



UNIVERSITÀ POLITECNICA DELLE MARCHE

FACOLTÀ DI INGEGNERIA

Corso di Laurea Triennale in Ingegneria Biomedica

**STUDIO E SVILUPPO DELLE CURVE DI
DISSOCIAZIONE DELL'EMOGLOBINA SU
NEONATI PRETERMINE: FILTRAGGIO DEI
DATI SPURI.**

**STUDY AND DEVELOPMENT OF
HEMOGLOBIN DISSOCIATION CURVE ON
PRETERM: FILTERING OF SPURIOUS DATA.**

Relatore:

Prof. Lorenzo Scalise

Corelatore:

Ing. Paolo Marchionni

Tesi di Laurea di:

Giulio Quaglia

Anno Accademico 2018/2019

Prima di procedere con la trattazione, vorrei dedicare qualche riga a tutti coloro che mi sono stati vicini in questo percorso di crescita personale e professionale.

Un sentito grazie al mio relatore Lorenzo Scalise per l'opportunità concessami e al mio tutor Paolo Marchionni che mi ha guidato dall'inizio alla fine di questo lavoro, senza la sua pazienza, competenza, genialità e amicizia tutto ciò non sarebbe stato possibile.

Grazie a mia mamma e all'amore della mia vita Giulia, per l'affetto e l'amore che spesso non merito e per sostenermi nei momenti più difficili.

Infine, grazie a Umberto per la sua amicizia e per aver diviso gioie e dolori di questa esperienza, nell'augurio che possa terminare presto nel migliore dei modi.

Grazie a tutti, senza di voi non ce l'avrei mai fatta.

Indice

Il metodo.....	7
Emoglobina.....	8
Struttura.....	8
Emoglobina nella respirazione.....	9
Curve saturazione.....	11
Parametri che influenzano le curve dissociazione.....	12
Emogasanalizzatore e emogasanalisi	15
Amperometriche.....	16
Potenziometriche	16
Sistema Ottico	17
Parametri	18
Parametri calcolati.....	20
Pre-Filtraggio.....	24
Metodo di Tukey per gli outliers	26
Funzione di predizione.....	32
Modello di regressione lineare semplice	32
Modello di regressione lineare multiple.....	34
Feature selection.....	37
Metodi di convalida	40
Bland-Altman plot.....	40
Coefficiente di Pearson	41
Durbin-Watson plot.....	42
Lilliefors Test	43
Studio incertezza con Montecarlo.....	45
Conclusioni	50
Bibliografia:.....	51

Il metodo

Il lavoro svolto durante il tirocinio presso il reparto di terapia intensiva neonatale dell'azienda ospedaliera "G. Salesi" è stato orientato allo studio e allo sviluppo di curve di dissociazione dell'emoglobina in neonati pretermine (bambini nati prima della 37^a settimana di gestazione). Grazie all'archivio ospedaliero in cui sono contenuti migliaia di file relative a misurazioni di emogasanalisi si è riusciti a stabilire una correlazione statistica tra alcuni parametri ematici e la pressione parziale dell'ossigeno, parametro fondamentale per la diagnostica in quanto dà indicazioni importanti sulla funzione respiratoria.

La prima fase è stata quella di acquisizione dati e filtraggio, che è quella in cui verrà posta maggiore attenzione in questo elaborato, mentre la seconda fase riguarda la feature selection e i metodi di convalida statistica.

Il lavoro complessivamente punta ad avere una stima della pressione parziale dell'ossigeno a partire da valori di saturazione del pulsossimetro, strumento che è in grado di riconoscere la quantità di ossigeno legato all'emoglobina sfruttando il diverso spettro di assorbimento, in maniera non invasiva.

Emoglobina

L'emoglobina(Hb) è una metalloproteina globulare con struttura quaternaria, la cui funzione principale è quella di legarsi all'ossigeno e trasportarlo nei vari distretti del corpo di quasi tutti i vertebrati.

Struttura

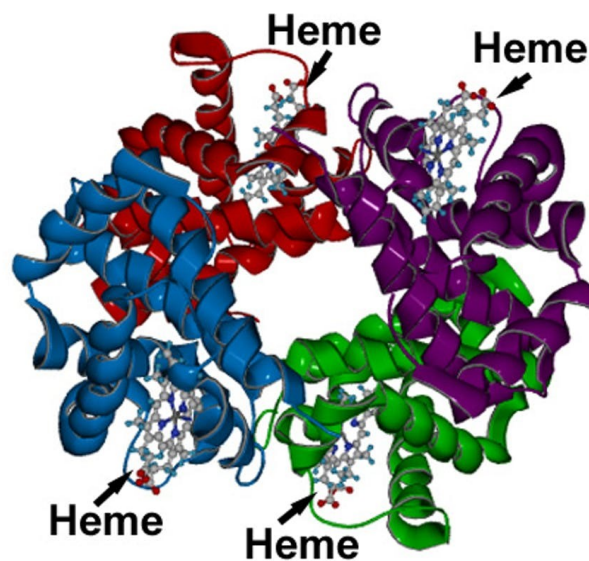


Figura 1 Struttura Emoglobina

L'Hb è formata da quattro subunità, le globine, a due a due identiche. Queste catene polipeptidiche vanno a loro volta a legarsi con il gruppo eme, un complesso costituito da uno ione Fe^{2+} (Fe^{3+}) con una molecola di protoporfirina responsabile del legame con l' O_2 .

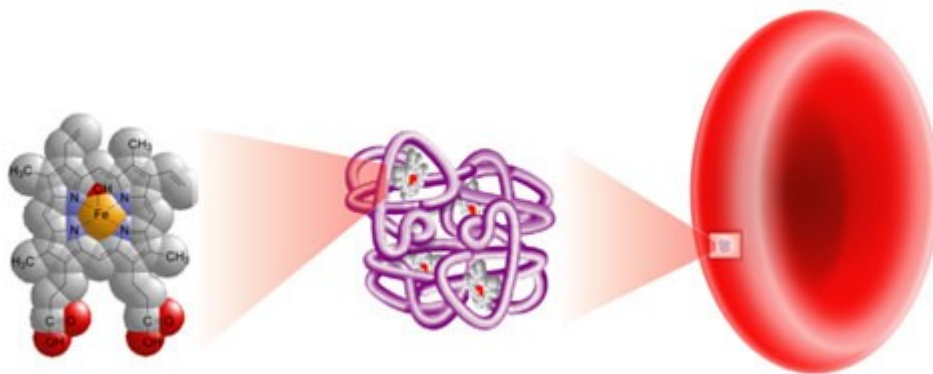


Figura 2 Eritrocita

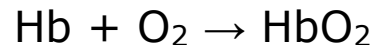
Le catene di globine nell'essere umano, determinate dall'espressione genica del polipeptide, sono principalmente quattro: α , β , γ , δ . La combinazione delle varie catene vanno a determinare diversi tipi di emoglobina con caratteristiche simili ma non identiche.

L'Hb nell'adulto HbA è costituita da due catene α e due catene β , rappresenta il 97% dell'emoglobina totale, chiamata anche HbA1, mentre due catene α e due catene δ , rappresentano la HbA2, circa il 2% del totale. Particolare attenzione nello studio in questione viene dato all'Hb costituita da due catene α e due catene γ , HbF o emoglobina fetale, presente nel feto e nel neonato fino a massimo 6 mesi.

Emoglobina nella respirazione

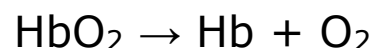
La quantità di O_2 che si scioglie nel sangue è circa il 2% (legge di Henry), troppo bassa per le esigenze metaboliche del corpo umano, per questo il ruolo dell'Hb è fondamentale. L'Hb può legare 1.34 mL di O_2 per grammo che aumenta di 70 volte la quantità di O_2 che è possibile trasportare con il sangue. Il responsabile del legame con l' O_2 è il gruppo eme, il cui ione di Fe passa da temporaneamente da 2+ a 3+:

Ossiemoglobina (1 legame O_2)



Per poi cedere l'atomo di O_2 e ritornare nello stato 2+

Deossiemoglobina



La formazione di HbO_2 avviene a livello dei capillari polmonari mentre il rilascio di O_2 avviene nei distretti periferici. Questo è possibile grazie alla variazione della pressione parziale dell' O_2 (P_{aO_2}) tra i settori periferici e quelli polmonari, infatti ad una alta P_{aO_2} corrisponde lo **stato R** dell'Hb detto rilassato, che aumenta l'affinità con O_2 , al contrario una bassa P_{aO_2} favorisce lo **stato T** detto teso, che diminuisce l'affinità e quindi dà luogo al rilascio di O_2 nei tessuti. Anche la pressione parziale della CO_2 (PCO_2) influenza lo stato dell'Hb in modo opposto a quella dell' O_2 .

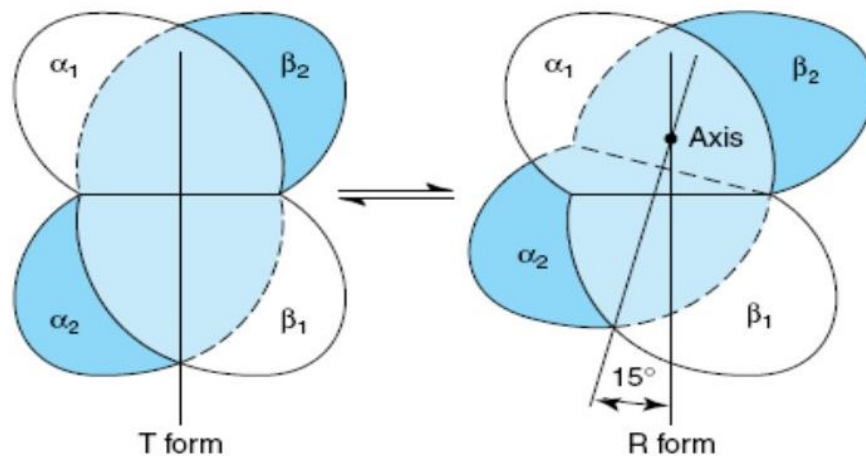


Figura 3 Forme dell'emoglobina

Esistono anche altre molecole che possono legare con l'O₂ come la mioglobina che possiede lo stesso gruppo eme dell'Hb, ma a differenza di questa risulta molto più stabile e quindi meno soggetta al rilascio di O₂, il ruolo è di accumulo di O₂ nei tessuti, mentre la Hb proprio perché molto più instabile riveste un ruolo fondamentale nell'ossigenazione e nella regolazione del pH a breve termine.

L'Hb può legarsi anche ad altri composti: carbodiossiemoglobina (CO₂), carbossiemoglobina (CO), cianoemoglobina (CN⁻), metaemoglobina (gruppo eme con stato ferrico (Fe³⁺) permanente che non permette il legame con l'O₂), solfoemoglobina (S). Può inoltre presentarsi in molte altre isoforme fisiologiche o patologiche (Hb di Portland, HbA, Hb Lepore, ...) che non verranno trattate.

Curve saturazione

La curva di saturazione dell'Hb mette in relazione la saturazione dell'O₂ (SO₂) cioè quale percentuale di O₂ rispetto al totale è legata all'Hb con la pressione parziale dell'O₂ (PO₂). La saturazione è massima quando tutta l'Hb è presente solo in forma di ossiemoglobina, al contrario la saturazione è minima quando è presente in forme diverse dall'ossiemoglobina, principalmente deossiemoglobina ma anche altre forme in piccolissima parte. Come si può notare dal grafico è una curva sigmoide infatti nel settore centrale per piccole variazioni di pO₂ si ottengono grandi variazioni di SO₂. La forma a sigmoide può essere vista anche come conseguenza globale dell'allosterismo dell'Hb, infatti l'affinità con l'O₂ aumenta per ogni gruppo eme che si lega, quindi una Hb con nessuno gruppo eme legato è meno affine rispetto ad una con un gruppo legato e a sua volta una con un gruppo legato è meno affine di una con due gruppi legati e così via.

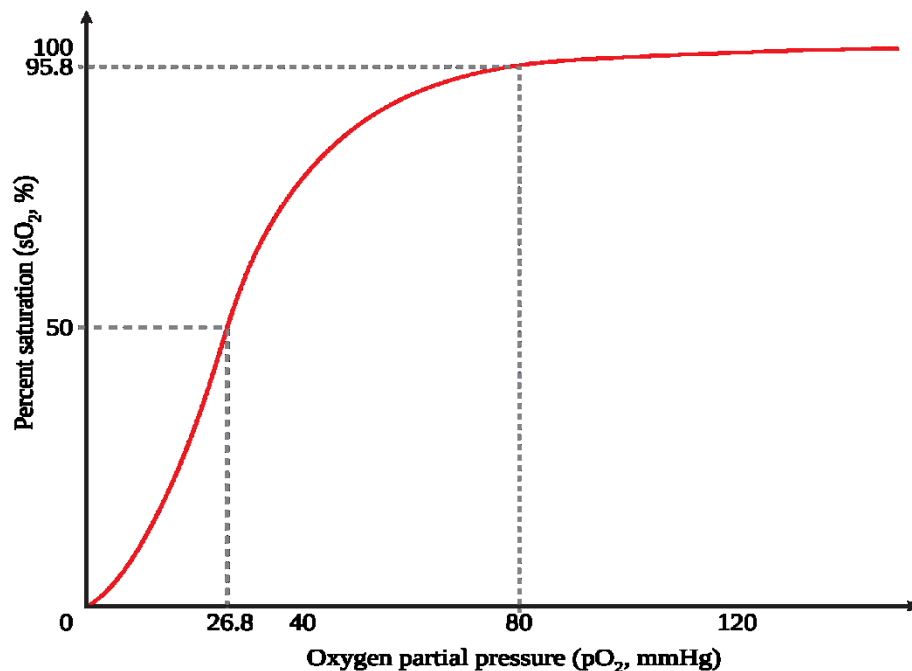


Figura 4 Curva Saturazione

Anche la mioglobina (Mb) riveste un ruolo fondamentale per il corpo dei vertebrati, ma come si evince dal grafico la pressione necessaria per raggiungere emisaturazione dei siti di legame è 1mmHg contro i 26.8 mmHg dell'Hb. La Mb quindi anche se ha un solo gruppo eme risulta molto più affine all'O₂ dell'Hb, principalmente dovuto al fatto che rivestono due ruoli complementari in cui quello della Mb è di accumulatore di O₂ nei tessuti.

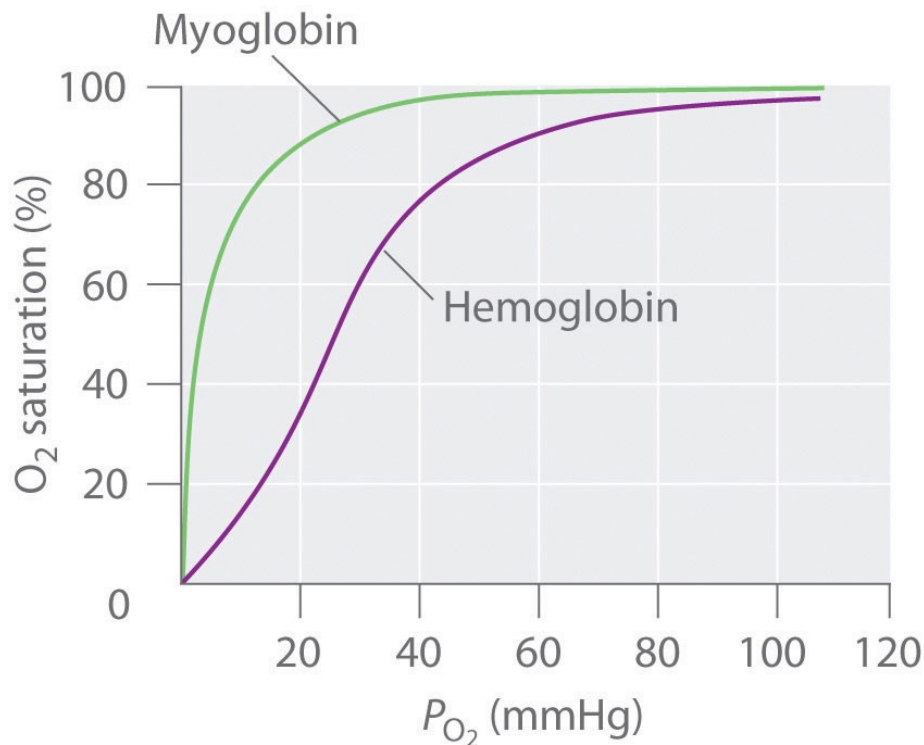


Figura 5 Mioglobina ed Emoglobina

Parametri che influenzano le curve dissociazione

Ci sono diversi fattori che possono influenzare la curva di saturazione dell'emoglobina tra i principali studiati in letteratura sono la temperatura, il 2,3-bisfosfoglicerato, la PCO₂ e il pH. Si studiano gli andamenti della curva di saturazione rispetto ai vari fattori che la influenzano tenendo costanti gli altri parametri:

- *2,3-DPG*, è un acido che risulta la base coniugata del 1,3-bisfosfoglicerico, si lega all'Hb stabilizzandola e impedendogli di legare con l'O₂, in particolare con le catene β. Mentre tende a legarsi (in maniera minore) con l'ossiemoglobina promuovendo il rilascio dell'O₂ per effetto allosterico. Condizioni di bassa ossigenazione tissutale (altitudine, ostruzioni vie aeree, insufficienza cardiaca) promuovono la generazione e l'accumulo di 2,3-DPG a causa dei cambiamenti nel pH e degli enzimi che lo degradano. D'altra parte, nelle gestanti l'accumulo di 2,3-DPG intracellulare diminuisce l'affinità dell'Hb materna che permette di trasportare maggior O₂ verso le arterie uterine e quindi verso il feto, proprio perché si lega con i gruppi β l'HbF non è sensibile alla molecola avendo al posto dei β due gruppi γ. In conclusione, nell'adulto maggiori livelli di 2,3-DPG determinano minore affinità dell'Hb, quindi uno shift verso destra della curva.

