PCA VINI ROSSI

Relazione di Francesco Malferrari

Il set di dati preso in analisi è proveniente da un insieme di vini rossi di diversa provenienza, ciascuno rappresentato nel dataset con acronimi indicati di seguito:

- Barolo (OLO);
- Barbera (ERA);
- Grignolino (GR).

In particolare, i vini vengono sottoposti ad un'analisi delle componenti principali (PCA) in base a 13 variabili:

- "Alcohol": il grado alcolico (% in volume);
- "Malic ac": il contenuto di acido malico in g/l (uno dei principali acidi organici presenti nelle uve da vino);
- "Ash": descrive la parte inorganica (ceneri e sali);
- "Alcalinity of ash": l'alcalinità delle ceneri. Esprime approssimativamente le quantità di acidi organici presenti nel vino sottoforma di sali (pH);
- "Mg": Magnesio in g/kg;
- "Phenols": composti fenolici in mg/L (sostanze naturali che danno colore al vino oltre che a sensazioni gustative);
- "Flavanoids": composti flavonoidi in mg/L (i polifenoli più abbondanti nel vino);
- "Nonflav phen": composti nonflavonoidi fanolici in mg/L (conferiscono caratteristiche specifiche al vino e creano anche aromi e sapori specifici durante la fermentazione e la vinificazione);
- "Proanthoc": proto-antocianine in mg/L (un tipo di fenolo antiossidante del vino rosso);
- "Color intensity": intensità del colore rosso calcolata sommando l'assorbimento a 420 nm (marrone), a 520 nm (rosso) e 620nm (blu);
- "Hue": saturazione del colore;
- "OD280/OD315": rapporto tra 2 lunghezze d'onda nel UV-Vis. È un metodo per determinare la concentrazione di proteine nei vini;
- "Proline": prolina in mg/L (un amminoacido).

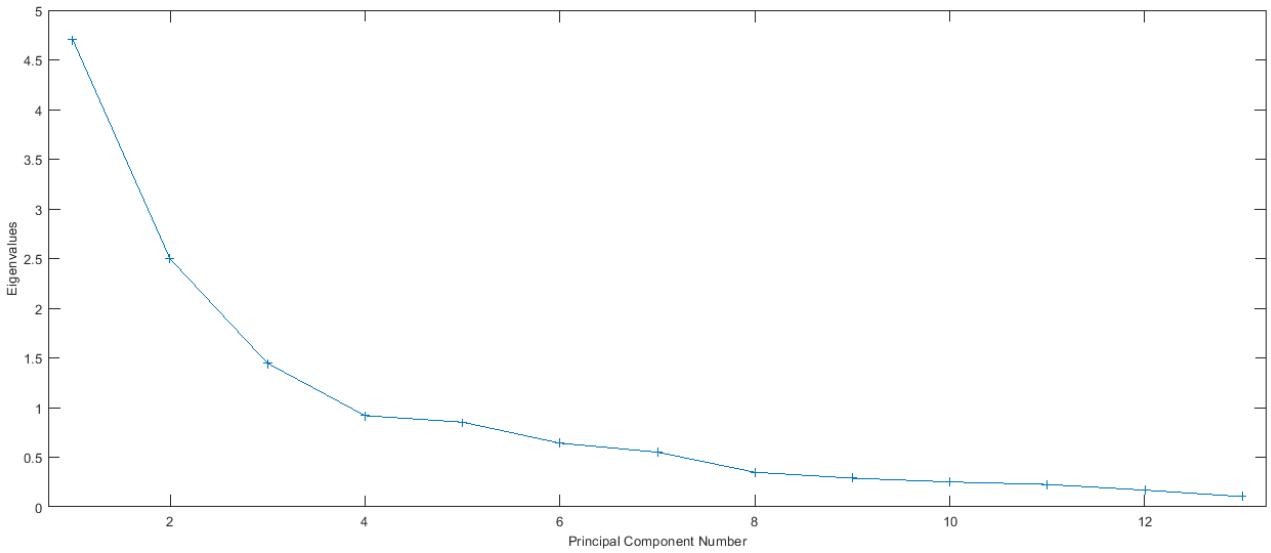
La tabella che colleziona i valori per ciascun vino è composta da 13 variabili e 178 campioni (59 per OLO, 71 per GR e 48 per ERA).

Quello che si vuole fare con questo dataset è individuare le differenze sistematiche tra le varie provenienze, per quali variabili e le eventuali relazioni tra i vari composti attraverso l'analisi esplorativa multivariata.

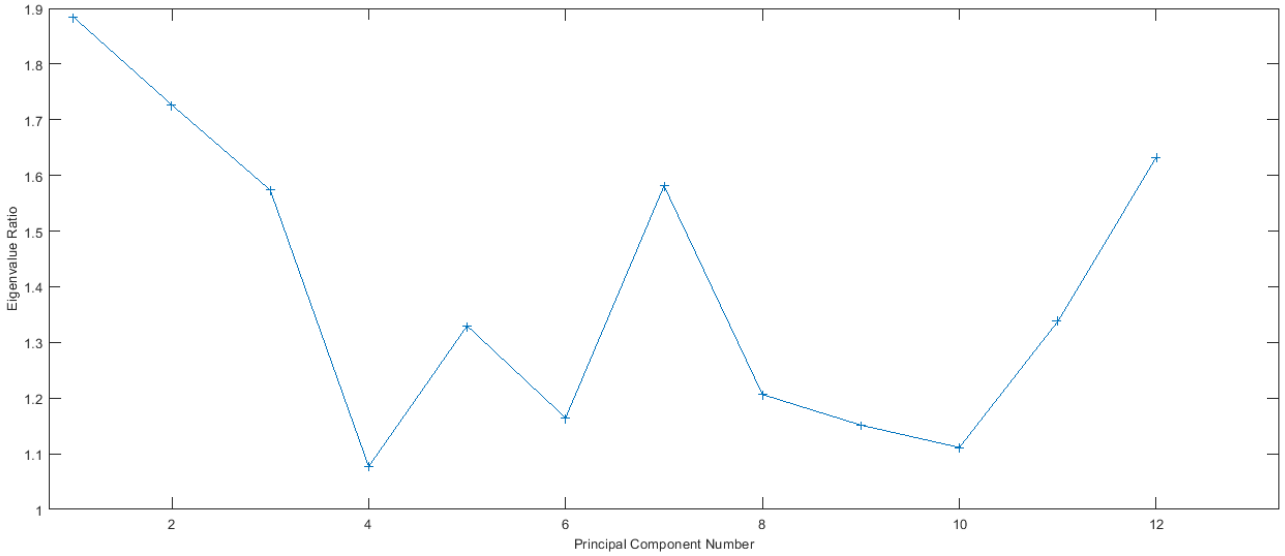
Per fare ciò è stato usato il PLS Toolbox, una suite completa e avanzata di strumenti di analisi chemiometrica multivariata presente nell'ambiente computazionale MATLAB.

