includono separatamente le grandezze di pressione con una correlazione $\leq 82\%$, combinate con "Temperatura", "Media Coppia Finale" e "Picco Coppia Finale". Le due variabili relative alla coppia finale continuano ad essere considerate separatamente per i motivi prima espresse.

Modello 3: Coefficiente $\sim$ Picco Press. (140) + Temperatura + Media Coppia Finale

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	0.0829343
Picco Press. (140)	<2e-16 ***	-0.0968194
Temperatura	<2e-16 ***	-0.0007664
Media Coppia Finale	<2e-16 ***	-0.0021630

Tutte le variabili sono statisticamente significative. La variabile più rilevante per prevedere il coefficiente di perdita è il picco di pressione in fase di controllo (140 rpm), caratterizzato da una relazione negativa con α. La ragione, come visto in precedenza, è facilmente deducibile: tale colonna è identicamente ditribuita alla media di pressione durante la medesima fase e correlata al 92% con la media di portata a 140 rpm, le due variabili con cui la Y è stata generata (equazione 4). l coefficiente di determinazione (R^2) associato al modello è pari a 0.938.

Modello 4: Coefficiente $\sim$ Picco Press. (140) + Temperatura + Picco Coppia Finale

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	0.082962
Picco Press. (140)	<2e-16 ***	-0.096812
Temperatura	<2e-16 ***	-0.000584
Picco Coppia Finale	<2e-16 ***	-0.003744

Tutte le variabili sono statisticamente significative e caratterizzate da una relazione negativa con il coefficiente di dispersione. Come ci si poteva attendere, "Picco Press. (140)" risulta ancora la più rilevante. La bontà di adattamento rimane eccellente: 0.938.

Modello 5: Coefficiente $\sim$ Media Press.(2300) + Temperatura + Media Coppia Finale

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	-0.007341
Media Press. (2300)	<2e-16 ***	0.048612
Temperatura	0.00271 **	-0.000396
Media Coppia Finale	<2e-16 ***	0.036867

Qui, la variabile "Temperatura" è leggermente meno significativa, e con molta probabilità questo fattore è causa della riduzione della capacità del modello di catturare la variabilità dei dati, offrendo un R^2 di $\sim$0.745.

Modello 6: Coefficiente $\sim$ Media Press.(2300) + Temperatura + Picco Coppia Finale
In questo caso, la temperatura non è addirittura significativa per la previsione del target, e viene dunque scartata dal modello. A seguito di quanto detto, è facile dedurre la ragione per cui il coefficiente di correlazione (R^2) subisca un'ulteriore

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	-0.00876
Media Press. (2300)	<2e-16 ***	0.04856
Temperatura	0.333	
Picco Coppia Finale	<2e-16 ***	0.02097

contrazione a ∼0.74.

Modello 7: Coefficiente ∼ Picco Press.(2300) + Temperatura + Media Coppia Finale

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	-0.0071346
Picco Press. (2300)	<2e-16 ***	0.0488265
Temperatura	0.000201 ***	-0.0004871
Media Coppia Finale	<2e-16 ***	0.0316346

Tutte le variabili sono statisticamente significative. Nonostante ciò, l'indice di bontò di adattamento (R^2) è basso: 0.7487.

Modello 8: Coefficiente ∼ Picco Press.(2300) + Temperatura + Picco Coppia Finale

Variabile	$Pr(>\|t\|)$	Coefficiente
Intercetta	<2e-16 ***	-0.0058528
Picco Press. (2300)	<2e-16 ***	0.0493986
Temperatura	0.00187**	-0.0004141
Picco Coppia Finale	<2e-16 ***	0.0194586

Tutte le variabili sono rilevanti, ma, come visto nel modello 7, la performance (R^2) è più scarsa (0.745).