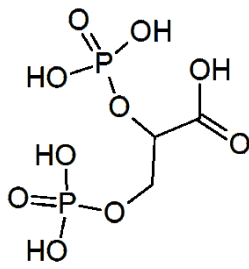


Figura 6 Struttura
2,3 DPG

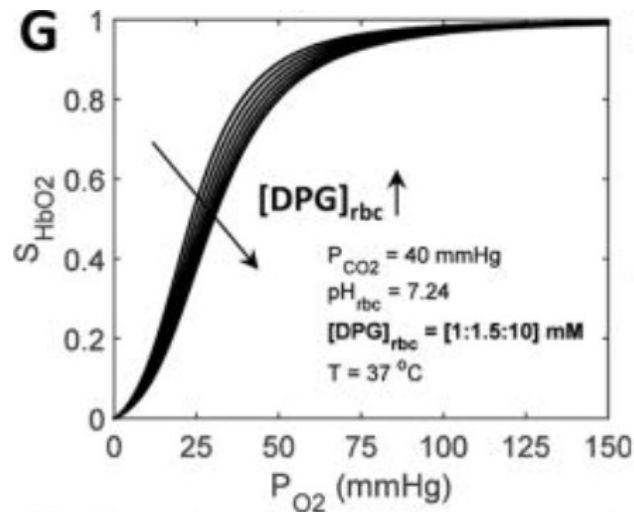


Figura 7 Influenza del 2,3 DPG

- *Temperatura*, se la temperatura incrementa (a parità di O₂ presenti) ha effetto sul legame dello ione Fe²⁺ con O₂ sfavorendo le condizioni per cui si verifichi, simultaneamente aumenta la pO₂. Quindi la curva avrà uno shift a destra dovuto all'aumento di pressione e uno shift verso il basso dovuto all'indebolimento del legame tra ferro e ossigeno.

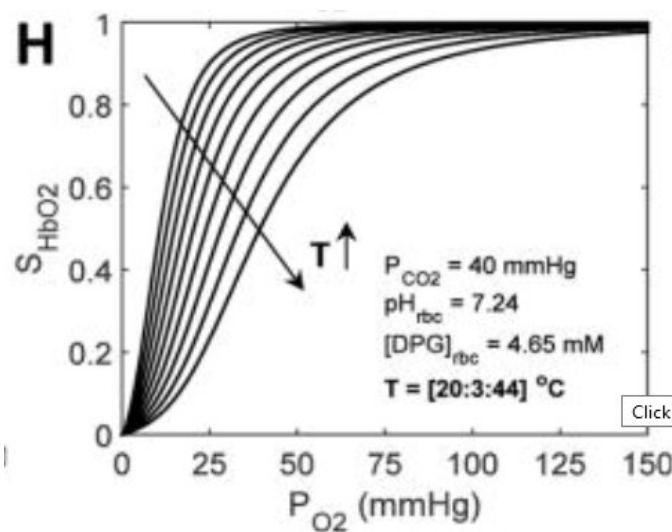


Figura 8 Influenza Temperatura

- *pH e Effetto Bohr*, la descrizione dell'influenza del pH sull'Hb passa attraverso la spiegazione dell'effetto Bohr, infatti nei tessuti durante azioni cataboliche delle sostanze nutritive viene prodotta CO₂ in forma di bicarbonato che genera ioni H⁺ (diminuzione del pH), la maggiore concentrazione di ioni idrogenioni va a protonare alcuni amminoacidi come l'istidina 146 che forma dei legami ionici (ponti salini) che stabilizzano l'Hb nella conformazione T che quindi va a rilasciare O₂ più facilmente. In conclusione, l'aumento del pH provoca uno shift della curva verso sinistra, e viceversa.

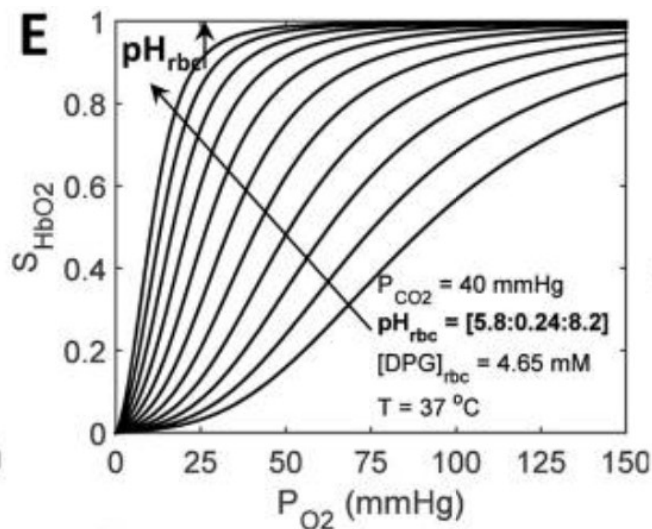


Figura 9 Influenza pH

- *pCO₂ ed effetto Haldane*, la CO₂ influisce in diversi modi sulla curva, non legando nei siti dell'O₂ e avendo un effetto allosterico quando si accumula CO₂ si forma carboaminoemoglobina e si ha un rilascio di O₂. L'effetto Haldane si verifica nella situazione opposta il rilascio di ossigeno dell'Hb aumenta la capacità di formazione di gruppi di carboaminoemoglobina. A livello polmonare l'Hb in ambiente più acido rilascia più facilmente CO₂ mentre a livello tissutale si avrà maggiore probabilità che si leghi. Il secondo effetto invece è collegato direttamente al pH infatti la formazione di bicarbonati rilascia ioni idrogenioni che favoriscono il rilascio di O₂, viceversa bassi livelli di carbonati alzano il pH che incrementano i legami con l'ossigeno. In conclusione, alti livelli di PCO₂ provocano uno shift verso destra della curva e viceversa.

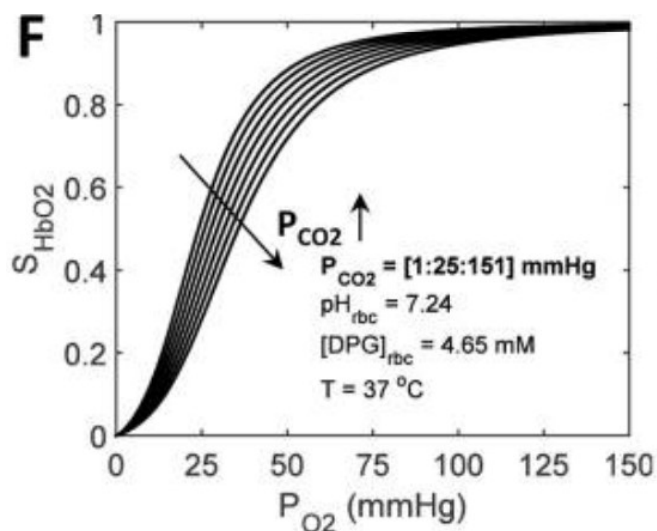


Figura 10 Influenza PCO2

Emogasanalizzatore e emogasanalisi



Figura 11 ABL 800

L'emogasanalizzatore è uno strumento che permette di misurare diversi parametri ematici: concentrazione di ioni idrogenioni (H^+) e quindi il pH, le pressioni parziali dei vari gas e concentrazioni di altre sostanze presenti nel sangue. L'emogasanalisi è molto importante a livello diagnostico perché è l'esame con il miglior rapporto costo/benefici, anche se invasivo.

L'esame con pochi μg di sangue arterioso o venoso permette di:

- controllare lo stato di ossigenazione del paziente, ossemia venosa o arteriosa dipendentemente dalla sede del prelievo
- variazioni di P_aO_2 come indici degli scambi gassosi e della ventilazione polmonare
- valutare eventuale ipossiemia
- informazioni per equilibrio idro-elettrolitico, osmolare, volemico, etc

Purtroppo, esclusa la taratura dello strumento ci sono altre cause di incertezza dell'esame che sono operatore dipendente.

Uno dei metodi di misurazione prevede l'impiego di elettrodi. Essi sono composti da:

- un contatto elettrico
- una parte colorata per distinguere il tipo di elettrodo
- un rivestimento esterno di protezione
- una soluzione elettrolitica tra il campione e l'elettrodo
- una membrana di separazione tra il campione e l'elettrodo sensibile alle variazioni delle sostanze in esame

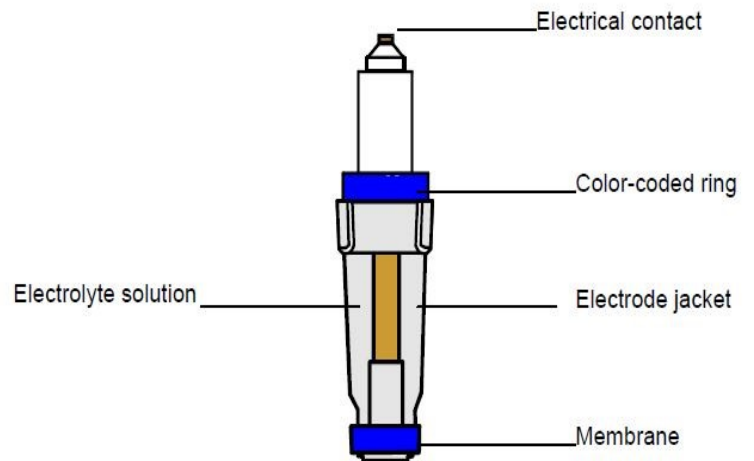


Figura 12 Elettrodo

Con gli elettrodi è possibile effettuare misurazioni di due tipi:

- Amperometriche
- Potenzimetriche

Amperometriche

Misura la corrente elettrica generata dalla riduzione sul catodo a tensione costante tramite un amperometro. La corrente è proporzionale alla concentrazione della sostanza in esame:

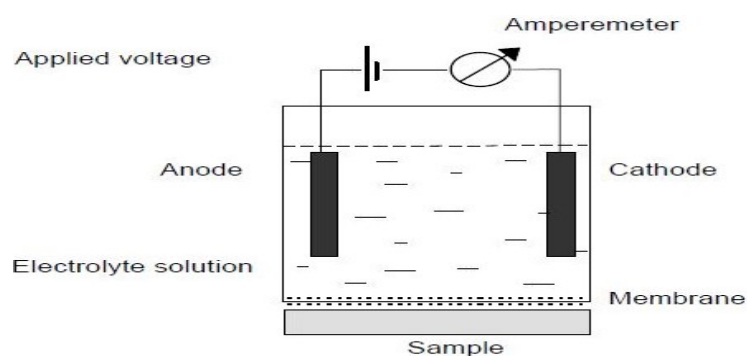


Figura 13 Metodo Amperometrico

Potenzimetriche

Misura la differenza di potenziale secondo lo schema in figura:

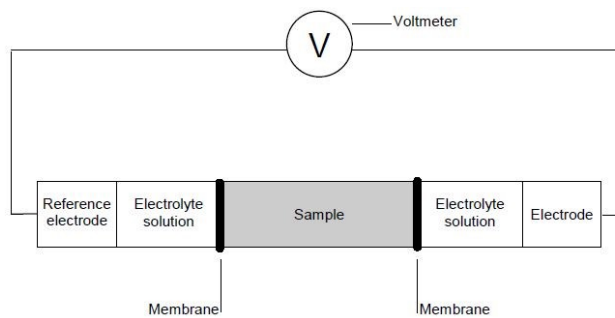


Figura 14 Metodo Potenzimetrico

L'elettrodo di riferimento è composto da una placca di argento in soluzione elettrolitica di formiato di sodio (HCOONa) e acido idrocloridrico (HCl) a pH 5,5. Per il calcolo si usa l'equazione di Nernst che permette di mettere in relazione il potenziale misurato con l'attività chimica della specie in esame che è proporzionale alla sua concentrazione.

$$E_{sample} = E_0 + \frac{2.3RT}{nF} \log a_x \quad a_x = \gamma c_x$$

Sistema Ottico

Il sensore per il sistema ottico è uno spettrofotometro a 128 lunghezze d'onda con un range di misura che va da 478 a 672nm, si tratta di uno spettroscopio elettromagnetico che indaga su lunghezze d'onda vicine allo spettro del visibile servendosi di un sensore, il fotometro, che determina l'intensità di luce e la mette in correlazione con la lunghezza d'onda. Infine

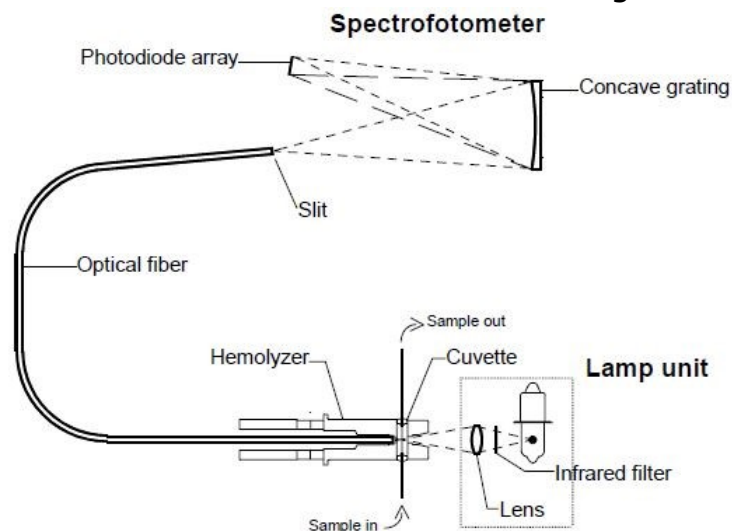


Figura 15 Metodo Ottico

collegato tramite fibra ottica c'è un emolizzatore e una cuvette in cui inserire il campione.

La quantità di sangue necessaria per l'analisi è molto piccola, 1µg, una volta inserito nella cuvette riscaldata a 37°C costanti l'emolizzatore agisce con una frequenza in ultrasuoni di circa 30k Hz e una pressione di 2 atm. A questo punto una luce costante di intensità e filtrata attraverso un filtro agli infrarossi, la luce viene riflessa su un set di 128 fotodiodi, uno per ogni lunghezza d'onda calcolabile dallo strumento. Ogni diodo converte la luce in una differenza di potenziale, l'insieme di tutte le ddp va a comporre lo spettro di assorbimento che poi viene analizzato al computer restituendo i parametri ossimetrici secondo la legge Lamber-Beer:

$$A_y^\lambda = \varepsilon_y^\lambda \cdot c_y \cdot l$$

l'assorbimento luminoso di un composto è proporzionale alla concentrazione del composto e alla lunghezza d'onda della luce che lo attraversa.

A = assorbimento luminoso

ε = coefficiente

c = concentrazione

l = lunghezza d'onda campione

λ = lunghezza d'onda

y = campione

dove:

$$A = \log \frac{I_0}{I}$$

I_0 = intensità della luce trasmessa attraverso acqua

I = intensità della luce trasmessa attraverso il campione.

Infine ci potrebbero essere delle sostanze che possono interferire con la misurazione perché hanno la stessa lunghezza d'onda del campione (intralipidi, solfoemoglobina) che devono essere compensate.

Parametri

I parametri possono essere misurati o calcolati: quelli misurati sono misurati direttamente tramite elettrodi o sistema ottico mentre quelli calcolati sono calcolati con equazioni basate sui parametri misurati.