Il preprocessing scelto è "Autoscale" in quanto si tratta di dati non omogenei, con unità di misura diverse e con differenti scale di misura (come elencato prima).

Per la scelta del numero di componenti principali (PC) è stato fondamentale guardare il plot degli autovalori e individuare il numero oltre al quale ciò analizzato è praticamente solo rumore.

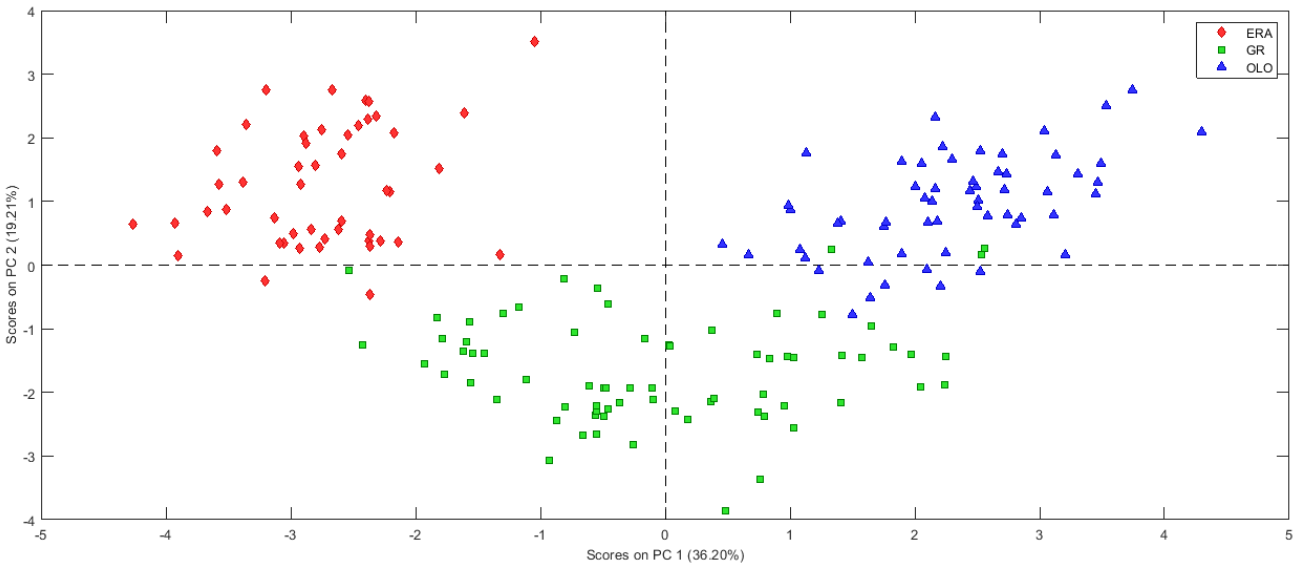


Da quanto si vede può essere 4, 5 o 6 il numero ideale, ma per togliere ogni dubbio si va a guardare anche il grafico del rapporto tra gli autovalori e si cerca quale numero tende ad avvicinarsi al valore 1 (ricordandoci che si comincia a contare da 2 perché facendo il rapporto si scala di uno).



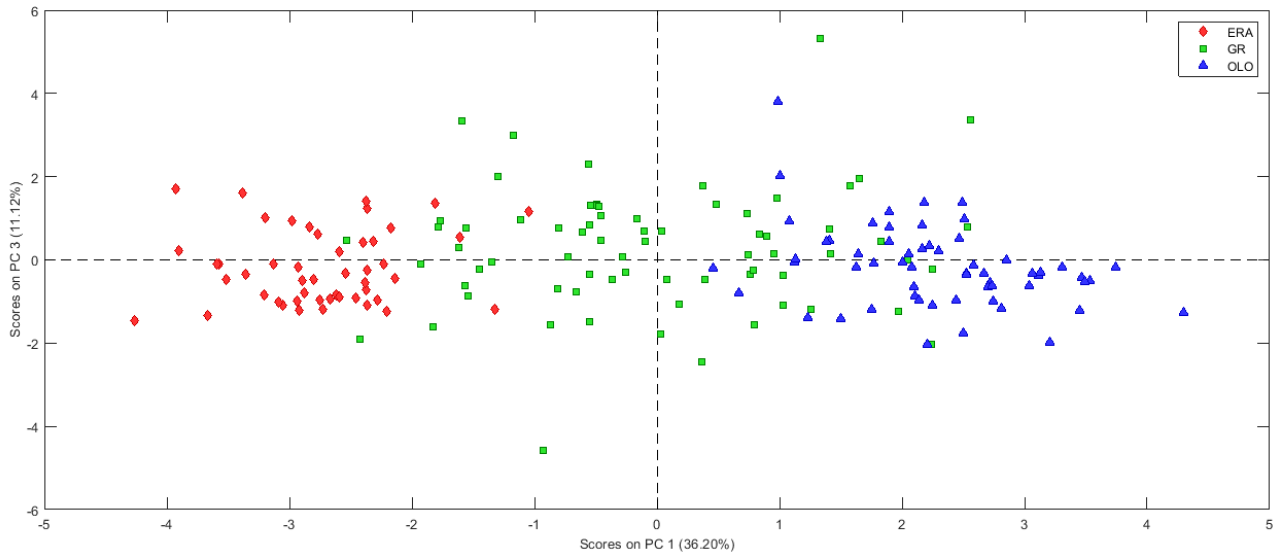
Da qua possiamo intendere che il numero di componenti principali ottimale è 5.