6.2 Tecniche di *shrinkage*

Poichè i dati presentano un fortissimo problema di multicollinearità, risulta impossibile includere la totalità delle colonne in un modello di regressione lineare per la previsione del coefficiente α, ed è quindi necessario ridurre la matrice del disegno. L'estrazione di un subset ottimale di *feature* può avvenire tramite vari tentativi di selezione manuale - come effettuato nella sezione precedente -, ma tale metodo può essere troppo dispendioso. In alternativa, è possibile eseguire una *feature selection* implicita tramite aggiunta di una componente di penalizzazione alla funzione di regressione. L'applicazione di questo meccanismo di regolarizzazione, ponendo un vincolo al valore dei coefficienti di regressione, contrae le stime β e quindi riduce la rilevanza dei candidati esplicativi. In questo modo, le colonne che meno contribuiscono a descrivere la variabilità di α, subiscono un'ulteriore riduzione della propria importanza. Qualora la contrazione comportasse l'annullamento di un coefficiente di regressione, allora il predittore associato sarebbe indirettamente scartato dal modello. In questo caso, le tecniche di *shrinkage* avranno successo se riusciranno a filtrare le poche colonne realmente importanti, eliminando l'informazione ridondante presente nelle distribuzioni eccessivamente correlate.
Sono state adottate due popolari tipologie di regolarizzazione: *Ridge* e *Lasso*.

6.2.1 Regressione *Ridge*

La stimatore lineare *Ridge* [2] è formulato come segue:

$$\hat{\beta}^{\lambda} = (X'X + \lambda I_p)^{-1} X'y \tag{4}$$

dove λ è il parametro di regolarizzazione pari alla norma L_2, tale per cui:

- se $\lambda = 0$, la soluzione coincide con il quella derivante dal metodo ai minimi quadrati ordinari;
- se $\lambda > 0$, si produce una penalizzazione sul punto di minimo della funzione di perdita $D^{(\beta,\lambda)}_{Ridge}$.

In questo modo, la complessità del modello diminuisce senza eliminare nessuna variabile.
Il valore atteso $\mathbb{E}[\hat{\beta}^{\lambda}]$ si dimostra essere pari a

$$(X'X + \lambda I_p)^{-1}(X'X)\beta \neq \beta \tag{5}$$

, dimostrando quindi che $\hat{\beta}^{\lambda}$ è distorto. Tuttavia, nonostante l'introduzione del *bias* appena menzionato, è possibile dimostrare come questa tecnica di *shrinkage*, oltre ad avere l'effetto di risolvere efficacemente la supercollinearità fra i regressori, riesca a produrre, per un'opportuna scelta di λ, un errore quadratico medio dello stimatore minore rispetto a quello offerto dal metodo OLS.
Dunque, per uno stimatore ridge, la selezione di un buon valore per λ è fondamentale: per trovare l'ottimo è stata utilizzato un metodo di ricerca a griglia (*Grid Search*), ottenendo un parametro quasi nullo, pari cioè a 0.01. Poichè la penalizzazione applicata è quasi nulla, è possibile dedurre che, nonostante le aspettative, la regolarizzazione offerta da una norma L_2 sia quasi inutile. Precisamente, nel modello sono state considerate congiuntamente tutte le variabili originarie - relative a pressione e portata, temperatura, e coppia finale. I risultati ottenuti sono riassunti in Figura 20. È possibile notare come la variabile "Picco Press. (140)" risulti la più rilevante, ovvero la meno contratta dalla penalizzazione. La ragione risiede, come spiegato in precedenza, nella fortissima dipendenza lineare fra tale regressore e le distribuzioni con cui è stato generato il coefficiente di dispersione. Al contrario, la "Temperatura" non contribuisce alla previsione, e il suo β è pertanto annullato. Per il resto, la relativa importanza del resto dei candidati appare nettamente ridimensionata, ad indicare la grande ridondanza informativa all'interno della matrice del disgeno. Infine, il coefficiente di determinazione (R^2) associato al modello è pari a 0.845, attestandosi pertanto su un livello accettabile, ma non eccellente.

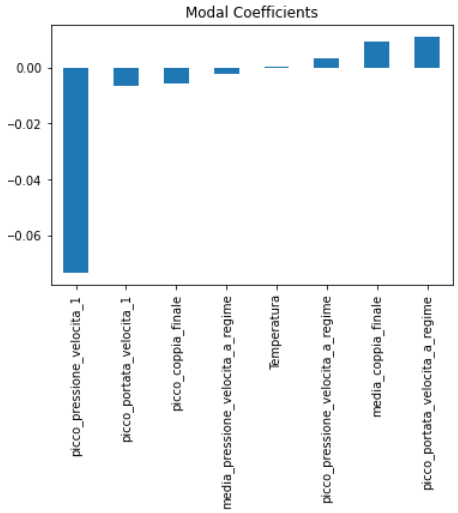


Figura 20. Stime *Ridge* dei coefficienti di regressione.