Parametro	Descrizione	Metodo	Range/unità di misura di misura
Pressione Atmosferica	Effetto della pressione barometrica sui campioni	barometro integrato	Metri slm /mmHg
pH	Indica l'acidità o l'alcalinità del campione	potenziometrico	da 6.3 a 8.0 pH
cH ⁺	Indica la concentrazione di ioni idrogeno	potenziometrico	da 10.0 a 199,9 nmol/L
pCO ₂	E' la pressione parziale del diossido di carbonio	potenziometrico	da 0,67 a 33,3 kPa
ctHb	Indica la concentrazione di emoglobina totale	ottico	da 0,00 a 17,2 nmol/L
FO ₂ Hb	Indica la frazione di ossiemoglobina sull'emoglobina totale	ottico	da 0 a 100%
FCOHb	Indica la frazione di carbossiemoglobina nell'emoglobina totale	ottico	da 0 a 100%
RHb	Indica la frazione di deossiemoglobina sull'emoglobina totale	ottico	da 0 a 100%
FMetHb	Indica la frazione di metemoglobina nell'emoglobina totale nel sangue.	ottico	da 0 a 100%
sO ₂	Indica la saturazione d'ossigeno, ovvero il rapporto tra la concentrazione di ossiemoglobina e l'emoglobina meno le disemoglobine.	ottico	da 0 a 100%
HbF	Indica la frazione di emoglobina fetale nell'emoglobina totale	ottico	da 0 a 100%
pO ₂	Indica la pressione parziale di ossigeno nel sangue, valori alti o bassi indicano rispettivamente iperossia e ipossia.	amperometrico	da 0,0 a 107 kPa
tBil	Indica la concentrazione di bilirubina totale nel plasma, che comprende la sua forma coniugata e non coniugata	ottico	da 0 a 585mg/L
K ⁺	Indica la concentrazione di ioni potassio nel plasma.	potenziometrico	da 0,5 a 25,0mmol/L.
Na ⁺	Indica la concentrazione di ioni sodio nel plasma.	potenziometrico	da 7 a 350 mmol/L
Ca ²⁺	Indica la concentrazione di ioni calcio nel plasma.	potenziometrico	da 0,20 a 9,99 mmol/L

Cl ⁻	Indica la concentrazione di ioni cloruro nel plasma.	potenziometrico	da 7 a 350 mmol/L
Glu	Indica la concentrazione di D-glucosio nel plasma	amperometrico	da 0,0 a 60 mmol/L
Lac	Indica la concentrazione di L-lattato nel plasma.	amperometrico	da 0,0 a 30 mmol/L

Tabella 1

Parametri calcolati

pH(T): pH del sangue alla temperatura del paziente.

cH⁺(T): Concentrazione degli ioni idrogeno nel sangue alla temperatura del paziente.

pCO₂(T): Pressione parziale del biossido di carbonio alla temperatura del paziente.

cHCO₃⁻(P): Concentrazione del carbonato di idrogeno nel plasma.

cBase(B) o ABE: Eccesso Attuale di Base, ovvero la concentrazione di base titolabile quando il sangue è titolato con una base forte o un acido, con un pH del plasma di 7,40, ad una pCO₂ di 5,33 kPa e 37°, all'attuale saturazione di ossigeno. Valori positivi indicano un relativo deficit di acidi nanocarbonici mentre valori negativi segnalano un relativo eccesso di acidi non carbonici.

cBase(B,ox): La cBase(B) del sangue pienamente ossigenato.

cBase(Ecf) o SBE: Eccesso Standard di Base, ovvero un'espressione "in vivo" dell'eccesso di base. Si riferisce ad un modello del fluido extracellulare ed è calcolato usando un valore standard della concentrazione di emoglobina del fluido extracellulare totale.

cBase(Ecf,ox): cBase(Ecf) del sangue pienamente ossigenato.

cHCO₃⁻(P,st): "Bicarbonato Standard", ovvero la concentrazione del carbonato di idrogeno nel plasma del sangue che è stato equilibrato con una mistura di gas con pCO₂=5,33kPa e PO₂≥13,33kPa a 37°C.

ctCO₂(P): Concentrazione del biossido di carbonio totale nel plasma.

ctCO₂(B): Concentrazione del biossido di carbonio totale nel sangue. Calcolato basato sulla concentrazione del CO₂ totale nelle due fasi: plasma e fluido eritrocitario.

pH(st): pH standard, definito come il pH del plasma del sangue equilibrato a $p\text{CO}_2=5,33\text{kPa}$.

Un valore costante di $p\text{CO}_2$ rimuove infatti l'influenza respiratoria sul pH, che quindi rifletterà lo stato metabolico del plasma del sangue.

VCO_2/V (aria secca): La frazione volumetrica del biossido di carbonio nell'aria secca.

Hct: Indica l'ematocrito, ovvero il rapporto tra il volume di eritrociti e il volume di sangue intero.

$p\text{O}_2(\text{T})$: Pressione parziale di ossigeno alla temperatura del paziente.

$p\text{O}_2(\text{A})$: Pressione parziale di ossigeno nell'aria alveolare.

$p\text{O}_2(\text{A},\text{T})$: Pressione parziale di ossigeno nell'aria alveolare alla temperatura del paziente.

$p\text{O}_2(\text{a})/\text{FO}_2(\text{i})$: Rapporto tra la pressione parziale d'ossigeno del sangue arterioso e la pressione parziale d'ossigeno nell'aria inspirata.

$p\text{O}_2(\text{a},\text{T})/\text{FO}_2(\text{i})$: Rapporto tra la pressione parziale d'ossigeno del sangue arterioso alla temperatura del paziente e la pressione parziale d'ossigeno nell'aria inspirata.

P50: Pressione parziale d'ossigeno a metà saturazione nel sangue. Valori alti e bassi denotano rispettivamente una ridotta e aumentata affinità dell'ossigeno all'emoglobina.

P50(T): Pressione parziale d'ossigeno a metà saturazione nel sangue a temperatura paziente.

P50(st): Pressione parziale d'ossigeno a metà saturazione nel sangue a condizioni standard:

- $T=37^\circ\text{C}$
- $\text{pH}=7,40$
- $p\text{CO}_2=5,33\text{kPa}$
- $\text{FCOHb}, \text{FMetHb}, \text{FHbF} = 0$

$p\text{O}_2(\text{A-a})$: Differenza nelle pressioni parziali d'ossigeno nell'aria alveolare e nel sangue arterioso. Indica l'efficacia del processo di ossigenazione nei polmoni.

$p\text{O}_2(\text{A-a},\text{T})$: Differenza nelle pressioni parziali d'ossigeno nell'aria alveolare e nel sangue arterioso a temperatura paziente.

$p\text{O}_2(\text{a/A})$: Rapporto tra la pressione parziale dell'ossigeno nel sangue arterioso e nell'aria negli alveoli. Indica l'efficacia del processo di ossigenazione nei polmoni.

pO₂(a/A,T): Rapporto tra la pressione parziale dell'ossigeno nel sangue arterioso e nell'aria negli alveoli alla temperatura del paziente.

pO₂(x): Pressione parziale d'estrazione d'ossigeno del sangue arterioso. Riflette gli effetti integrati dei cambiamenti di pO₂(a), ctO₂, and p50 sull'abilità del sangue arterioso di rilasciare ossigeno ai tessuti.

pO₂(x,T): Pressione parziale d'estrazione d'ossigeno del sangue arterioso a temperatura paziente.

ctO₂(B): Concentrazione totale d'ossigeno nel sangue.

ctO₂(a-v): Differenza tra la concentrazione d'ossigeno del sangue arterioso e venoso misto.

BO₂ : Capacità d'ossigeno dell'emoglobina; la massima concentrazione di ossigeno legato all'emoglobina nel sangue saturato, in modo che tutta la deossiemoglobina è convertita in ossiemoglobina.

ctO₂(x): Concentrazione dell'ossigeno estraibile dal sangue arterioso. Definito come la quantità di ossigeno che può essere estratta per litro di sangue arterioso con una pressione parziale di ossigeno di 5,0kPa, mantenendo costanti pH e pCO₂.

DO₂ : Distribuzione d'ossigeno; la quantità totale di ossigeno distribuita a tutto l'organismo per unità di tempo.

Q_t : Output cardiaco; il volume di sangue in uscita dal ventricolo sinistro per unità di tempo.

VO₂ : Consumo di ossigeno; quantità totale di ossigeno utilizzata da tutto l'organismo per unità di tempo.

FShunt: Shunt fisiologico relativo o shunt basato sulla concentrazione.

Calcolato dall'equazione dello shunt polmonare se viene usato sia sangue arterioso che venoso misto.

Può essere stimato da un campione arterioso assumendo una differenza costante tra le concentrazioni di ossigeno totale nel sangue arterioso e venoso misto: ctO₂ (a – v) = 2.3 mmol / L (5.1 mL / dL)

FShunt(T): FShunt a temperatura paziente.

RI: Indice Respiratorio; rapporto fra la differenza di tensione d'ossigeno tra aria alveolare e sangue arterioso e la tensione d'ossigeno nel sangue arterioso.

RI(T): Indice Respiratorio; rapporto fra la differenza di tensione d'ossigeno tra aria alveolare e sangue arterioso e la tensione d'ossigeno nel sangue arterioso a temperatura paziente.

$VO_2/V(\text{aria secca})$: Rapporto volumetrico dell'ossigeno nell'aria secca.

Q_x : Fattore Cardiaco di compensazione dell'ossigeno del sangue arterioso, definito come il fattore con il quale l'output cardiaco dovrebbe aumentare per permettere un rilascio di 2,3mmol/L di ossigeno ad una pO_2 venosa mista pari a 5,0kPa.

$V(B)$: Volume di sangue, calcolato quando sono immessi i valori di FCOHb e $V(CO)$.

$cCa^{2+}(7,40)$: Concentrazione di ioni calcio nel plasma a pH 7,40.

Anion Gap, K^+ : Differenza di concentrazione tra $cK^+ + cNa^+$ e $cCl^- + cHCO_3^-$.

Anion Gap: Differenza di concentrazione tra cNa^+ e $cCl^- + cHCO_3^-$.

mOsm: Osmolalità del plasma.

Pre-Filtraggio

Dopo l'estrazione iniziale dal database i dati si presentano così:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
	Paziente	Codice	id	Analizzatore	DataAnalisi	Tipo	P50_st_MmHg	RQ	O2HB	BMmHg	cBase_B_MmolL	PH	PCO2MmHg	PO2MmHg	HCl
10	'maschio '	'09864311135'	98153	'ABL800 AC 1'	2012-10-16 10:02:00	Non Specific...	26.8500	0.8600	48.9000	753	-4.8000	7.2690	49.8000	23.5000	
11	'merav9gli...	'0131109'	116579	'ABL800 AC 1'	2011-09-16 08:13:00	Capillare	26.8500	0.8600	89.5000	755	-4.4000	7.4170	29.8000	50.4000	
12	'merav9gli...	'0131109'	117051	'ABL800 AC 1'	2011-09-16 08:13:00	Capillare	26.8500	0.8600	89.5000	755	-4.4000	7.4170	29.8000	50.4000	
13	'francesco '	'cesaretti'	62116	'ABL800 3'	2014-08-03 16:28:00	Venoso	NaN	NaN	1.4000	NaN	NaN	7.4170	37.4000	49	
14	'ricci rebec...	'01310082601'	131109	'ABL800 AC 1'	2010-08-28 07:10:00	Capillare	26.8500	0.8600	88	750	0.3000	7.4250	36.9000	44.8000	
15	'ricci rebec...	'01310082601'	131559	'ABL800 AC 1'	2010-08-28 07:10:00	Capillare	26.8500	0.8600	88	750	0.3000	7.4250	36.9000	44.8000	
16	'..... '	'01223654778'	152483	'ABL800 AC 1'	2009-01-11 12:21:00	Arterioso	26.8500	0.8600	89.2000	767	-3.6000	7.3100	45.5000	49.5000	
17	'..... '	'01223654778'	152911	'ABL800 AC 1'	2009-01-11 12:21:00	Arterioso	26.8500	0.8600	89.2000	767	-3.6000	7.3100	45.5000	49.5000	
18	'lilla '	'01310081702'	131198	'ABL800 AC 1'	2010-08-20 09:43:00	Capillare	26.8500	0.8600	94.3000	761	-2.3000	7.4610	27.3000	67.8000	
19	'lilla '	'01310081702'	131648	'ABL800 AC 1'	2010-08-20 09:43:00	Capillare	26.8500	0.8600	94.3000	761	-2.3000	7.4610	27.3000	67.8000	
20	'0 0'	'0114100522...	138701	'ABL800 AC 1'	2010-01-20 15:18:00	Capillare	26.8500	0.8600	90.7000	757	-10.4000	7.2300	41.2000	50.3000	
21	'0 0'	'0114100522...	139142	'ABL800 AC 1'	2010-01-20 15:18:00	Capillare	26.8500	0.8600	90.7000	757	-10.4000	7.2300	41.2000	50.3000	
22	'0 maxwell'	'0130910050...	142235	'ABL800 AC 1'	2009-10-11 08:02:00	Capillare	26.8500	0.8600	89	754	-2.8000	7.3160	46.9000	47.9000	
23	'0 maxwell'	'0130910050...	142262	'ABL800 AC 1'	2009-10-10 08:54:00	Capillare	26.8500	0.8600	90.1000	756	-1.5000	7.3490	44.3000	50.1000	
24	'0 maxwell'	'0130910050...	142662	'ABL800 AC 1'	2009-10-11 08:02:00	Capillare	26.8500	0.8600	89	754	-2.8000	7.3160	46.9000	47.9000	
25	'0 maxwell'	'0130910050...	142689	'ABL800 AC 1'	2009-10-10 08:54:00	Capillare	26.8500	0.8600	90.1000	756	-1.5000	7.3490	44.3000	50.1000	
26	'0 maxwell'	'0130910050...	142705	'ABL800 AC 1'	2009-10-09 12:02:00	Capillare	26.8500	0.8600	85.6000	758	-0.5000	7.3680	43.6000	39.9000	
27	'0 maxwell'	'0130910050...	142734	'ABL800 AC 1'	2009-10-07 06:46:00	Capillare	26.8500	0.8600	92.6000	761	-4.2000	7.3550	37.3000	52	
28	'0 maxwell'	'0130910050...	143122	'ABL800 AC 1'	2009-10-09 12:02:00	Capillare	26.8500	0.8600	85.6000	758	-0.5000	7.3680	43.6000	39.9000	
29	'0 maxwell'	'0130910050...	143151	'ABL800 AC 1'	2009-10-07 06:46:00	Capillare	26.8500	0.8600	92.6000	761	-4.2000	7.3550	37.3000	52	
30	'0 piccola'	'0131401090'	84259	'ABL800 AC 1'	2014-02-13 11:46:00	Venoso	26.8500	0.8600	60.8000	755	-1.1000	7.3970	38	23.2000	
31	'0 0'	'1141005220...	62029	'ABL800 3'	2010-01-20 15:18:00	Venoso	NaN	NaN	NaN	NaN	NaN	7.2300	41.2000	50.3000	
32	'0 maxwell'	'13091005040'	61610	'ABL800 3'	2009-10-10 08:54:00	Venoso	NaN	NaN	NaN	NaN	NaN	7.3490	44.3000	50.1000	
33	'0123 0123'	'01234567899'	87233	'ABL800 AC 1'	2013-11-17 11:34:00	Scalpo Fetale	26.8500	0.8600	61.1000	763	-6.4000	7.2850	42	33.6000	

Figura 16 Dataset

Un array bidimensionale di 126635 righe con più di 40 colonne diverse relative sia ai valori biometrici dell'EGA sia a parametri identificativi del paziente, del prelievo, dell'orario e et simili.

La prima selezione sui dati ha riguardato i dati anagrafici, data e ora e nome analizzatore, visto che sono stati presi dallo stesso apparecchio. Si è scelto di fare uno studio sui dati relativi a prelievi di tipo arterioso poiché di maggiore interesse per la PO₂, ma si sarebbe potuto fare lo stesso con i dati derivanti da prelievo venoso.

La tabella evidenzia alcuni parametri che hanno come risultato NaN, not a number, il calcolatore non riconosce il dato come un numero, quindi sono degli errori di misura, questi vengono subito selezionati e scartati.