Ora si osservano i grafici degli score, a partire dalla relazione tra PC1 e PC2.

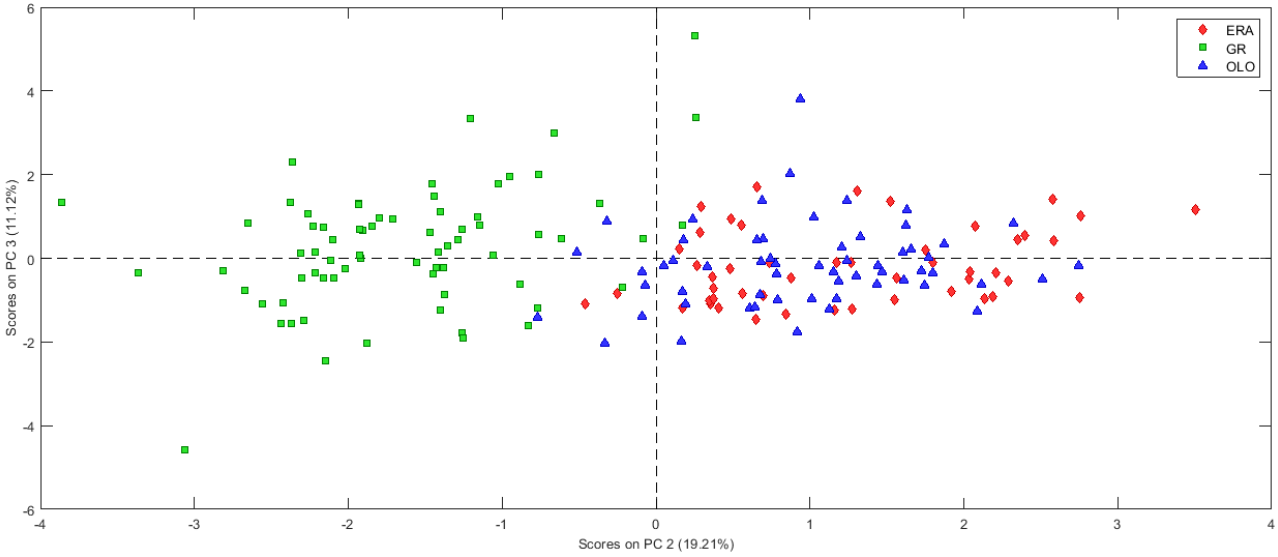


Da questa prima analisi i tre vini si distinguono molto bene fatta eccezione di qualche campione GR ed ERA sparso.

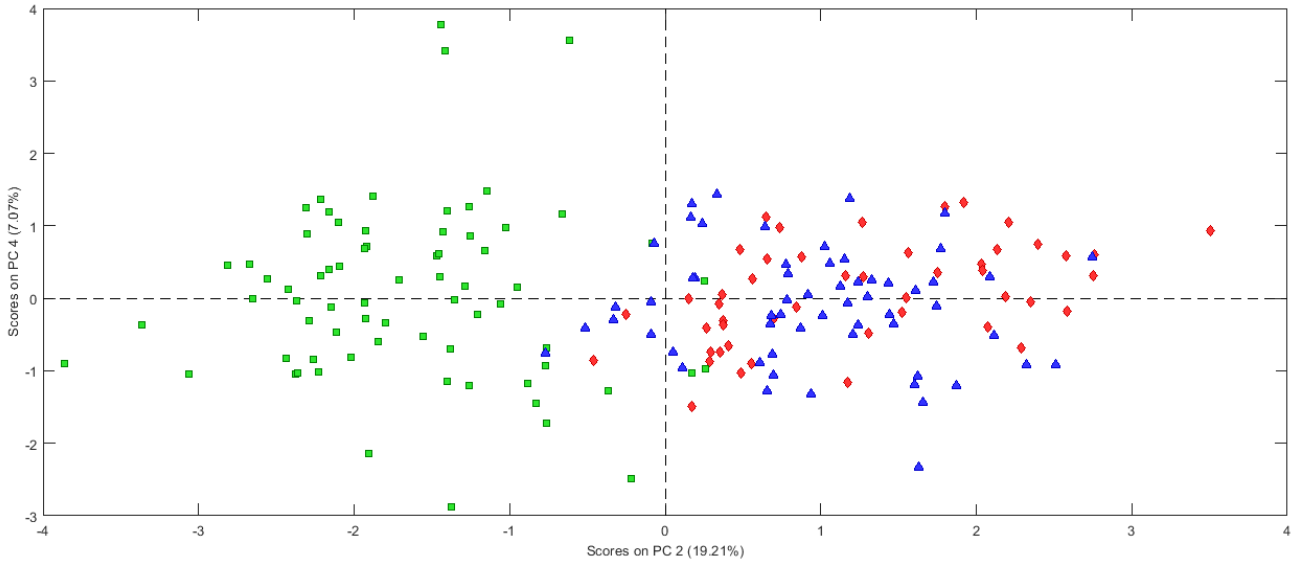
Ora si guarda la relazione tra PC1 e PC3, dove si può notare i raggruppamenti distinti sulla prima componente principale.