6.2.2 Regressione *Lasso*

Nonostante sia un'ottima soluzione per contrarre la varianza della funzione di regressione, il metodo *Ridge* è, in realtà, raramente in grado di effettuare una vera e propria *feature selection*, ovvero di annullare i coefficienti che moltiplicano le variabili meno rilevanti per descrivere il fenomeno di studio. Questa limitazione deriva dalla scelta di una norma L_2 come componente di penalizzazione. La regressione *Lasso* [3] può essere pertanto utilizzata come alternativa ai metodi di *feature selection*, perché, utilizzando una funzione di regolarizzazione non derivabile - norma L_1, è in grado di ridurre ulteriormente i coefficienti di regressione verso lo zero. Anche qui, la selezione del valore ottimale di λ è stata effettuata utilizzando un metodo di ricerca a griglia (*Grid Search*), ottenendo un parametro ottimo praticamente nullo (0.00001). Le stime *Lasso* dei coefficienti sono rappresentate graficamente in Figura 21. L'unica variabile ritenuta importante per la previsione di α

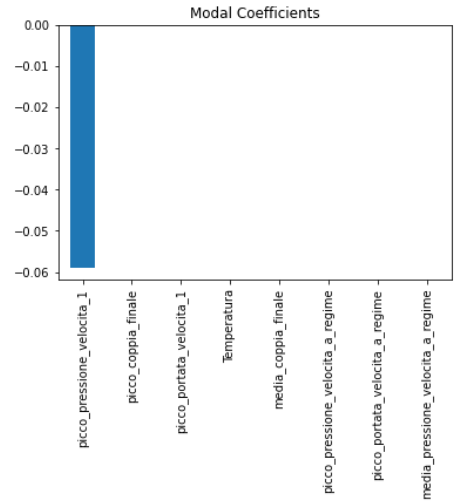


Figura 21. Stime *Lasso* dei coefficienti di regressione.

è sempre il picco di pressione in fase di controllo (140 rpm). Ogni altro candidato è scartato, come conseguenza logica dell'eccessiva dipendenza fra tutte le principali misurazioni che descrivono il processo di produzione delle pompe GP5. L'indice di bontà di adattamento (R^2) è scarso, pari a 0.7058.

6.3 Regressione Multilevel

I *Multilevel Model* [4][5], noti anche come gerarchici lineari, sono modelli statistici caratterizzati da parametri che variano su più livelli. Come anticipato nella Sezione 3.3, i dati a disposizione - riguardanti principalmente le performance del prodotto - variano a seconda dei sottogruppi di pompe. La regressione *Multilevel* appare quindi perfetta per il fenomeno di studio, permettendo di considerare nel modello finale i fattori contestuali che subentrano nei diversi programmi di produzione, ovvero le differenti relazioni, da gruppo a gruppo, fra la variabile target e i regressori X.
Sono stati pertanto costruiti vari modelli multilivello, considerando separatamente le variabili di pressione e portata - sempre con lo scopo di gestire il problema di multicollinearità - , oltre agli altri candidati esplicativi già menzionati. La variabile di *grouping* "Programma"[16], è indicata nella formulazione del modello con la scrittura "$(1|Programma)$"[17].