Sono stati eliminati alcuni valori irrealistici:

- Valori di saturazione inferiori a 60%, già valori inferiori all'80% indicano una grave crisi ipossica in atto
- Valori di emoglobina fetale inferiori a zero e negativi, per ovvi motivi
- Valori di bilirubina inferiori a zero e negativi, per ovvi motivi
- Valori di PO₂ tra 20mmHg e 200mmHg per una rappresentazione più chiara

Dopo questa fase iniziale di filtraggio si rappresentano i valori di saturazione e pressione parziale:

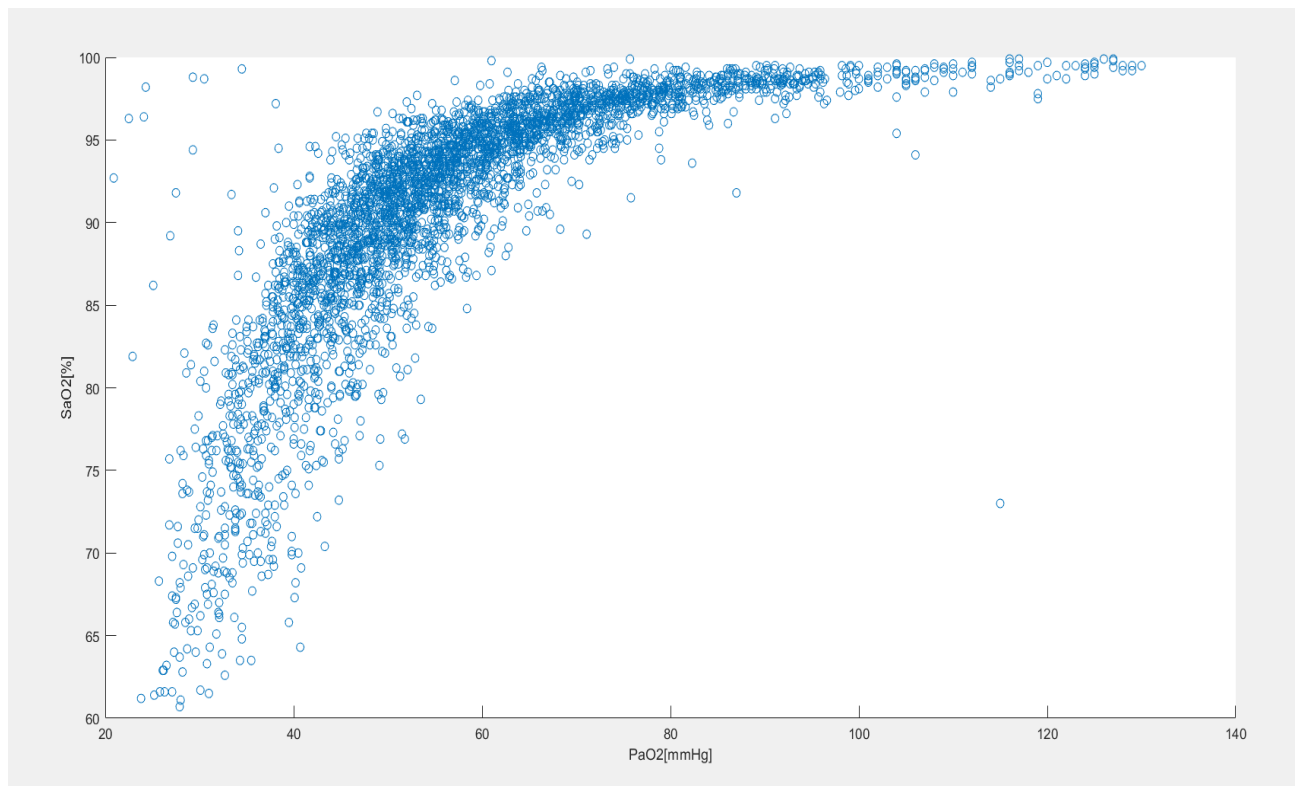


Figura 17 Plot dati PaO2 SaO2

La curva caratteristica di dissociazione dell'emoglobina è quasi rispettata, infatti la forma è intermedia tra la sigmoide dell'emoglobina e l'iperbole della mioglobina indicante un'affinità con l'ossigeno intermedia alle due molecole; questo non ci sorprende infatti come descritto nei capitoli precedenti l'emoglobina fetale con due catene γ ha più affinità con l'ossigeno dell'emoglobina $2\alpha 2\beta$ e quindi la forma è nei limiti dei risultati teorici.

Nella distribuzione si evidenziano purtroppo anche dei valori che vanno di molto fuori dal range della curva teorica, questi saranno l'oggetto del prossimo capitolo.

Metodo di Tukey per gli outliers

I valori cerchiati in rosso sono valori anomali che prendono il nome di outliers.

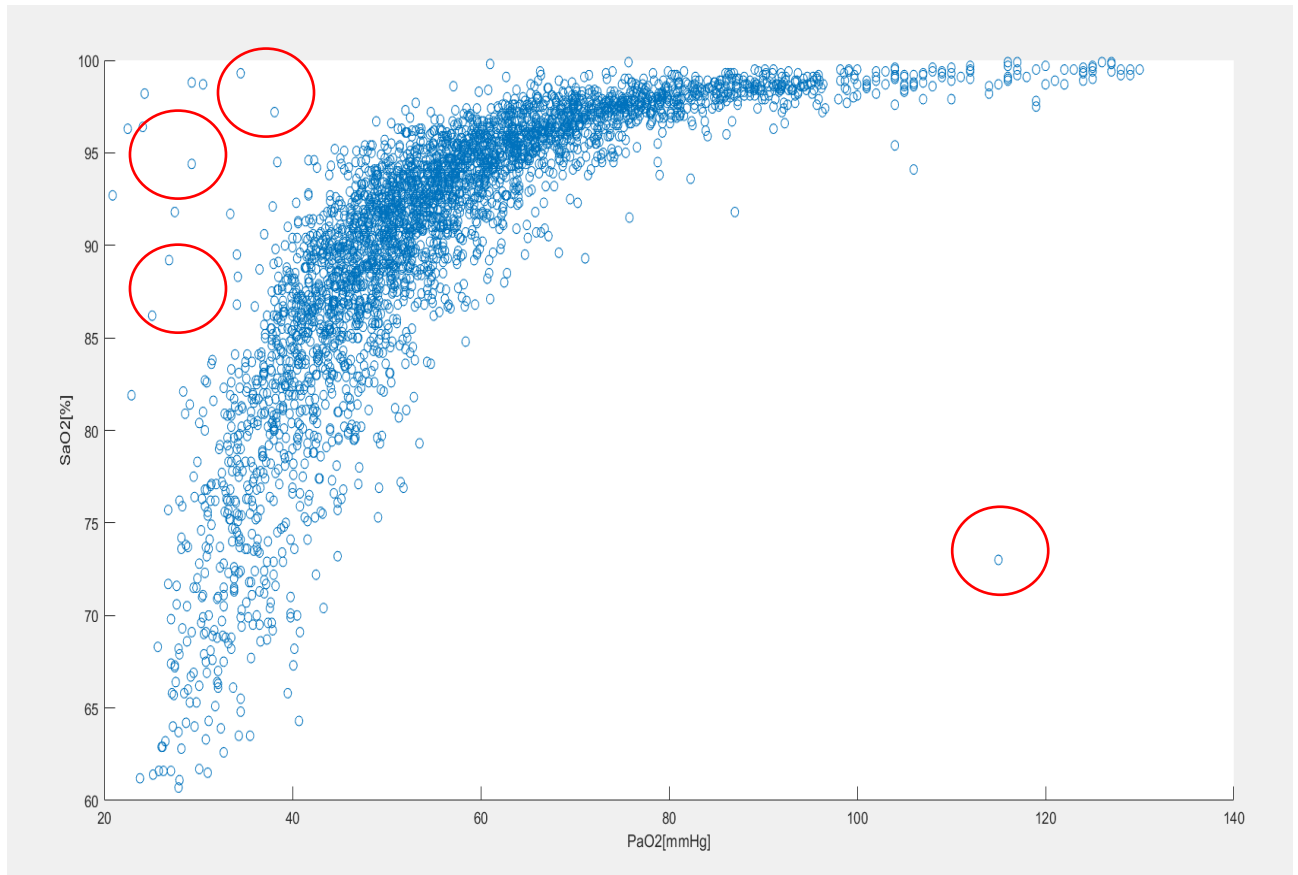


Figura 18 Plot dati - outliers

Possono verificarsi in ogni distribuzione e sono dovuti ad un errore sperimentale o ad un errore di misura, in ogni caso la presenza di uno o diversi outliers comporta gravi problemi nell'analisi statistica. L'approccio è di scartarli o usare metodi statistici che sono resistenti agli outliers, "robusti". Il metodo statistico utilizzato in questo studio risulta abbastanza sensibile alla presenza di outliers quindi si cercherà di eliminarli.

Per illustrare come gli outliers influenzano negativamente analisi statistiche si farà un semplice esempio:
si supponga di voler calcolare l'altezza media di 10 persone, 9 misurazioni sono tra 1.60 m e 1.90 m mentre c'è una misura che risulta 3.5 m, la mediana risulterà tra 1.60 e 1.90 ma la media sarà oltre 2 m.

In base alla distribuzione che si prende in considerazione si può stimare la quantità di outliers che ci dovrebbero essere, il caso più comune è quello della distribuzione gaussiana data da:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- con μ valore atteso stimato dalla media con N il numero di osservazioni

$$\mu \approx \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

- con σ la deviazione standard del campione corretta

$$\sigma \approx \sigma_x = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}}$$

La forma classica della distribuzione è questa:

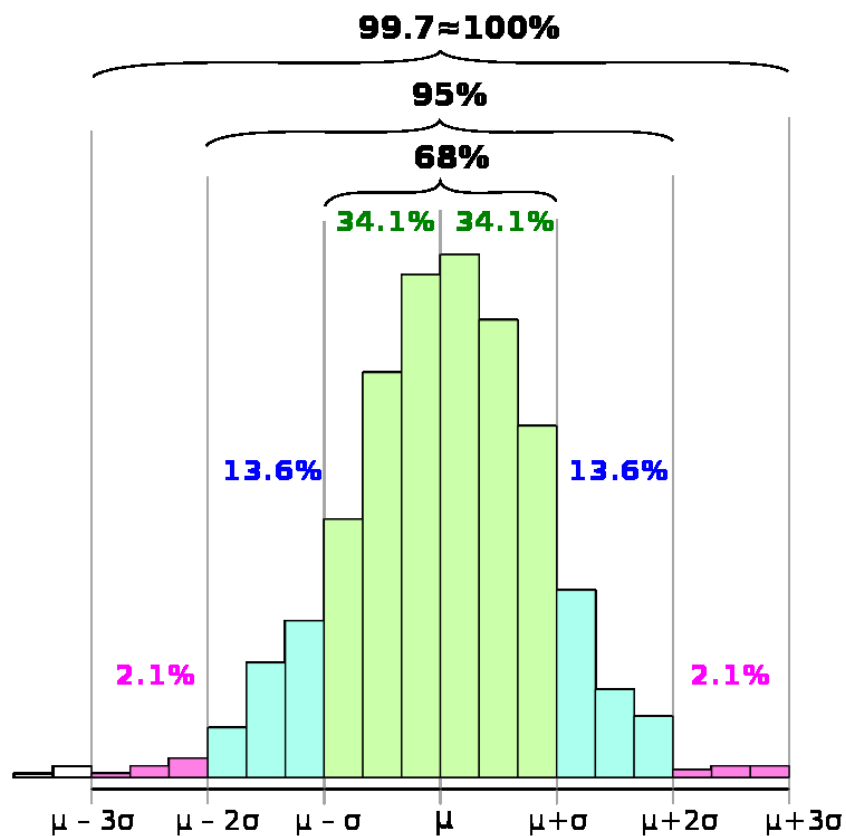


Figura 19 Distribuzione Normale

Quindi la regola del tre sigma dice che la quasi totalità delle osservazioni ricadranno entro l'intervallo $\mu - 3\sigma$:

- il 68% entro la prima deviazione standard
- il 95% entro la seconda deviazione standard
- il 99,7% entro la terza deviazione standard

Da un set di 1000 osservazioni ci aspettiamo che 3 osservazioni cadano fuori dalla terza deviazione standard, da un set di 126635 osservazioni circa 380 osservazioni potrebbero cadere fuori dalla terza deviazione standard. Il caso in esame purtroppo non presenta solo distribuzioni normali quindi potrebbero essere anche di più.

Ci sono diversi approcci per l'eliminazione degli outliers e proprio perché non c'è una vera e propria definizione matematica per l'outlier possono essere sia grafici che algebrici che ibridi, il metodo trattato usa la visualizzazione tramite box plot.

Il boxplot o diagramma a baffi è un metodo di visualizzazione di un dataset attraverso 5 numeri riassuntivi: massimo e minimo valore, mediana (50esimo percentile), primo quartile (25esimo percentile) e terzo quartile (75esimo percentile).

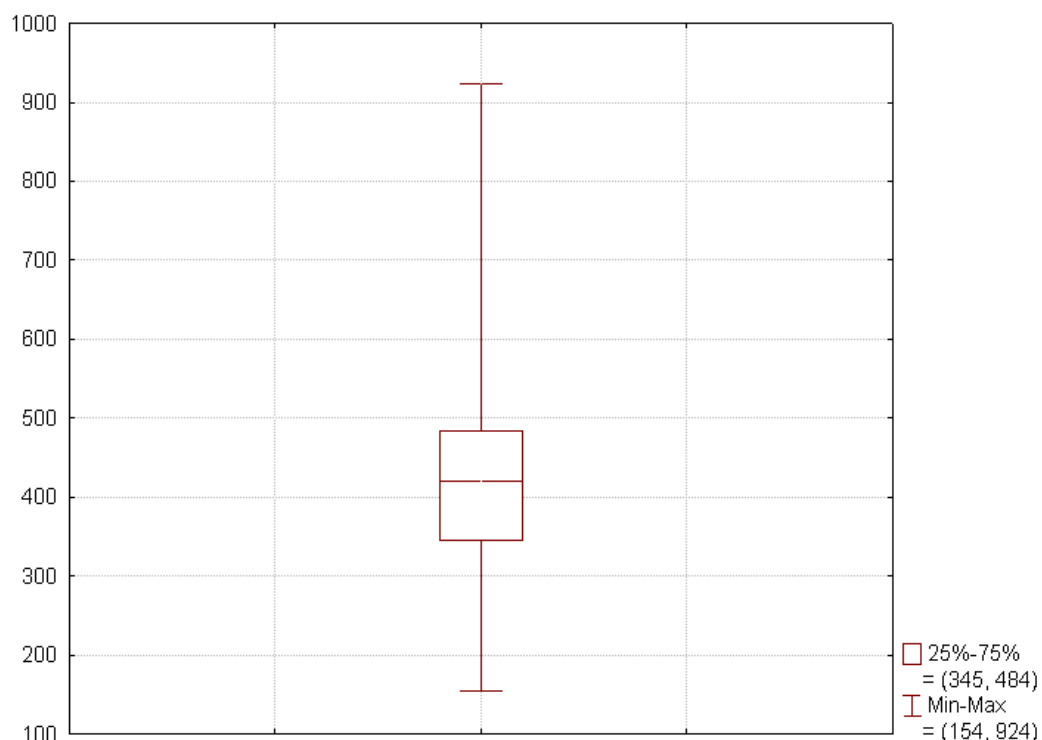


Figura 10 Boxplot

Il plot è composto da un rettangolo che costituisce il IQR o range interquartilico definito semplicemente come $Q_3 - Q_1$, la parte orizzontale nel mezzo del box rappresenta la mediana; la parte inferiore e superiore con una linea verticale sono i cosiddetti baffi, cioè la distanza tra il quartile e il massimo o minimo.

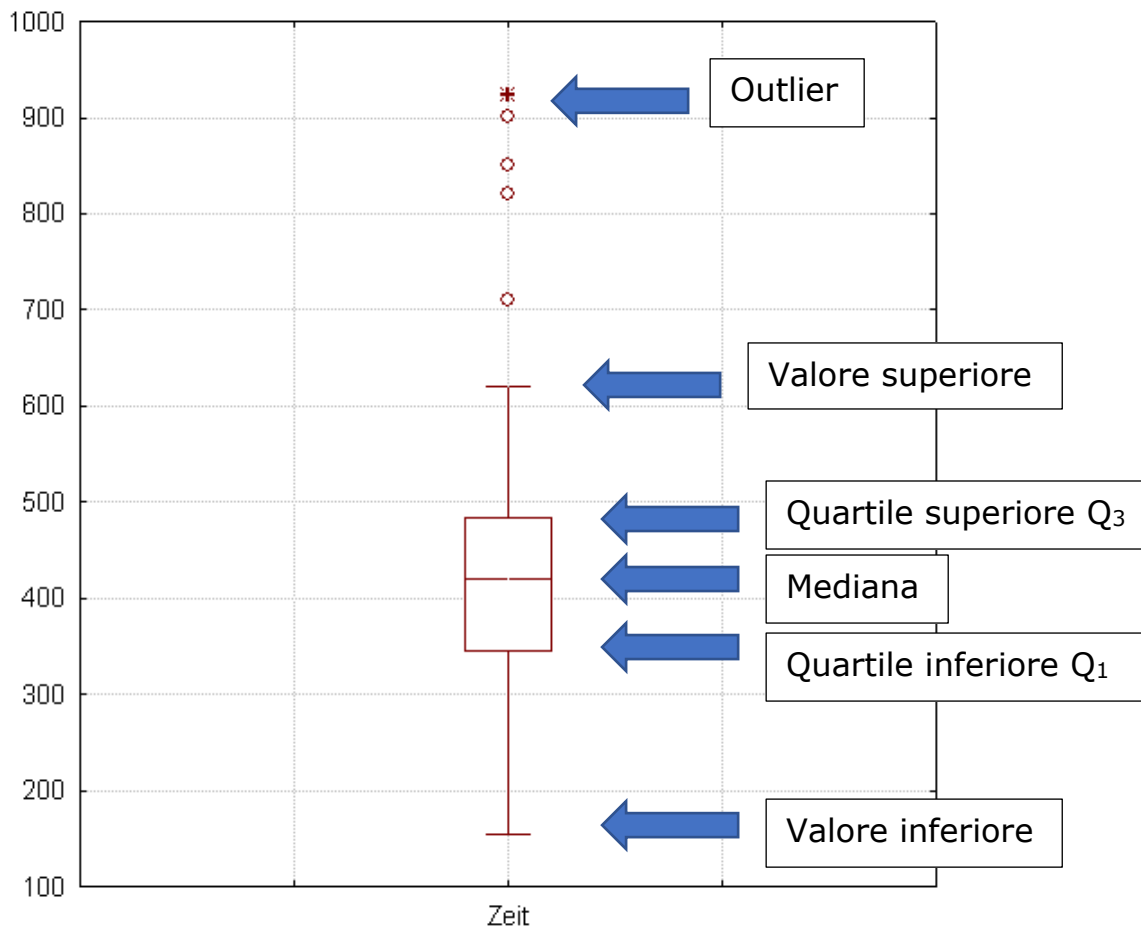


Figura 11 Boxplot Tukey

Secondo il metodo di Tukey la stessa rappresentazione può essere fatta prendendo Q_1 come $Q_1 - 1.5 \cdot IQR$ e Q_3 come $Q_3 + 1.5 \cdot IQR$. Gli outliers vengono plottati come pallini, croci o stelle.