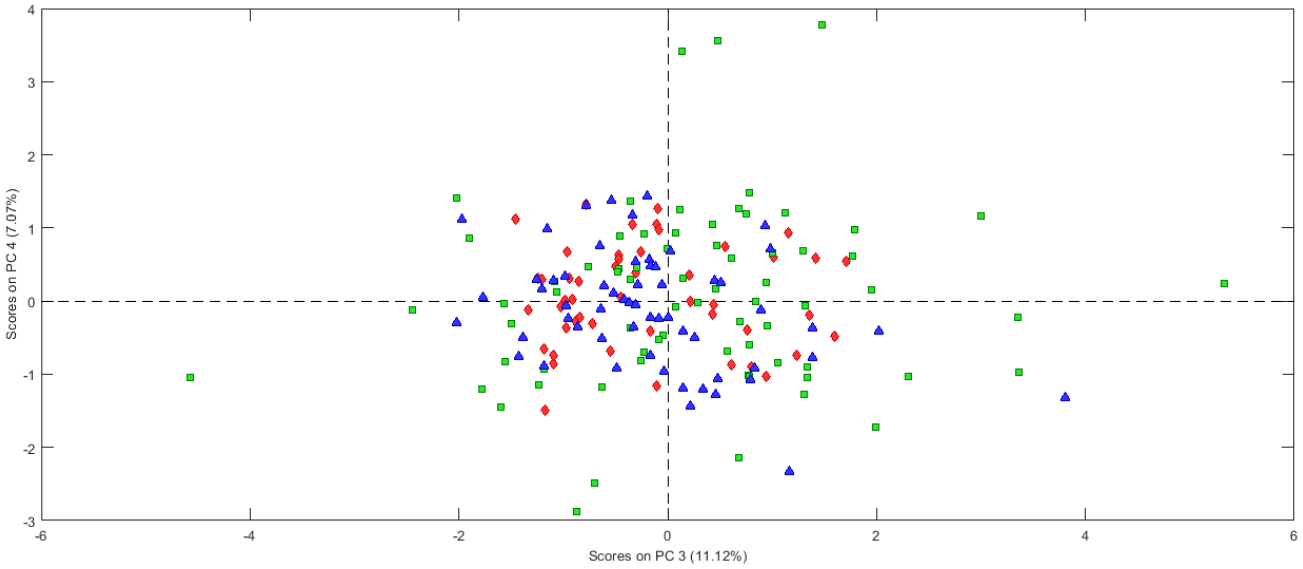
Con il grafico PC2-PC3 invece si può isolare solo il vino GR sulla seconda componente principale. Quindi PC3 non da nessun contributo in particolare.

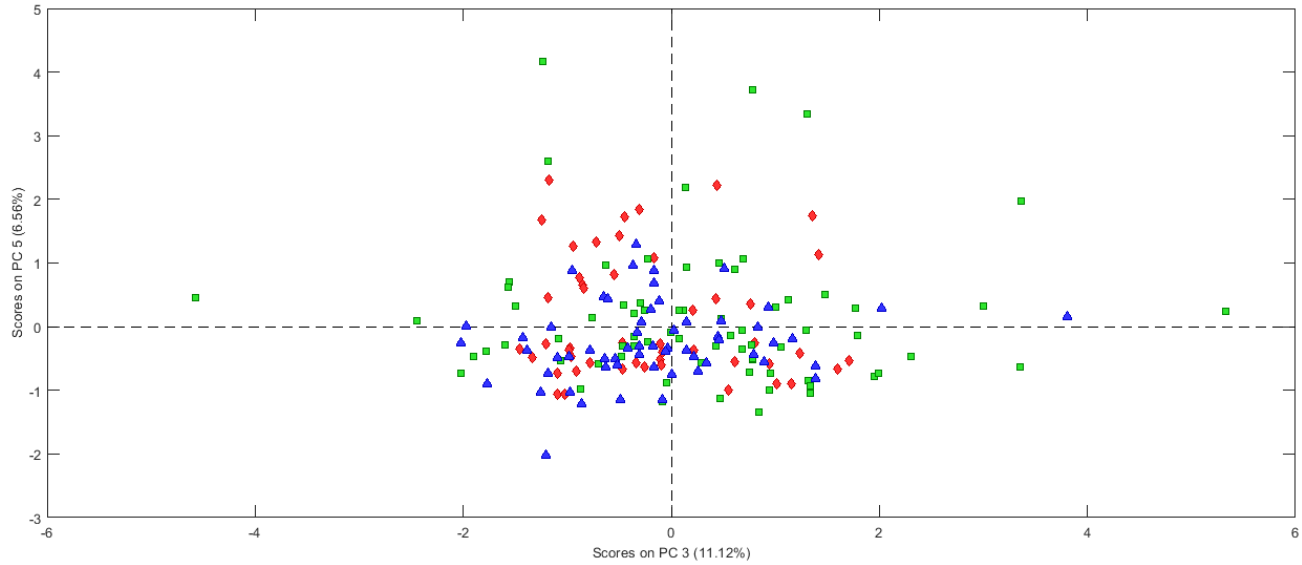


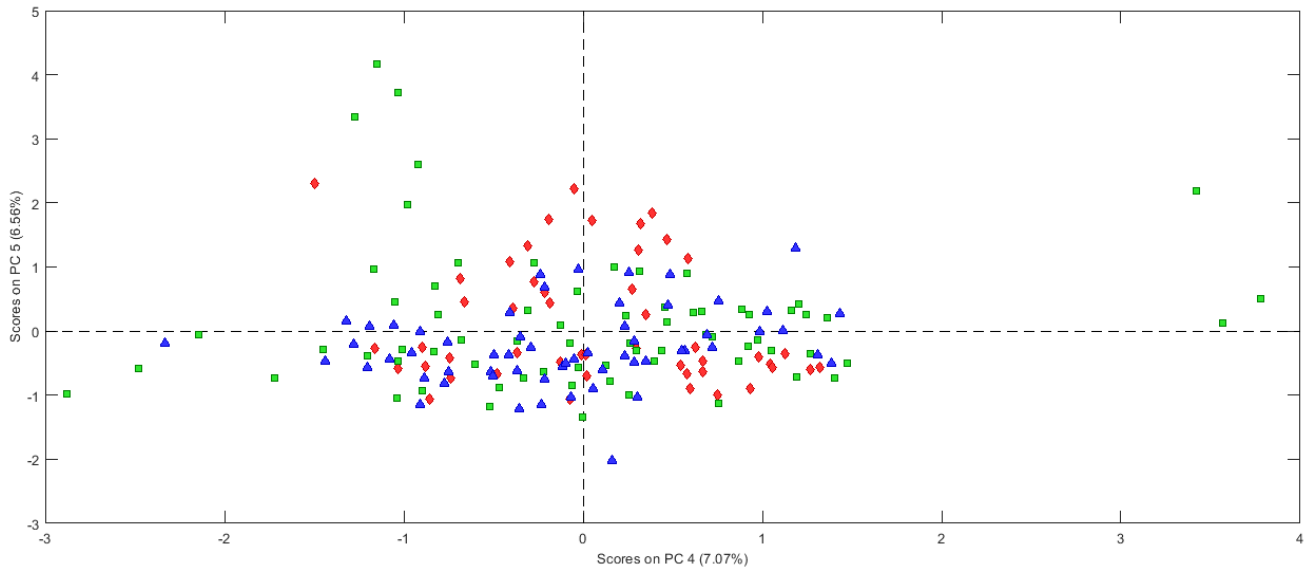
Con il plot PC2-PC4 si può sempre raggruppare il vino GR sulla seconda componente principale come prima ma i dati risultano meno compatti sulla quarta componente, in maniera non organizzata.