6.3.1 Modelli multilivello con variabili di portata

Modello 1: Coefficiente ∼ Picco Port. (140) + Media Port.(2300) + Picco Port. (2300) + Temperatura + Media Coppia Finale + (1|Programma)

Variabile	$Pr(>\|t\|)$
Intercetta	<2e-16 ***
Picco Port. (140)	<2.2e-16 ***
Media Port. (2300)	<2.2e-16 ***
Picco Port. (2300)	<2.2e-16 ***
Temperatura	<2.2e-16 ***
Media Coppia Finale	<2.2e-16 ***

Tutte le variabili risultano statisticamente significative. Il coefficiente di determinazione (R^2) associato al modello è elevato, pari a 0.918. È possibile misurare la differenza dei gruppi di programma, in termine di diversa funzione di regressione che mette in relazione le variabili indipendenti con α, considerando il valore del coefficiente di dispersione, la Y, come somma di tre componenti:

$$y_{ij} = \mu + a_j + \varepsilon_{ij} \tag{6}$$

dove

- y_{ij} è l'i-esima osservazione nel j-esimo gruppo
- μ è interpretabile come valore atteso globale di tutti i dati originari

[16] Con "variabile di *grouping*" s'intende la colonna con cui vengono raggruppate le osservazioni, con lo scopo di ottenere una stima gerarchica, ovvero estrarre i diversi vettori dei coefficienti di regressione per ciascun gruppo.
[17] Il valore 1 prima del simbolo | sta ad indicare la stima di un'intercetta.

- a_j è l'effetto casuale che caratterizza tutte le osservazioni di un dato gruppo, interpretabile pertanto come l'influenza che i fattori contestuali hanno sulle misurazioni della classe j, ovvero lo scostamento del gruppo j-esimo rispetto a trend globale

- ε_{ij} è il rumore casuale proprio della data osservazione.

Nella stima di un classico modello di regressione lineare, il fattore contestuale a_j è assunto nullo, poichè non viene ipotizzata l'esistenza di una gerarchia. Nel caso di una regressione multilivello, invece, il grado in cui i fattori contestuali comportano dinamiche diverse all'interno dei gruppi è misurato dal cosiddetto coefficiente di correlazione intraclasse (ICC), che consiste nel rapporto fra la varianza degli effetti casuali σ_a e la varianza totale dei dati (come somma di σ_a e la varianza del rumore casuale di ogni osservazione σ_ε):

$$\frac{\sigma_a}{\sigma_a + \sigma_\varepsilon} \tag{7}$$

Pertanto, al crescere di questo indice, cresce la certezza nell'affermare l'esistenza di una gerarchia sistematica all'interno dei dati.
In questo caso, il valore del coefficiente di correlazione intraclasse (ICC), pari a 0.912, conferma l'ipotesi formulata, ovvero che il coefficiente di dispersione α possa essere descritto similmente nell'ambito di ciascun programma e in maniera significativamente diversa tra le varie classi. La presenza di gruppi significativamente diversi rispetto al trend globale, ovvero di classi GP5 che subiscono scostamenti rilevanti a causa dei propri fattori contestuali di produzione, può essere verificata graficamente come in Figura 22.

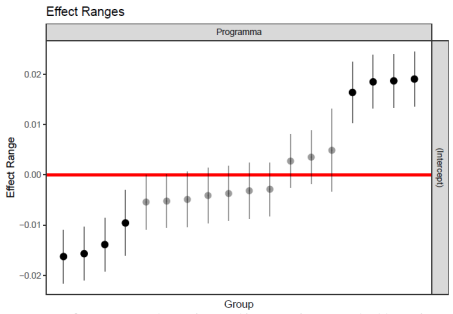


Figura 22. Grafico per la visualizzazione della significatività dell'effetto casuale di gruppo a_j.

I punti in tale rappresentazione indicano il valore dell'effetto casuale di ciascun gruppo a; se l'intervallo di confidenza attorno all'effetto casuale a_j comprende lo zero - ovvero interseca la linea rossa in Figura -, allora significa che la classe j-esima non subisce uno scostamento significativo rispetto al trend globale, cioè che i propri fattori contestuali non causano la formazione di un cluster di produzione a sé stante. Al contrario, se l'intervallo di confidenza si allontana dallo zero, allora il gruppo può considerarsi parte di una gerarchia sistematica sottostante ai dati. Qui, il numero di cluster di produzione è pari a 8 su 18.