Le barriere interne stanno alla distanza di $1.5 \cdot IQR$ sotto Q_1 e sopra Q_3 mentre le barriere esterne stanno $3 \cdot IQR$ sotto Q_1 e sopra Q_3 . I valori tra le barriere interne e quelle esterne sono possibili outliers mentre quelli oltre le barriere esterne sono probabili outlier.

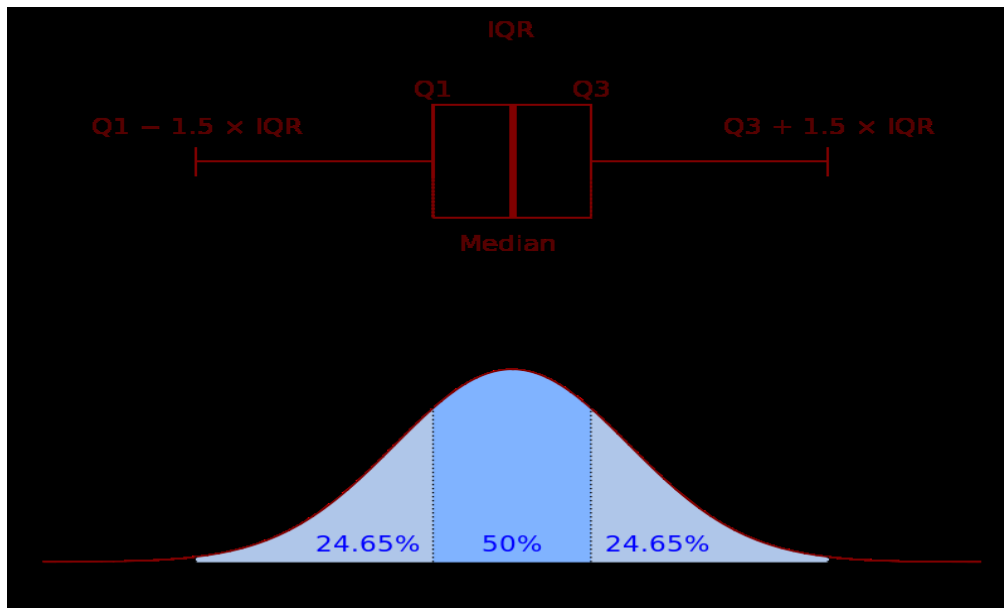


Figura 22 Tukey con Distribuzione Normale

Si evidenzia come il metodo di Tukey si rapporta ad una distribuzione normale:

Nel caso in esame si illustra il boxplot del pH prima e dopo l'applicazione dell'algoritmo di tukey:

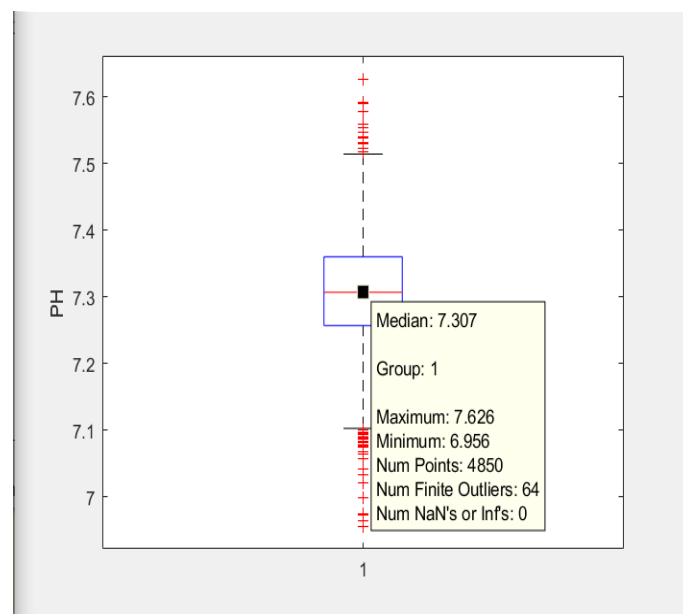
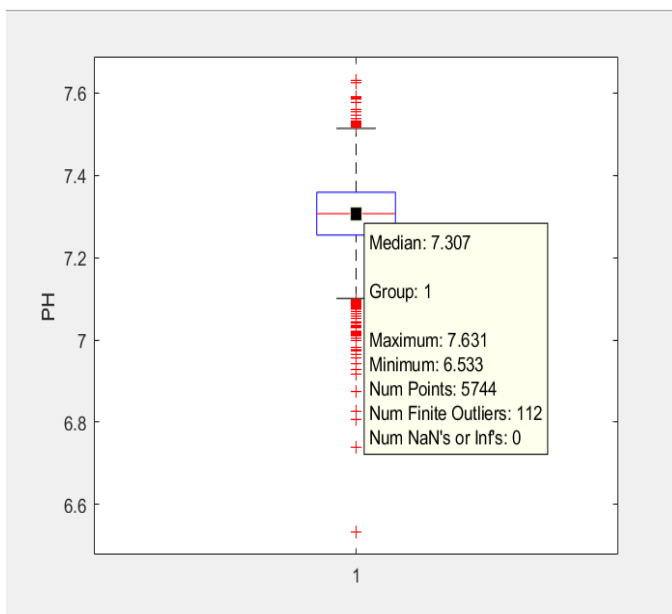


Figure 12a-13b Boxplot pH

Per i valori del pH il numero degli outliers passa da 112 a 64. Dopo aver applicato a tutto il dataset l'algoritmo andiamo a vedere come si modifica la curva di dissociazione.

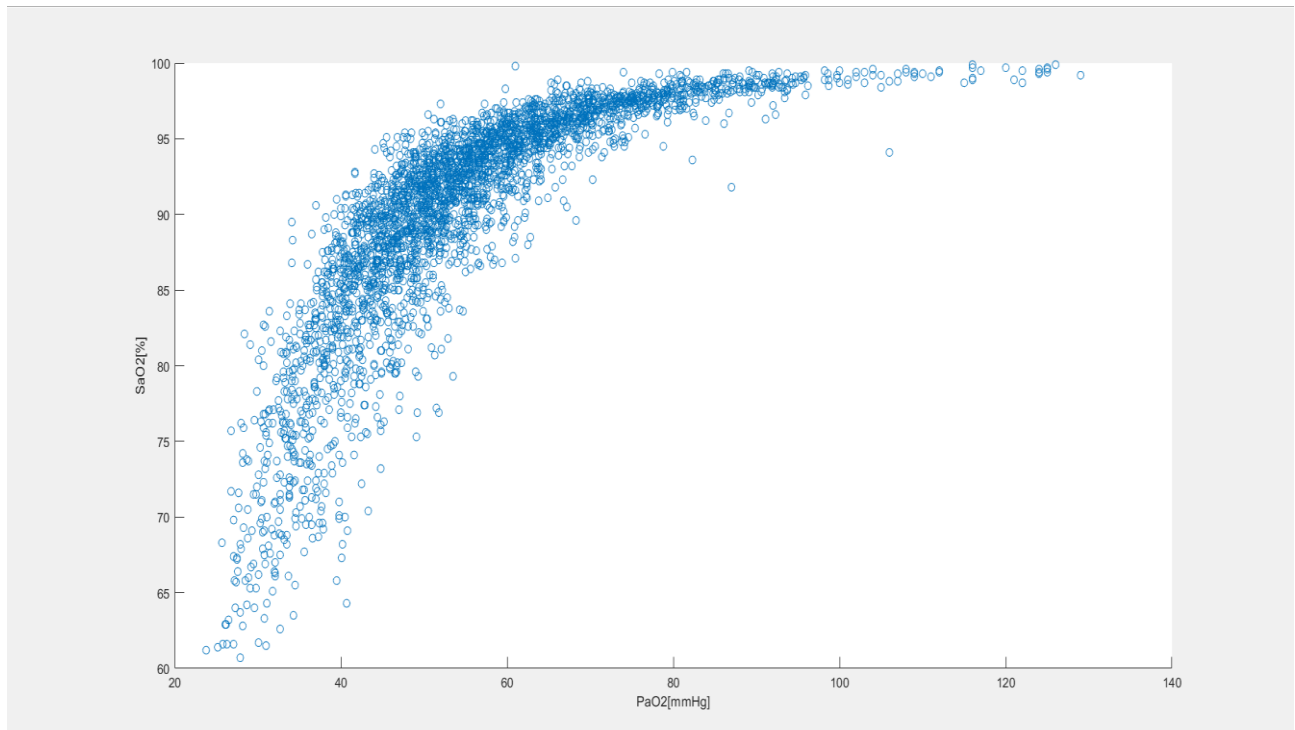


Figura 13 Plot dopo Tukey

Come evidenzia la figura il dataset è stato filtrato da molti outliers. Da 126635 campioni iniziali dopo il pre-filtraggio e dopo Tukey i campioni in esame sono circa 5mila cioè 4843. Si sarebbe potuto filtrare ulteriormente il dataset cercando di eliminare più outliers ed avere una distribuzione più pulita ma in questo modo si sarebbero persi anche dati non outlier, perciò ci si è fermati ad una equilibrata via di mezzo tra numero di dati per l'indagine statistica e numero di outliers.

Funzione di predizione

Dopo la fase di filtraggio si può fare inferenza statistica, cioè ipotizzando dei legami di causa-effetto tra due o più parametri, si vanno ad applicare dei metodi matematici per poter esplicitare la funzione che li lega, "regredire" alla regola matematica che collega i due fenomeni, da qui il metodo delle regressioni.

La specificazione di un modello consiste nell'esplicitare un legame tra i fenomeni di interesse:

$$Y = f(X_1, X_2, \dots, X_p)$$

Dove Y è la variabile da esplicitare, mentre X_1, X_2, \dots, X_p sono le variabili scelte per spiegare Y tramite la funzione $f(x)$

Inoltre, non è quasi mai plausibile ipotizzare un legame deterministico quindi si deve aggiungere un fattore di errore ε :

$$Y = f(X_1, X_2, \dots, X_p) + \varepsilon$$

Y= variabile dipendente

X_1, \dots, X_p = variabili esplicative

ε : variabile casuale errore

Oggi il termine regressione è divenuto significato di "relazione funzionale tra variabili ottenuta con metodi statistici".

I modelli statistici che possiamo avere, divisi tra lineari e non lineari possono essere:

- Semplici, una X e una Y
- Multipli, più di una X ma una sola Y
- Multivariati, più di una X e più di una Y

Modello di regressione lineare semplice

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Identifica una retta, nota come la retta di regressione:

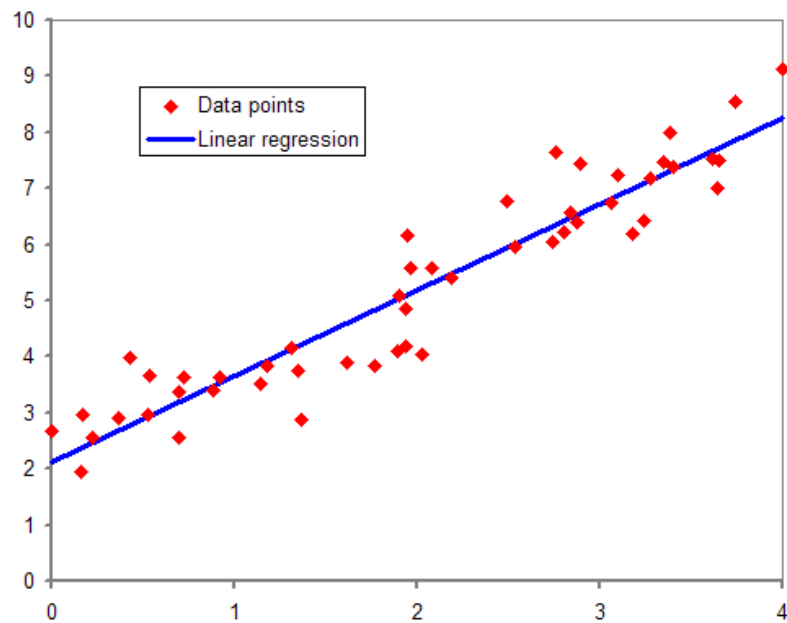


Figura 14 Retta di Regressione

β_0 = intercetta, il valore di Y_i quando $x_i=0$

β_1 = pendenza, di quanto cambia Y_i quando x_i incrementa di un'unità

ε_i = l'errore che si commette nella spiegazione della variabile y_i tramite una funzione lineare di x_i

Quando la retta ha pendenza positiva covariano positivamente, quando negativa, negativamente, quando la retta ha pendenza 0 non covariano.

La covarianza misura l'attitudine a covariare di due caratteri:

$$Cov(X, Y) = \frac{\sum_{i=1}^n (x - \bar{x})(y - \bar{y})}{n - 1}$$

È utile costruire una misura standardizzata che esprima quanto i due caratteri covariano:

$$p = \frac{cov(X, Y)}{std(X)std(Y)}$$

Il coefficiente di correlazione di *pearson* va da -1 (massima correlazione negativa) a +1 (massima correlazione positiva), se il coefficiente risulta 0 non c'è correlazione.

Per le regressioni lineari semplici si può calcolare tra tutte le rette la retta che minimizza la somma degli scarti dalla retta di regressione (**minimi quadrati**).

Siano definite le medie come:

$$\bar{x} = \frac{1}{n} \sum x_i \qquad \bar{y} = \frac{1}{n} \sum y_i$$

E la varianza:

$$Var(X) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Allora i coefficienti β che minimizzano la somma degli scarti della regressione devono essere calcolati in questo modo:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{Cov(X, Y)}{Var(X)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Modello di regressione lineare multiple

Quando ci sono più x ma una sola y si parla di regressione multipla il modello è il seguente:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$$

Y_i = risposta ai valori, ossia rappresenta il risultato previsto dal modello;

β_0 = intercetta, ossia il valore di \hat{y} quando gli x_i sono tutti uguali a 0;

β_1 = è il coefficiente di X_1 (la prima caratteristica);

β_n = è il coefficiente di X_n (l'ennesima caratteristica);

X_1, X_2, \dots, X_n = le variabili indipendenti del modello;

ϵ_i = errore statistico;

I β sono numeri reali e vengono chiamati anche coefficienti di regressione stimati del modello.

Possiamo scrivere in forma matriciale:

$$\beta = [\beta_0, \beta_1, \dots, \beta_k]'$$
$$X_i = \begin{bmatrix} 1 \\ X_{1i} \\ \vdots \\ X_{ki} \end{bmatrix}$$
$$y_i = [\beta_0, \beta_1, \dots, \beta_k]' \begin{bmatrix} 1 \\ X_{1i} \\ \vdots \\ X_{ki} \end{bmatrix}$$
$$\epsilon_i = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Quindi per ogni y :

$$Y_i = \beta' X_i + \epsilon_i$$

Il vettore Y raccoglie tutte le osservazioni della variabile dipendente. La matrice X raccoglie le osservazioni sulle variabili esplicative. Ogni colonna di X contiene tutte le osservazioni per la singola variabile esplicativa.

È molto utile costruire la matrice di covarianza, contenente sulla diagonale principale la varianza della variabile i -esima e nelle altre celle la covarianza tra le variabili, è una matrice diagonale, simmetrica.

$$\begin{bmatrix} \sigma_{11}^2 & \cdots & cov_{1m} \\ \vdots & \ddots & \vdots \\ cov_{m1} & \cdots & \sigma_{mm}^2 \end{bmatrix}$$

Nello studio in questione abbiamo deciso di rapportare tutto alla variabile di pressione parziale dell'ossigeno quindi il modello che utilizziamo è quello di regressione multipla ma per completezza di trattazione esiste anche il modello di regressione multivariata con il quale potevano essere prese in considerazione tutte le variabili e trattate insieme a scapito dell'aumento di complessità dei sistemi da trattare.

Una volta definito il modello di regressione adatto al problema in esame si applicherà un approccio "step-wise":

- forward, Parto dal modello con una variabile e man mano ne aggiungo un'altra. Se l'inclusione della variabile è significativa, la tengo; altrimenti, la scarto.
- Backward, Parto dal modello completo e man mano ne levo una. Se l'esclusione della variabile è significativa, la tengo; altrimenti, la scarto.