Per ultimi da osservare, i grafici PC3-PC4, PC3-PC5 e PC4-PC5 non rendono distinguibile nessun sottogruppo.

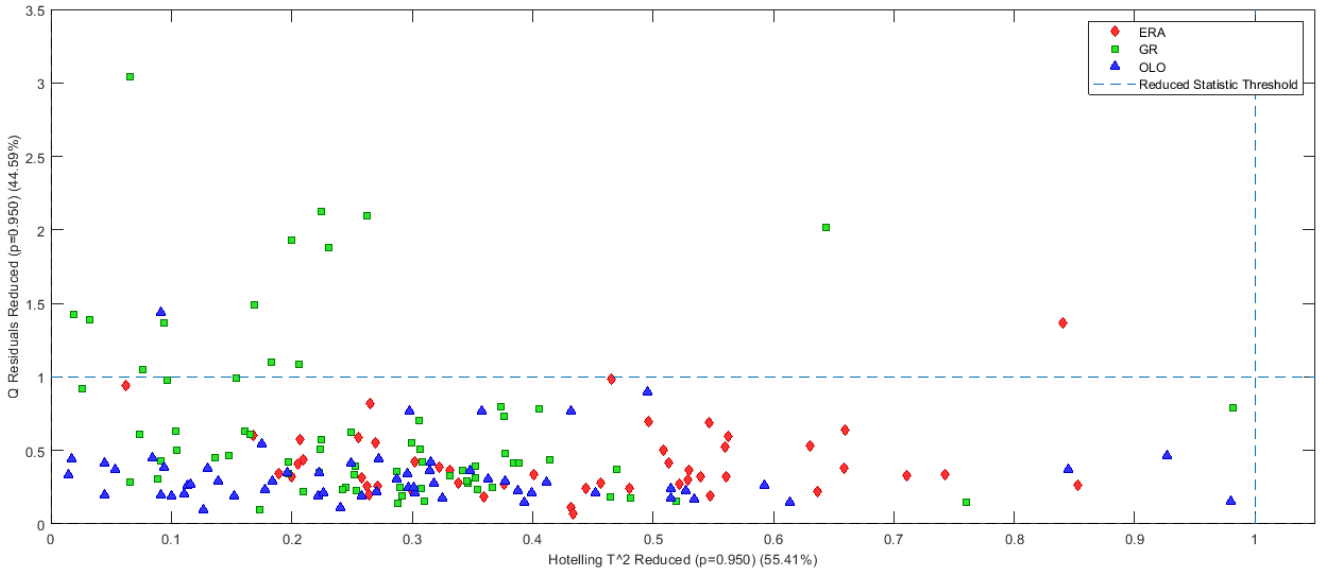




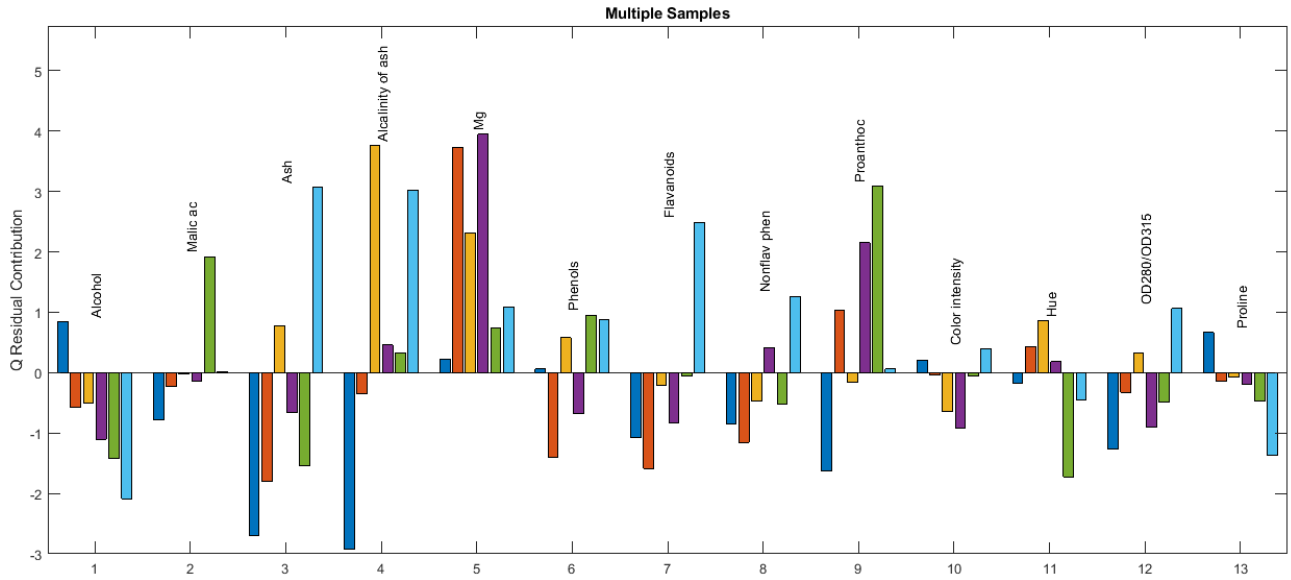


Da questi ultimi risultati, si può affermare che due componenti principali sono sufficienti a descrivere ciò che è il nostro obbiettivo, ovvero separare le categorie.