Modello 2: Coefficiente $\sim$ Picco Port. (140) + Media Port.(2300) + Picco Port. (2300) + Temperatura + Picco Coppia Finale + (1|Programma)

Variabile	$Pr(>\|t\|)$
Intercetta	<2e-16 ***
Picco Port. (140)	<2.2e-16 ***
Media Port. (2300)	<2.2e-16 ***
Picco Port. (2300)	<2.2e-16 ***
Temperatura	<2.2e-16 ***
Picco Coppia Finale	<2.2e-16 ***

Tutte le variabili risultano statisticamente significative. Il coefficiente di determinazione (R^2) associato al modello è pari a 0.921. L'esistenza di una gerarchia nei dati è riconfermata, anche in questo caso, dall'elevato coefficiente di correlazione intraclasse (ICC): 0.915, e dalla presenza di effetti contestuali significativi in 9 cluster, visibili in Figura 23.

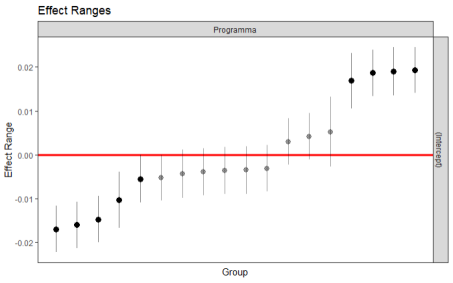


Figura 23. Grafico per la visualizzazione della significatività dell'effetto casuale di gruppo a_j.

6.3.2 Modelli di previsione con variabili di pressione

Modello 1: Coefficiente $\sim$ Picco Press. (140) + Media Press.(2300) + Picco Press. (2300) + Temperatura + Media Coppia Finale + (1|Programma)

Variabile	$Pr(>\|t\|)$
Intercetta	<2e-16 ***
Picco Press. (140)	<2.2e-16 ***
Media Press. (2300)	<2.2e-16 ***
Picco Press. (2300)	<2.2e-16 ***
Temperatura	0.9759
Media Coppia Finale	0.1177

Le variabili *Temperatura* e *Media Coppia Finale* non risultano statisticamente significative e sono dunque scartate. Nonostante il minor numero di predittori, il coefficiente di determinazione (R^2) associato al modello è superiore a quanto visto nell'ambito delle misurazioni di portata, pari a 0.939. Il valore del coefficiente di correlazione intraclasse (ICC) è, tuttavia, nettamente inferiore: 0.797. Ciò indica che le relazioni fra le grandezze di pressione - in combinazione con gli altri candidati - e la variabile target α variano in maniera decisamente

minore fra le varie classi di pompa rispetto a quanto osservato per quanto concerne la portata. Nonostante ciò, la Figura 24, segnala un numero maggiore di effetti contestuali significativi: 10 su 18.

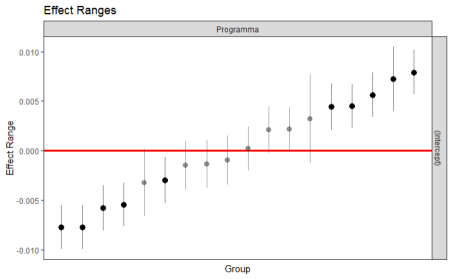


Figura 24. Grafico per la visualizzazione della significatività dell'effetto casuale di gruppo a_j.

Modello 2: Coefficiente $\sim$ Picco Press. (140) + Media Press.(2300) + Picco Press. (2300) + Temperatura + Picco Coppia Finale + (1|Programma)

Variabile	$Pr(>\|t\|)$
Intercetta	<2e-16 ***
Picco Press. (140)	<2.2e-16 ***
Media Press. (2300)	<2.2e-16 ***
Picco Press. (2300)	<2.2e-16 ***
Temperatura	0.6196549
Picco Coppia Finale	0.0001283 ***

La variabile "Temperatura", in questo preciso caso, non è statisticamente significativa per la previsione. L'indice di bontà di adattamento (R^2) è sempre eccellente (0.939), e il coefficiente di correlazione intraclasse (ICC) è ancora moderato (0.797), confermando la presenza di una gerarchia meno sistematica nei dati riguardati la pressione. Anche qui, nonostante il minor livello di varianza intraclasse, possono essere individuati 10 cluster di produzione.