Feature selection

Per capire quali variabili tenere e quali scartare è opportuno applicare metodi di feature selection.

La selezione delle principali componenti di analisi è un algoritmo che consente di selezionare, a partire dal dataset iniziale, un sub-set secondario, ottenuto come risultato di n iterazioni della stima dell'errore di precedenti sub-set inizialmente più grandi, fino a trovare quello con il coefficiente di errore inferiore. I metodi per applicare una selezione delle componenti sono molteplici, nel caso in esame si è applicato un algoritmo di feature selection detto NCA, neighborhood component analysis.

Dopo numerosi allenamenti la feature selection permette di passare da 35 feature iniziali a 7 feature principali che danno un errore medio di 3.6mmHg rispetto alla pressione parziale dell'ossigeno, le feature sono:

- **BM** (mmHg), pressione atmosferica
- **pH** concentrazione di ioni idrogeno
- **PCO₂** (mmHg), pressione parziale dell'anidride carbonica
- **THb** (g/dL), concentrazione di emoglobina totale in tutte le forme
- **SO₂** (%), saturazione dell'ossigeno
- **HbF** (%), percentuale di emoglobina fetale sul totale
- **tBil** (mg/dL), concentrazione di bilirubina totale in tutte le forme

Con l'approccio step-wise, sia backward che forward, mano a mano che viene allenata la funzione con determinate feature si rende necessario uno strumento di controllo, ovviamente quello immediato è il controllo del MSE mean squared error che ci da un indice chiaro di quanto siamo lontani da avere un modello perfetto, sono stati utilizzati anche altri metodi grafici:

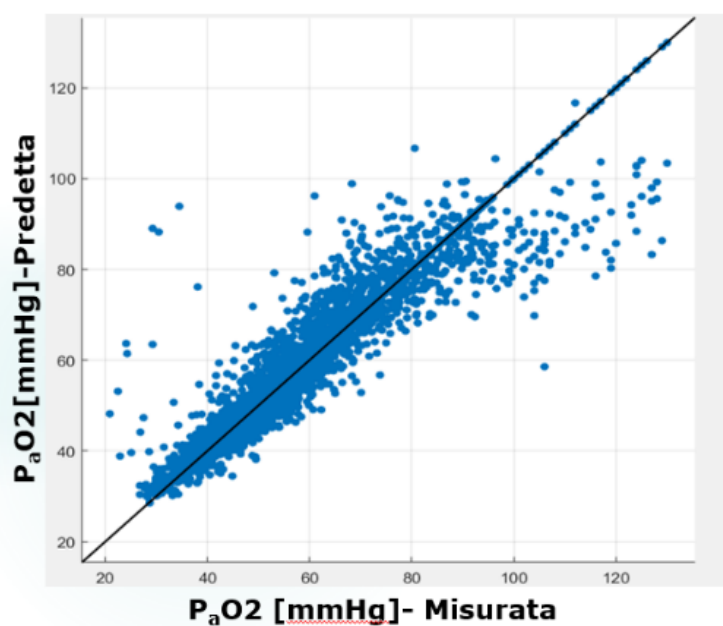


Figura 26 Plot Predetto Misurato

L'algoritmo di feature ci permette di determinare una funzione di predizione col metodo delle regressioni, si mostra in questo grafico qual è l'andamento della nostra funzione, in particolare vengono messe in confronto la pressione parziale misurata e quella derivante dalla funzione di predizione.

La retta di regressione è il risultato grafico della funzione che minimizza i residui delle distanze tra i punti misurati e quelli predetti, più i punti sono vicini alla retta e più il modello è giusto.

Si osserva che la maggior parte dei punti si accostano alla retta mentre solo una piccola parte risulta significativamente lontana.

I residui (differenza tra misurato e predetto) sono visualizzati come rette arancioni che uniscono due punti, la media di tutte queste distanze ci da il MSE.

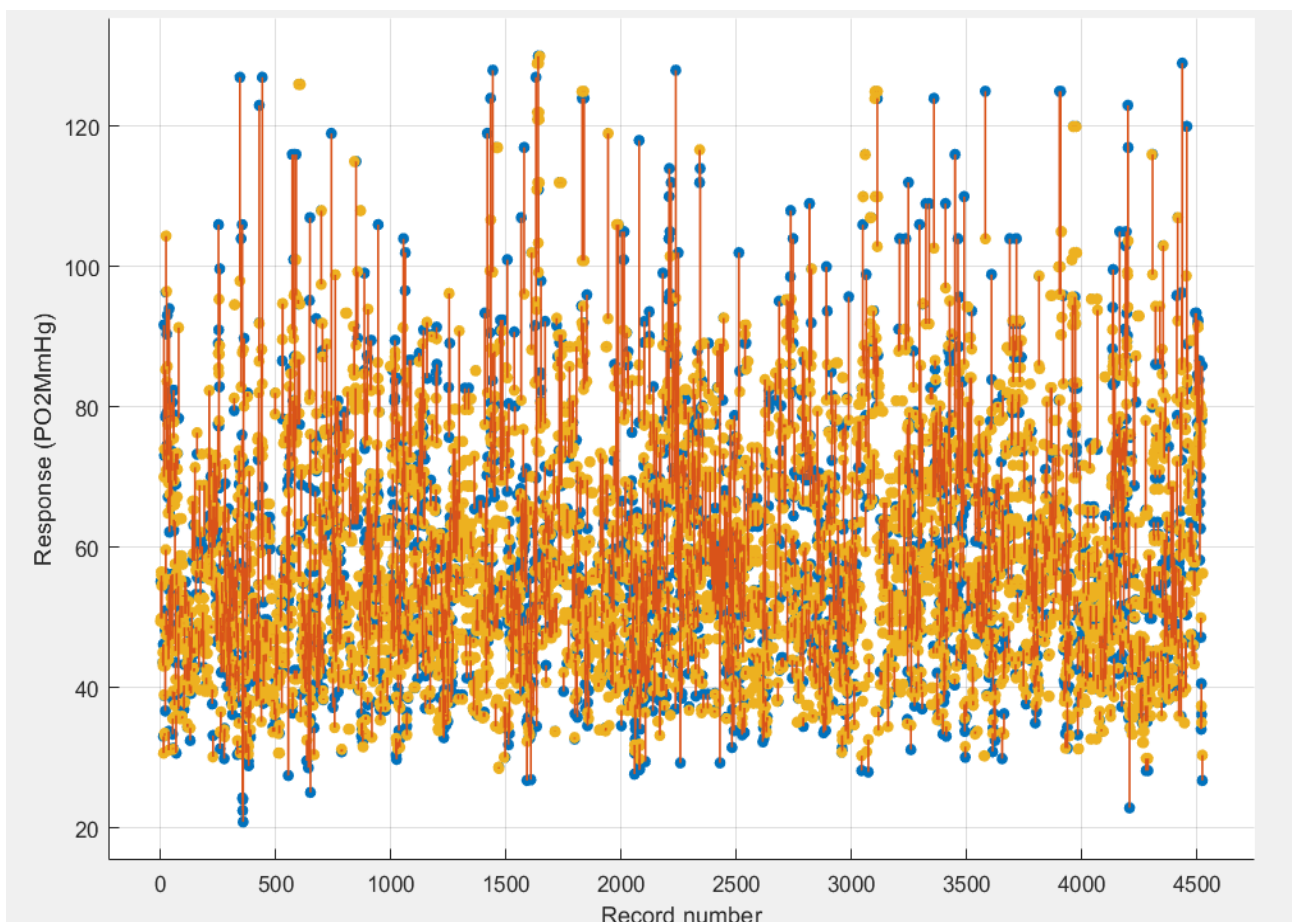


Figura 27 Plot Residui

Maggiore è la distanza e maggiore è il modulo del RMSE, quindi i valori con minor errori sono considerati quelli più intercorrelanti.

I residui possono essere rappresentati anche in modo alternativo alla rappresentazione precedente.

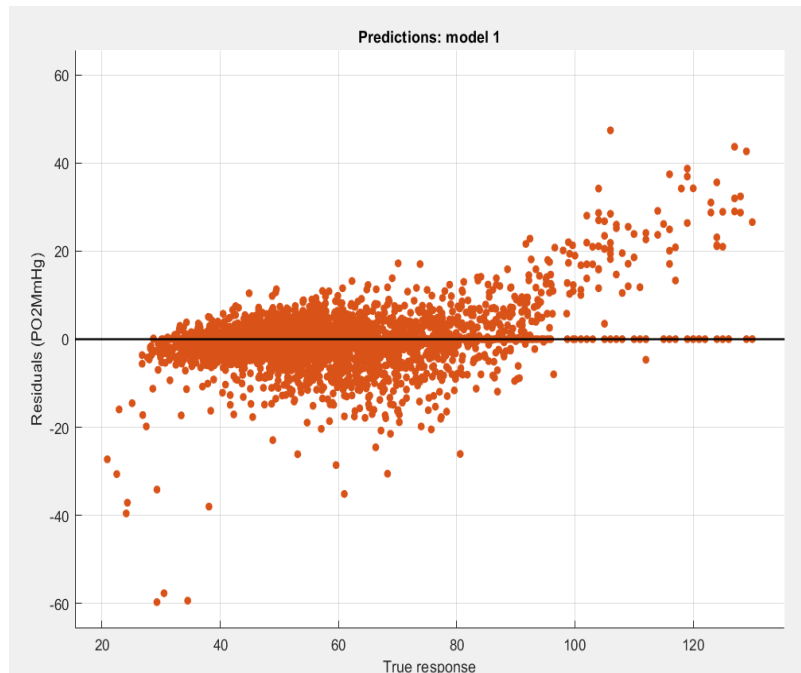


Figura 28 Plot Response Residui

Si osserva che una minor parte dei dati ha un grosso discostamento dallo 0, questi dati da soli concorrono nella gran parte del modulo del RMSE, mentre la maggior parte concorre in maniera molto inferiore se non nulla.

Metodi di convalida

La convalidazione è il processo che permette di determinare se uno specifico metodo di misura è strutturato correttamente osservando se esistono sostanziali discrepanze nei residui. La si svolge adottando un set campionario grande il 10% di tutto il set di partenza.

Metodi di validazione usati:

- Bland-Altman plot
- Coefficiente di Pearson
- Durbin-Watson plot

Bland-Altman plot

Stima della bontà del modello di predizione attraverso il confronto del dataset con un dataset fittizio generato randomicamente. Se la differenza tra i due metodi rientra in un determinato range assumiamo il nostro modello come accettabile.

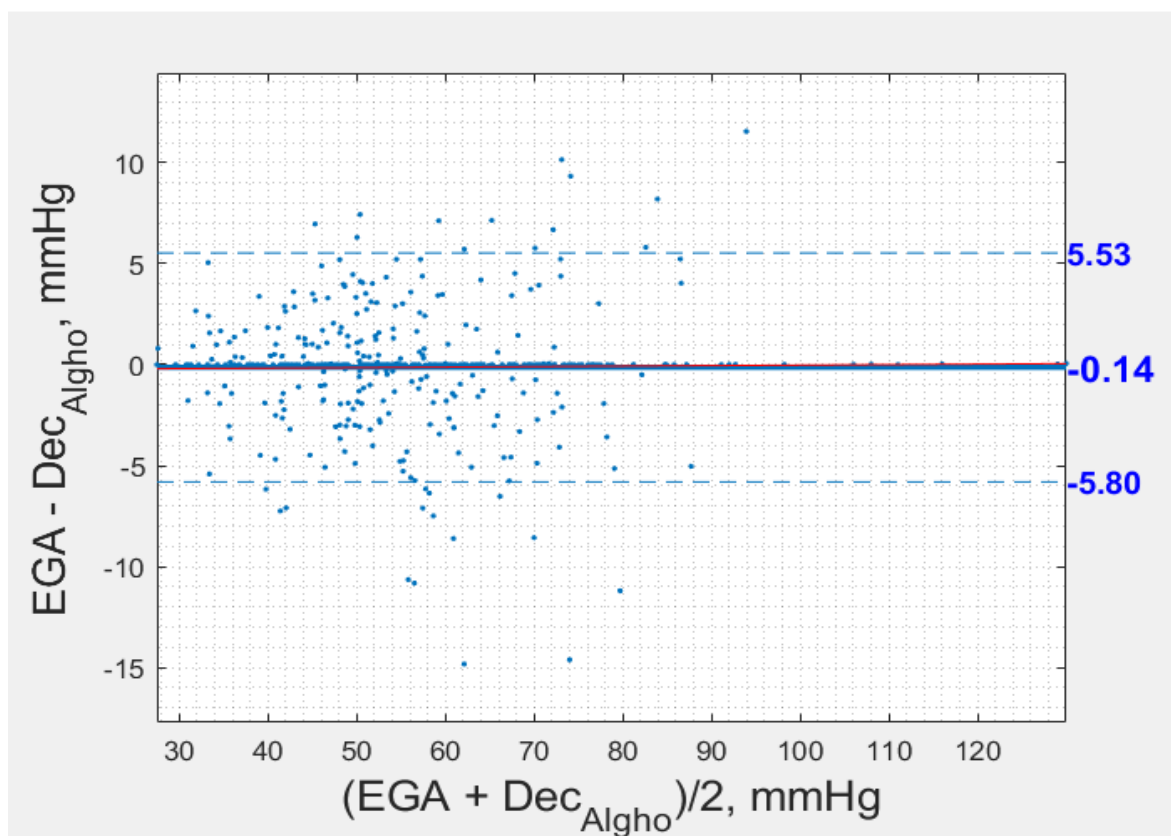


Figura 29 Bland-Altman Plot

Rappresentiamo le medie tra i valori del dataset A e del dataset B in relazione alla loro relativa differenza. Nel caso in esame EGA rappresenta il 10% del dataset dei predetti preso a caso e Dec invece un dataset random di misurazioni. Il dato attendibile si trova all'interno dell'intervallo di confidenza (-5,80;5,53). La media delle n coppie rappresenta il dato statistico in sé, la differenza invece è la stima della bontà del modello. Per calcolare l'intervallo di confidenza si usa la deviazione standard, + 1.96*std per il limite superiore, -1.96*std per il limite inferiore.

Come si può notare, non esiste un trend significativo esplicativo dell'andamento dei residui, la disposizione è casuale e non esiste quindi una sistematicità, cioè i due metodi forniscono le stesse misure.

Coefficiente di Pearson

Il coefficiente p è un indice di correlazione tra i dati, questo spiega che grado correlazione esiste tra la misura e la predizione, è definito nell'intervallo tra (-1;1). In particolare dato l'indice di correlazione così definito:

$$p_{AB} = \frac{cov(A, B)}{\sigma_A \sigma_B}$$

- se l'indice > 0 allora si avrà una correlazione diretta tra le variabili ovvero correlate positivamente.
- se l'indice $= 0$ non si avrà nessuna correlazione
- se l'indice < 0 si avrà una correlazione indiretta ovvero le variabili sono correlate negativamente.

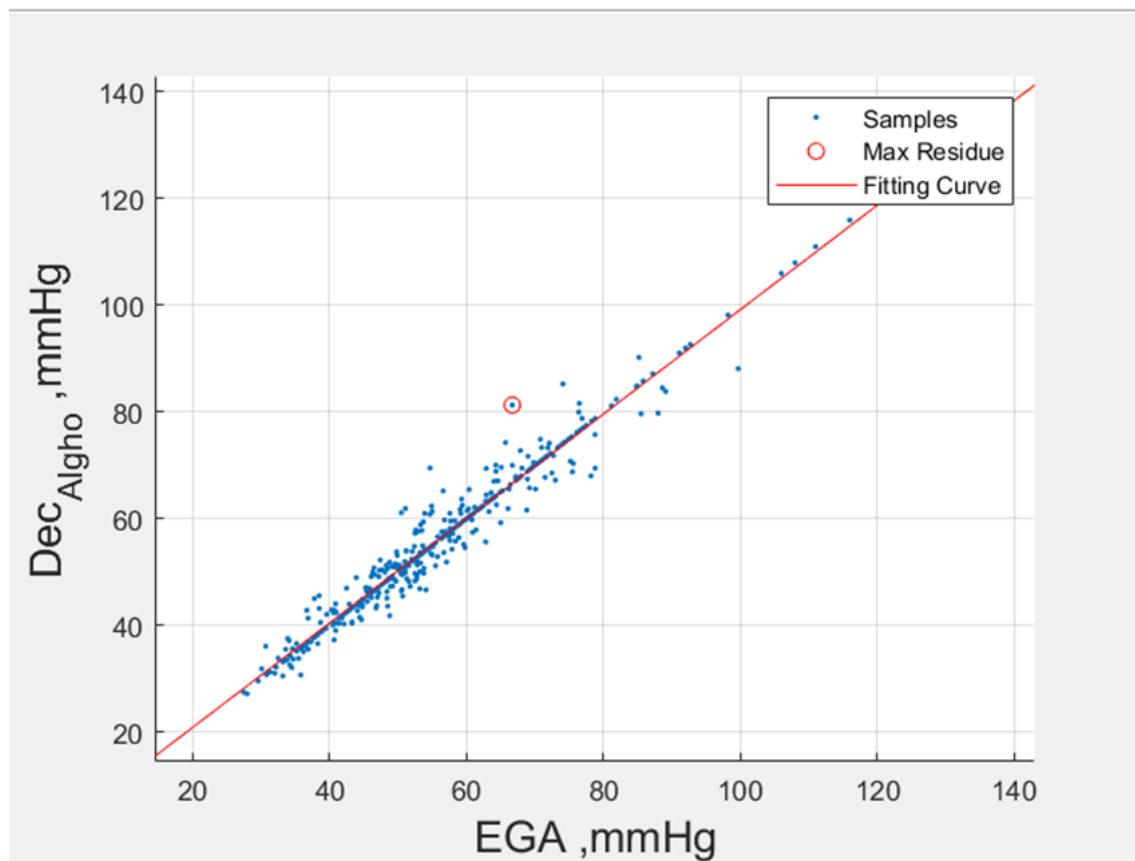


Figura 30 Pearson Plot

Anche in qui abbiamo preso un set campionario EGA e un set randomico di misurazioni Dec.