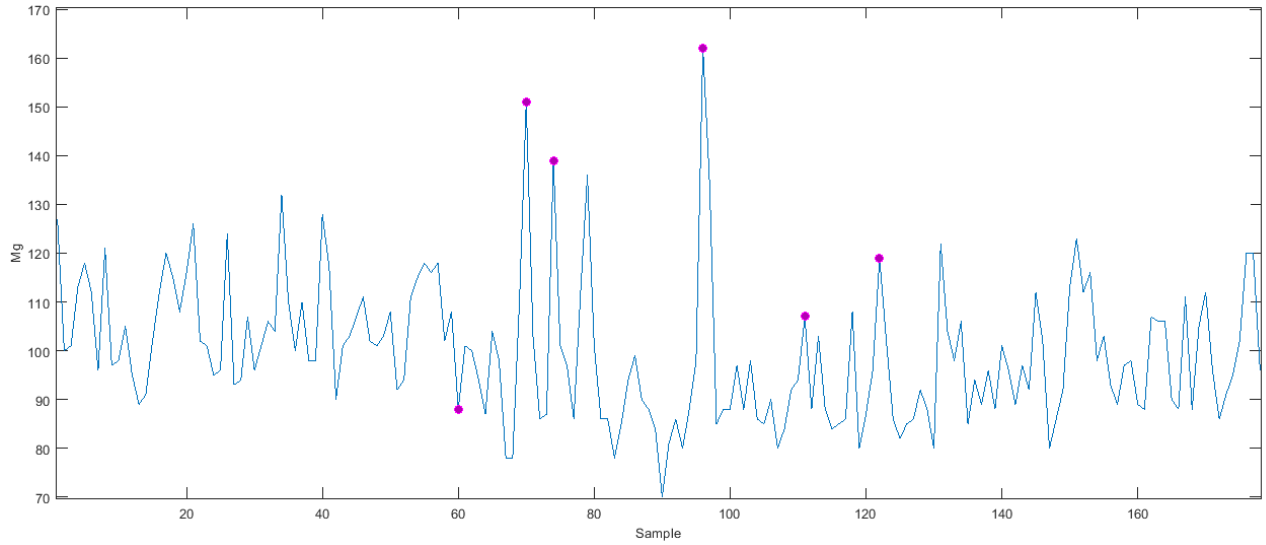
Selezionando le componenti sul nuovo numero scelto (ovvero due) si ritorna a fare il grafico degli score ma questa volta per analizzare la presenza di eventuali campioni outlier, estremi o con residui molto alti.



Da qua si può notare che quelli con un Q elevato sono principalmente GR. Analizzando meglio i sei campioni più in alto si può affermare che hanno valori elevati di Mg principalmente più qualche altra variabile sparsa, come il Proanthoc.



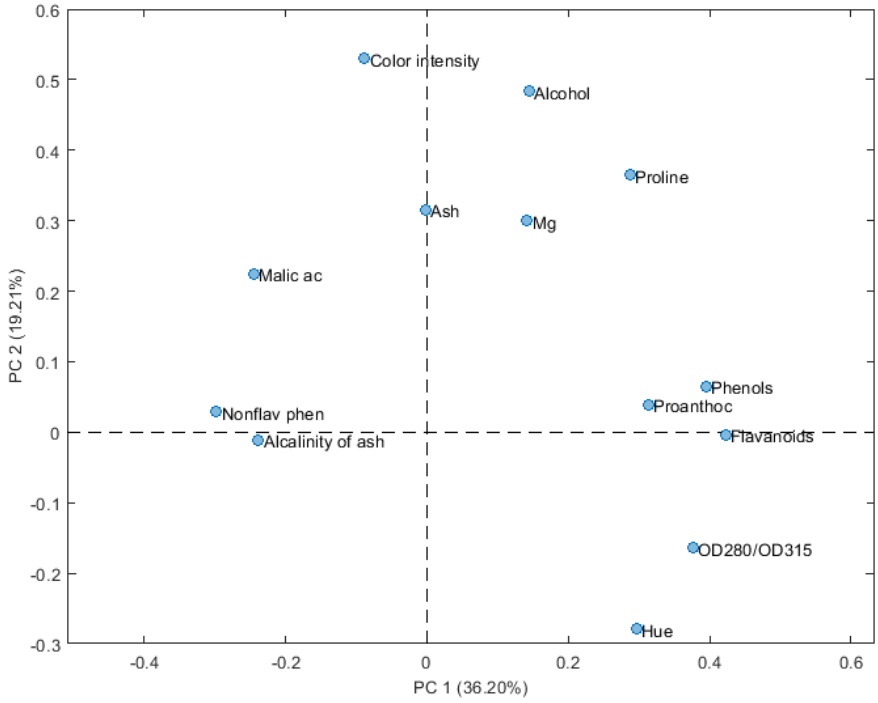
Tracciando i singoli grafici di Mg, si può sottolineare il fatto che metà dei campioni ha valori che superano di poco il normale range.