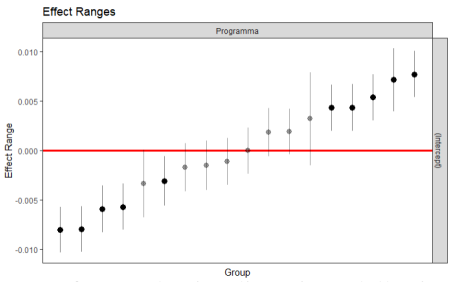


Figura 25. Grafico per la visualizzazione della significatività dell'effetto casuale di gruppo a_j.

6.4 Conlclusioni sui modelli predittivi

Tutti i modelli proposti sono in grado di prevedere il coefficiente di perdita con un'accuratezza accettabile: il minimo valore dell'indice di bontà di adattamento è 0.7058, ottenuto con una regolarizzazione di tipo *Lasso*. Le regressioni regolarizzate risentono negativamente di un'eccessiva collinearità della matrice del disegno - come dimostrato empiricamente da Weisberg [6][7] -, e questo viene confermato dal fatto che anche l'R^2 prodotto al metodo *Ridge* non sia eccellente: 0.845. Per questa ragione, le tecniche di *shrinkage* sono sconsigliate. Il modello migliore ($R^2 = 0.9508$) per la previsione di α consiste nella regressione lineare classica basato sulla combinazione di alcune variabili di pressione ("Picco Press. (140)", "Media Press.(2300)","Picco Press. (2300)"), temperatura e picco di coppia finale. Segue il modello OLS che include le stesse variabili di quello appena citato, ecceto per quanto riguarda il picco della coppia finale, sostituito dalla media della coppia finale. Pertanto, le regressioni lineari multiple basate sulle misurazioni di pressione in entrambe le fasi si rivelano essere le metodologie più accurate per la descrizione della variabile target. La seconda classe di metodi, in ordine di prestazione, è quella dei modelli multilivello, che, riuscendo a catturare gli effetti randomici propri dei diversi contesti di produzione, ottengono valori di R^2 compresi fra un minimo di 0.918 (quando la regressione è basata su grandezze di portata) e un massimo di 0.939 (in corrispondenza delle colonne di pressione). Nell'ambito di tali modelli misti, è interessante nota come quelli basati sulle misurazioni di pressione mostrino sempre un coefficiente ICC molto minore, quindi una struttura gerarchica meno evidente, ma un numero di cluster di produzione sempre maggiore. In ogni caso, il grande problema di multicollinearità non ha permesso di utilizzare congiuntamente sia le variabili di portata e pressione che quelle relative alla coppia finale, né di condurre una *feature selection* esplicita, obbligando una dispendiosa serie di tentativi di selezione manuale.

Codice

L'intero codice, implementato con linguaggi `Python` e `R`, è disponibile al link: `https://github.com/RCrvro/Industry-Lab---Progetto`.

Riferimenti bibliografici

[1] T. Hastie, R. Tibshirani, J. Friedman; *"The Elements of Statistical Learning: Data Mining, Inference, and Prediction"*, Stanford University, Stanford, 2009.

[2] A.E. Hoerl, R.W. Kennard; *Ridge regression: Biased estimation for nonorthogonal problems*, in *Technometrics*, vol. 42, n. 1, pp. 80–86, 1970.

[3] R. Tibshirani; *Regression Shrinkage and Selection via the lasso* in *Journal of the Royal Statistical Society. Series B (methodological)*, 1996.

[4] S. W. Raudenbush, A. S. Bryk; *Hierarchical linear models : applications and data analysis methods (2. ed., [3. Dr.] ed.)*, Thousand Oaks, CA [u.a.]: Sage Publications, 2002.

[5] B. G. Tabachnick, L. S. Fidell; *Using multivariate statistics (5th ed.)*, 2007.

[6] Discussione di Weisberg, su B. Efron, T. Hastie, I. Johnstone, R. Tibshirani; *"Least Angle Regression"*, nella rivista *"Annals of Statistics."*, 32 (2): pp. 407–499.

[7] S. Weisberg; *Applied Linear Regression*, Wiley, New York, 1980.

[8] "Documentazione ufficiale di Plotly-Dash", 2020.

[9] Apache Software Foundation, "Documentazione ufficiale di Apache Kafka", 2017.