In questo caso $p = 0,98$ e questo indica una quasi totale correlazione positiva tra i dati.

Mentre la retta fittante è definita da un coefficiente angolare di 0,978 e la sua intercetta si trova in $y=1,31$ mmHg.

Durbin-Watson plot

L'analisi dei residui fatta con Durbin-Watson è necessaria per capire quale sia la distribuzione degli stessi residui rispetto al set campionario. Per una buona inferenza statistica è necessario che siano distribuiti in modo quanto più possibile normale.

Il risultato è definito in un intervallo tra (0;4), un valore di 2 indica nessuna correlazione dal momento che il nostro indice $d=0,781$ è un valore relativamente piccolo ed indica che i residui sono vicini tra loro in successione, hanno perciò una correlazione positiva, indice di normalità della distribuzione dei residui.

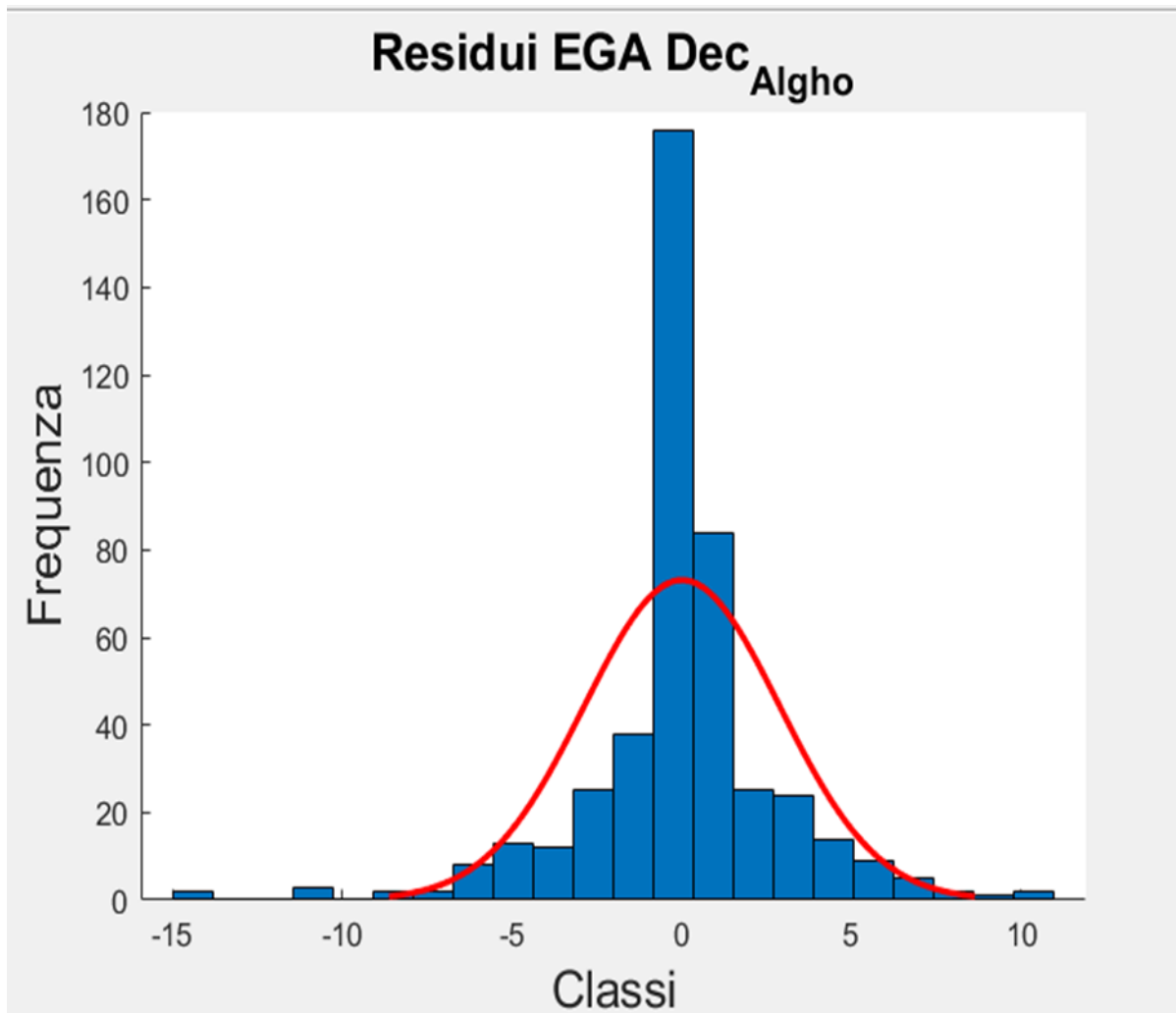


Figura 31 Durbin Watson Plot

Lilliefors Test

Una versione migliorata del Kolmogorov–Smirnov test, (K-S test), è un test non parametrico per verificare la distribuzione normale dei nostri dati, nel particolare lo usiamo per verificare la normalità dei nostri residui.

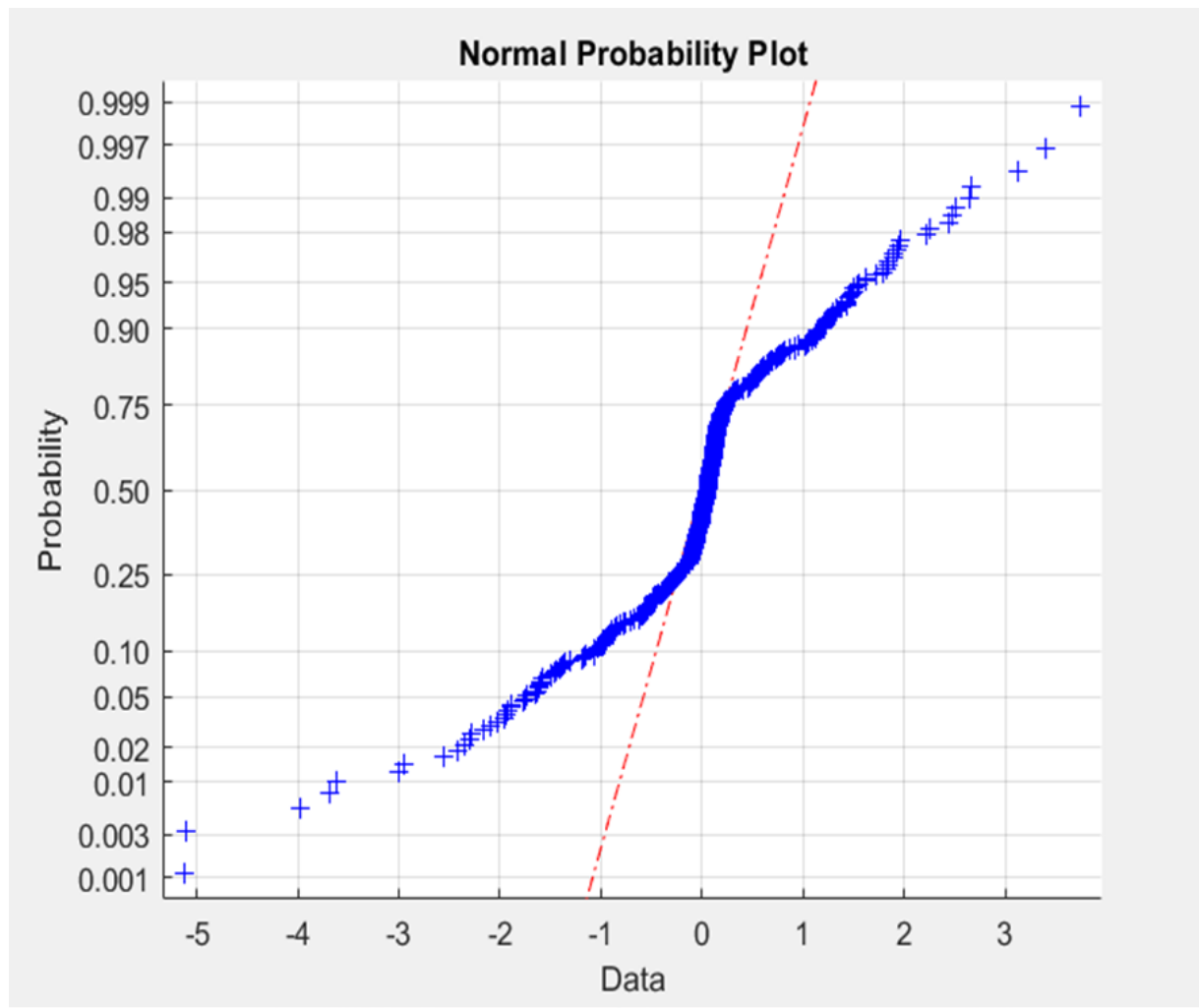


Figura 32 Normal Probability Plot

Da 0.25 a 0.75 i dati fittano perfettamente a retta, quindi sono distribuiti in modo normale, nei tratti da 0.10 a 0.25 e 0.75 a 0.90 sono abbastanza vicini alla retta per essere considerati sostanzialmente normali, nelle code i valori si distaccano anche di molto dalla retta e quindi non sono distribuiti in modo normale ma costituiscono una parte minima dei nostri residui quindi il test sulla normalità dei residui può essere considerato soddisfatto positivamente. A riconferma c'è anche da ricordare che questo test è molto poco sensibile sui valori di coda, al contrario è sensibile sui valori centrali dove il test risulta quasi ideale.

Studio incertezza con Montecarlo

Il metodo Monte Carlo è una tecnica matematica che consente di valutare e studiare determinati problemi attraverso l'uso di una simulazione.

- Consente di valutare l'incertezza anche nel caso di relazioni ingresso-uscita fortemente non lineari
- Non richiede la definizione analitica della reazione fra la misura indiretta e quella diretta
- Non richiede ipotesi sulla tipologia di distribuzione statistica delle misure in ingresso
- Permette di valutare anche uscite multidimensionali (analisi multivariata)

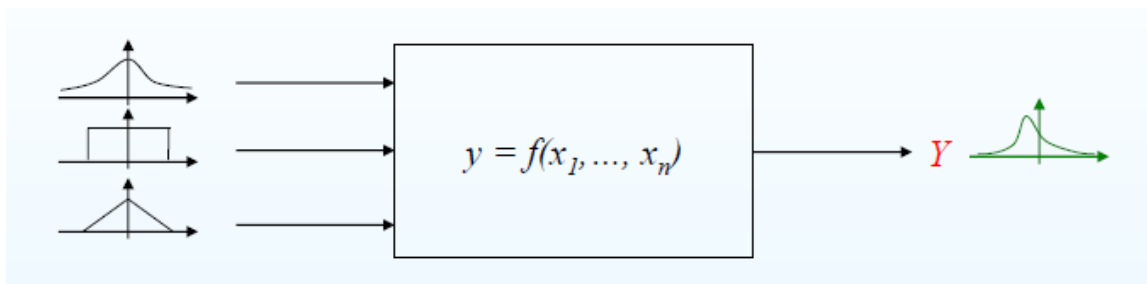
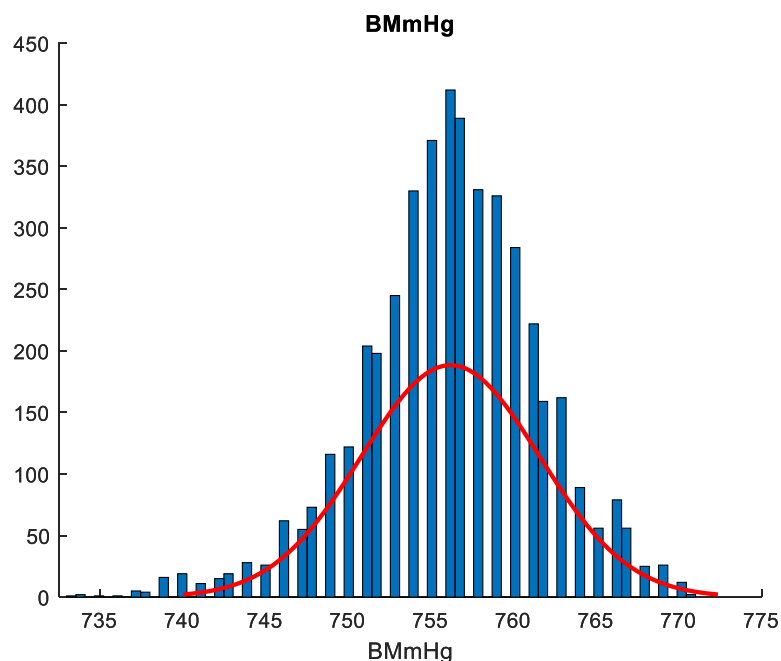


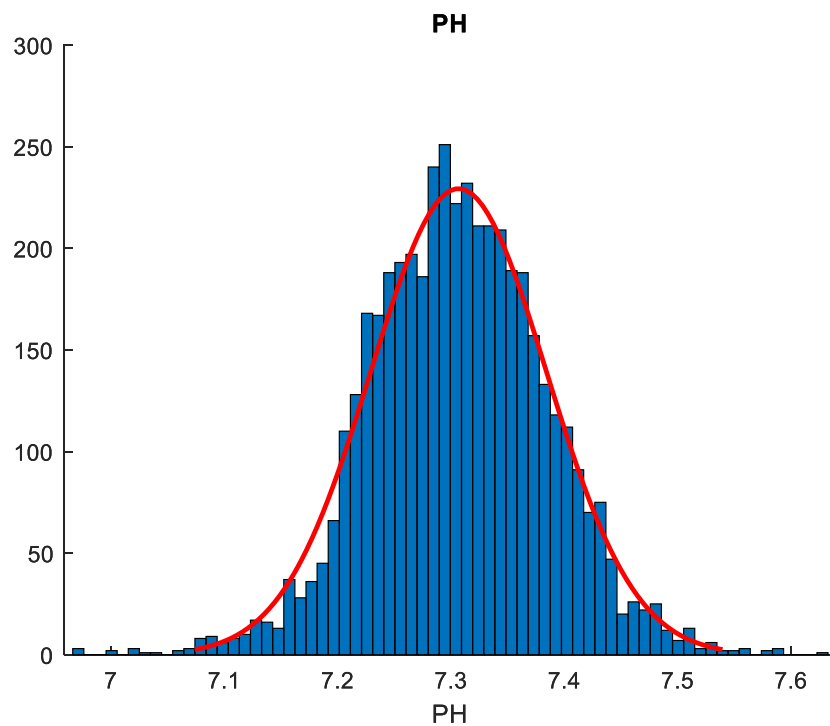
Figura 33 Metodo Montecarlo chart

Nel caso in esame abbiamo simulato le varie distribuzioni delle variabili e date in ingresso alla funzione di predizione per poi valutare l'incertezza del sistema:

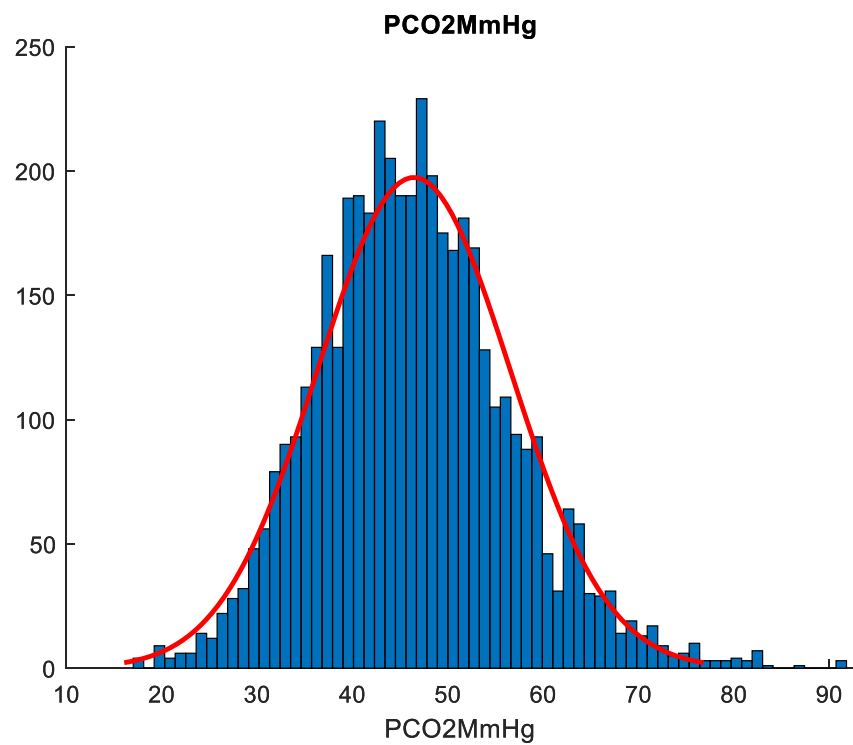
- distribuzione della pressione atmosferica BM (mmHg)