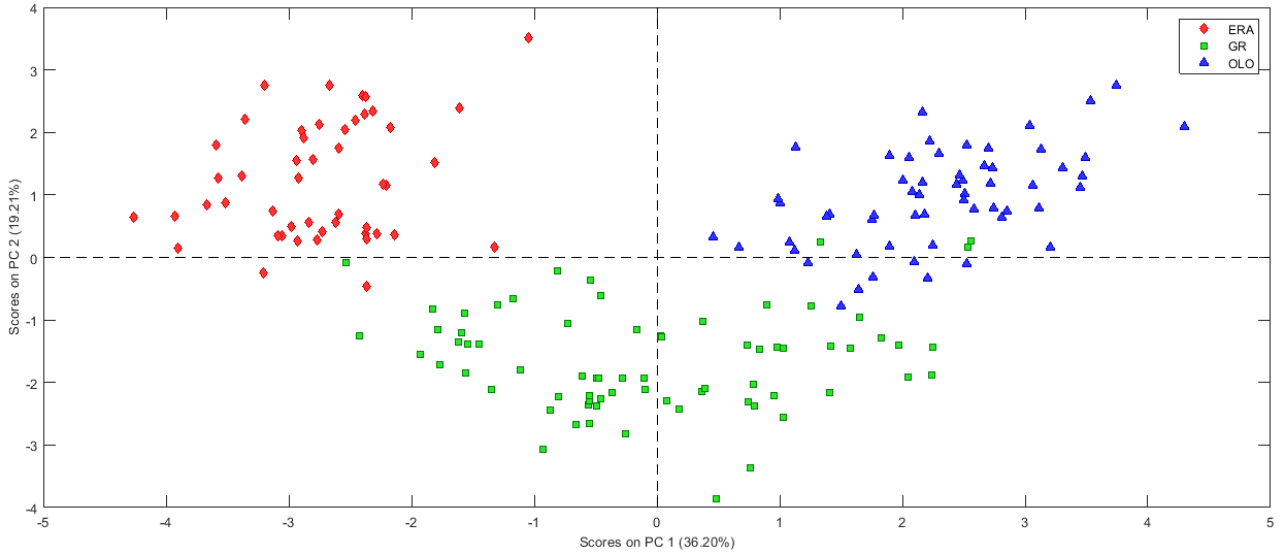
Tuttavia, essendo estremi in Q non mi cambia l'andamento delle componenti principali quindi non si ritiene necessario rimuovere i punti appena descritti. Quelli che potrebbero perturbare sono gli estremi in T^2 e gli outliers che fortunatamente non ci sono.

Ora è il momento di confrontare i loadings con gli score.

- Loadings: presenta correlazioni tra Nonflav phen e Alcalinity of Ash e tra Phenols, Proanthoc e Plavanoids.



- Score (era già stata mostrata, ma è stata rimessa a motivi di praticità nel confronto)

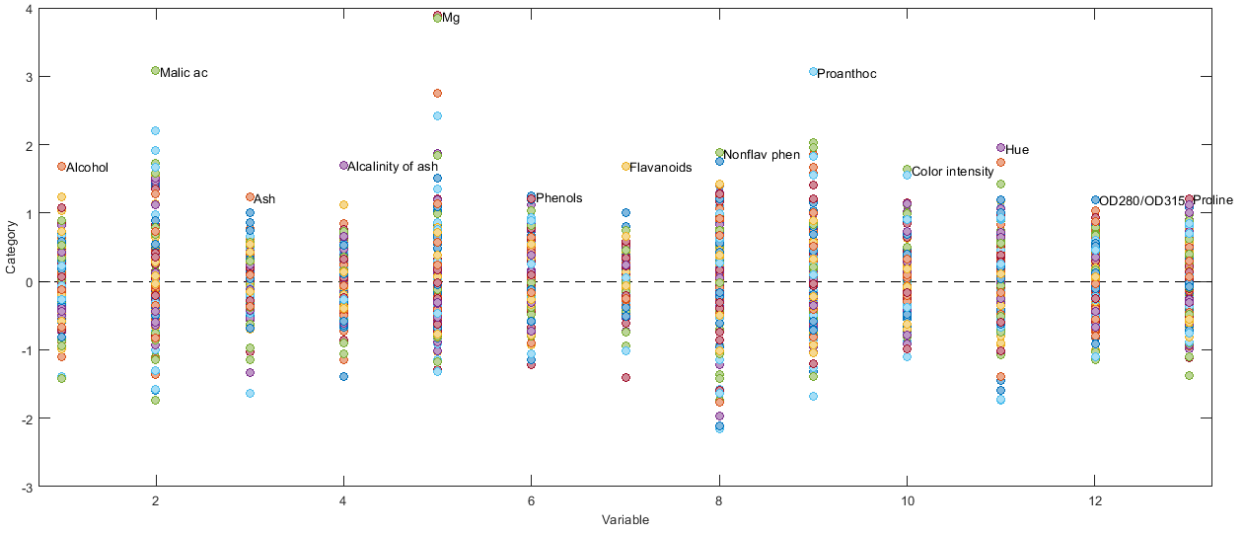


Il vino OLO che è situato nel quadrante in alto a destra (quindi influenzato da PC1 e PC2) presenta maggiori quantità di Phenols, Proanthoc e Flavanoids e minori quantità di Alcalinity of ash e Nonflav phen rispetto agli altri vini.

Il vino GR è influenzato principalmente da PC2 quindi avrà maggiori quantità di Hue, OD280/OD315, Alcalinity of ash e Flavanoids e minori quantità dei rimanenti.

Il vino ERA che è situato nel quadrante in alto a sinistra (come per l'OLO è influenzato da PC1 e PC2) presenta maggiori quantità di Maliac ac, Nonflav phen e Alcalinity of ash mentre minori quantità di OD280/OD315 e Flavanoids.

Infine, si vuole capire se i valori delle variabili sono distribuiti in modo normale e per fare ciò si crea il plot dei residui.



Si può affermare, fatta eccezione per qualche punto in alcune variabili, che la distribuzione è approssimativamente normale.

Concludendo, i risultati più importanti traibili da questa relazione sono i seguenti:

- Sono sufficienti due componenti principali ai fini di distinguere le categorie di vino;
- PC1 è ottimo a distinguere tutti e tre i vini, mentre PC2 distingue accuratamente solo il GR;
- Non ci sono campioni da rimuovere e non ci sono estremi in T^2 ed outliers;
- La distribuzione dei residui è approssimativamente bilanciata;
- Vi sono correlazioni tra Nonflav phen e Alcalinity of Ash e tra Phenols, Proanthoc e Plavanoids;
- Il vino OLO presenta valori più alti rispetto ad altri vini di Phenols, Proanthoc e Flavanoids e minori quantità di Alcalinity of ash e e Nonflav phen;
- Il vino GR presenta valori più alti rispetto ad altri vini di Hue, OD280/OD315, Alcalinity of ash e Flavanoids e minori dei rimanenti;
- Il vino ERA presenta valori più alti rispetto ad altri vini di Maliac ac, Nonflav phen e Alcalinity of ash mentre minori quantità di OD280/OD315 e Flavanoids.