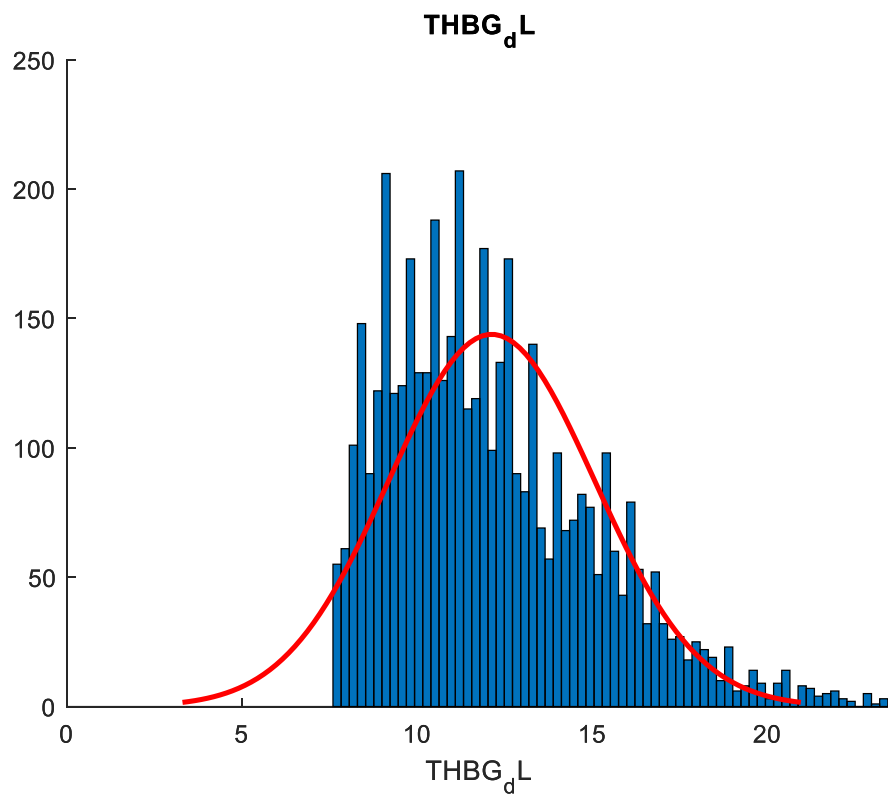
- distribuzione del pH



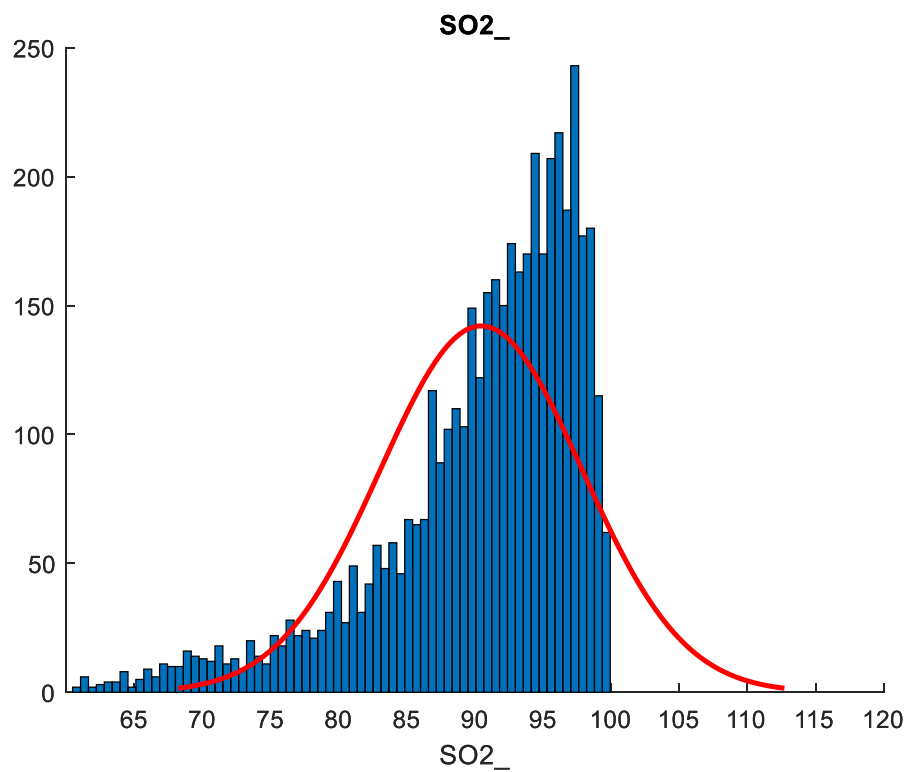
- distribuzione della pressione parziale dell'anidride carbonica $p\text{CO}_2$ (mmHg)



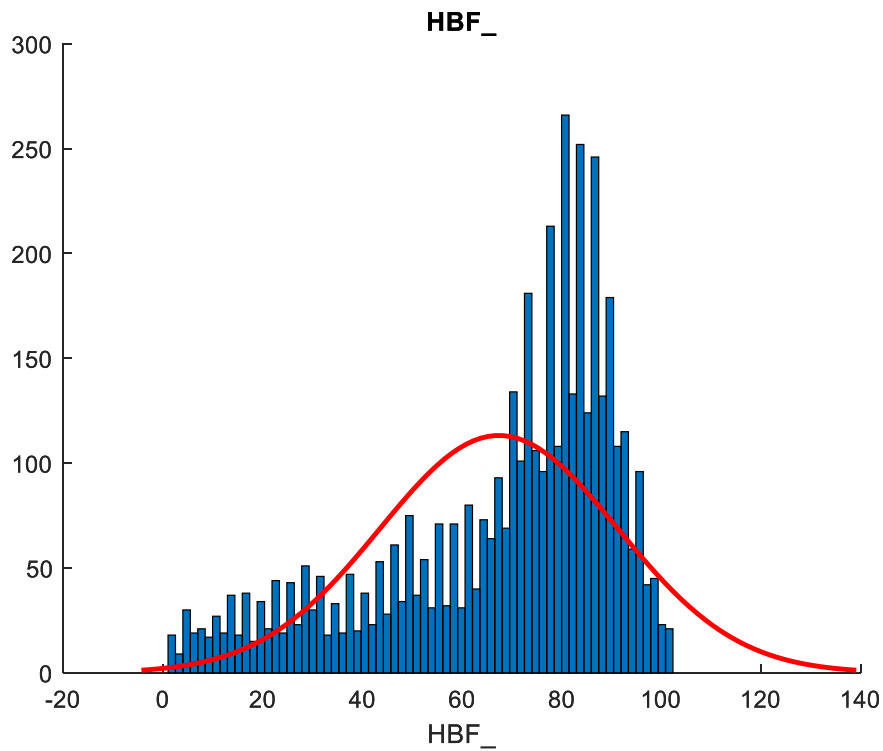
- distribuzione della concentrazione totale di emoglobina THB (g/dL)



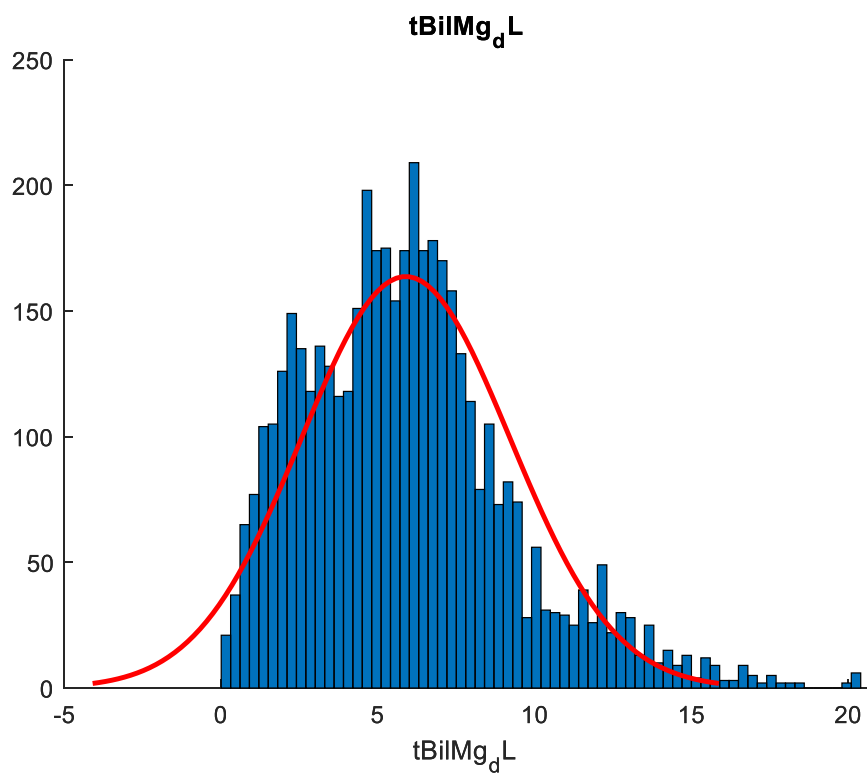
- distribuzione della saturazione SO₂(%)



- distribuzione dell'emoglobina fetale HbF (%)



- distribuzione della concentrazione di bilirubina tBil (mg/dL)



Le modellazioni randomiche delle features della funzione di predizione vengono messe in ingresso al sistema e viene calcolata la funzione di trasferimento IN/OUT che ci da una misura dell'incertezza del nostro sistema.

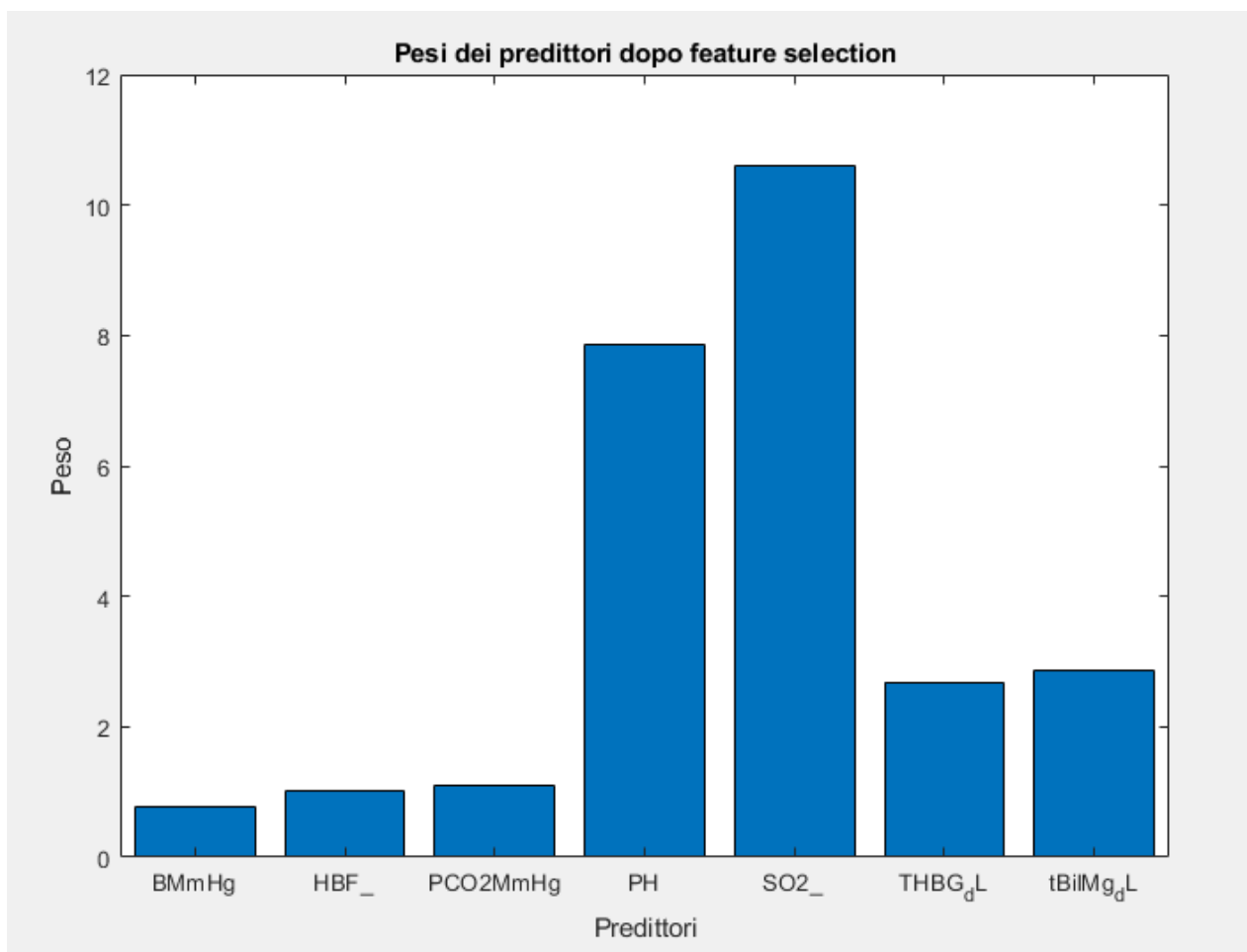
BM_{mmHg}	pH	pCO₂mmHg	THB_{g/dL}	SO₂%	HBF%	tBil_{mg/dL}
0,4271	26,1478	0,2386	0,9501	2,4762	0,3547	0,8680

Conclusioni

Per concludere lo studio, possiamo dire che la stima predetta della pressione parziale dell'ossigeno nel distretto arterioso può essere fatta partendo da un set di variabili indipendenti, tra questi, concorrono maggiormente il pH e la saturazione dell'ossigeno.

Quest'ultima la si può misurare per mezzo di un pulsossimetro.

Con un'incertezza di ± 3.5 mmHg, possiamo stimare e quindi monitorare in tempo reale la pressione parziale dell'ossigeno e tracciare la curva di saturazione.



Bibliografia:

Regressioni

- Ronald Fisher (1925) *Statistical Methods for Research Workers*.
- Ronald Fisher (1922) The goodness of fit of regression formulae, and the distribution of regression coefficients, *Journal of the Royal Statistical Society* 85, 597-612;

Algoritmo Tukey per outliers

- Seo, Songwon (2006) *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets*. Master's Thesis, University of Pittsburgh.
- David C. Hoaglin, Boris Iglewicz & John W. Tukey (1986) Performance of Some Resistant Rules for Outlier Labeling, *Journal of the American Statistical Association*, 81:396, 991-999, DOI: 10.1080/01621459.1986.10478363
- Hoaglin, D. C. 1983. "Letter Values: A Set of Selected Order Statistics,". In *Understanding Robust and Exploratory Data Analysis*, Edited by: Hoaglin, C., Mosteller, Frederick and Tukey, W. John. 33-57. New York: John Wiley.

Bland-Altman

- J.M. Bland, D.G. Altman Statistical methods for assessing agreement between two methods of clinical measurement *Lancet*, 1 (1986), pp. 307-310
- M.L. Greenfield, J.E. Kuhn, E.M. Wojtys A statistics primer. Correlation and regression analysis

Am J Sports Med, 26 (1998), pp. 338-343

Pearson

- Sheldon M. Ross, *Introduzione alla statistica*, 2^a ed., Maggioli Editore, 2014

Durbin Watson

- Greene, William H., 2003, "Econometric Analysis," Prentice Hall International Edition

Curve dissociazione emoglobina

- John W. Severinghaus. Simple, accurate equations for human blood O₂ dissociation computations. *J. Appl. Physiol: Respirat. Environ. Exercise Physiol.* 46(3):599-602, 1979.
- Oxygen saturation calculation procedures: a critical analysis of six equations for the determination of oxygen saturation H.-W. M. Breuer, H. Groeben, J. Breuer and H. Worth
- Blood HbO₂ and HbCO₂ Dissociation Curves at Varied O₂, CO₂, pH, 2,3-DPG and Temperature Levels RANJAN K. DASH1 and JAMES B. BASSINGTHWAIGHTE
- Burtis, Carl A. Tietz, Norbert W. Bruns, David E. Ashwood, Edward R. - Tietz textbook of clinical chemistry and molecular diagnostics-Elsevier Saunders (2006)