Preprocessing

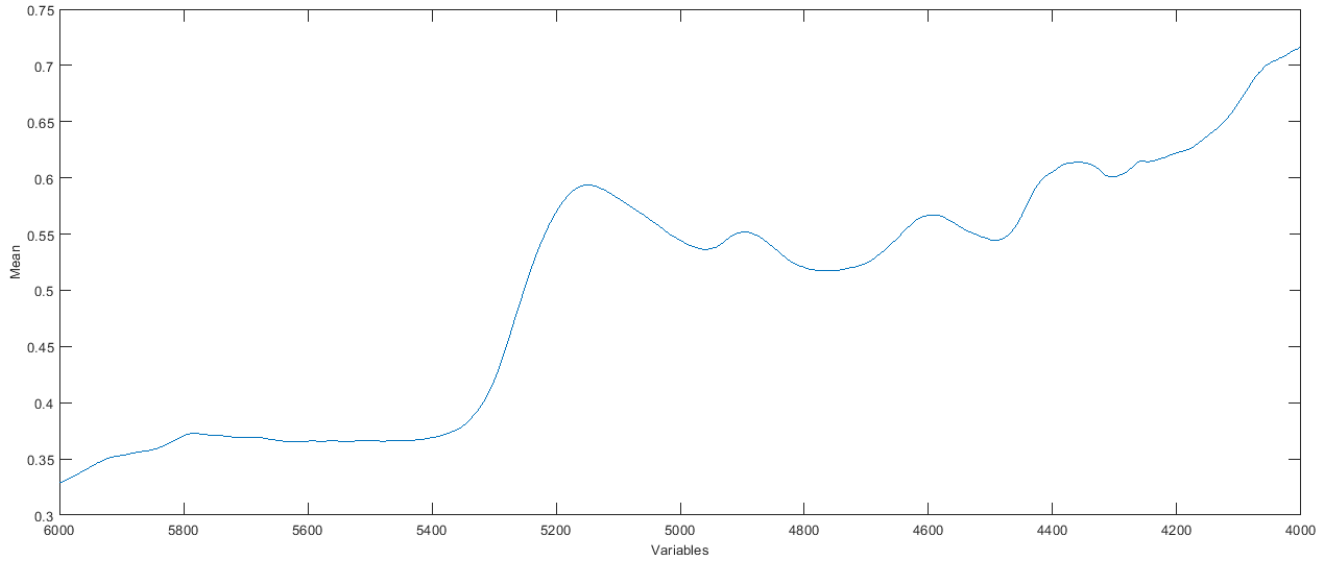
Relazione di Francesco Malferrari

In questa relazione l'obiettivo è trovare un preprocessing tale da rendere i vari tipi di farina riconoscibili sulle prime componenti principali. Le farine possono essere di:

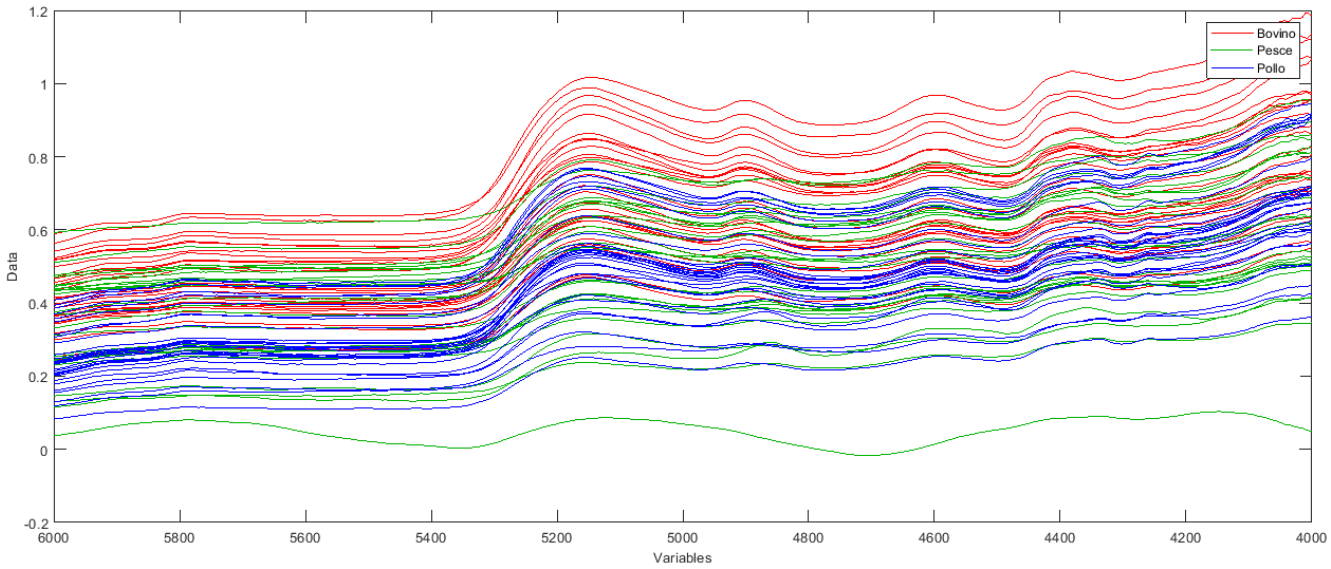
- Bovino (28)
- Pesce (26)
- Pollo (30)

I dati sono stati acquisiti attraverso una riflettanza diffusa di 16 scan con il metodo FT-NIR da 4000 a 6000 cm^{-1}.

Innanzitutto, è importante vedere il plot dei dati. Il primo mostrato rappresenta la media del segnale.

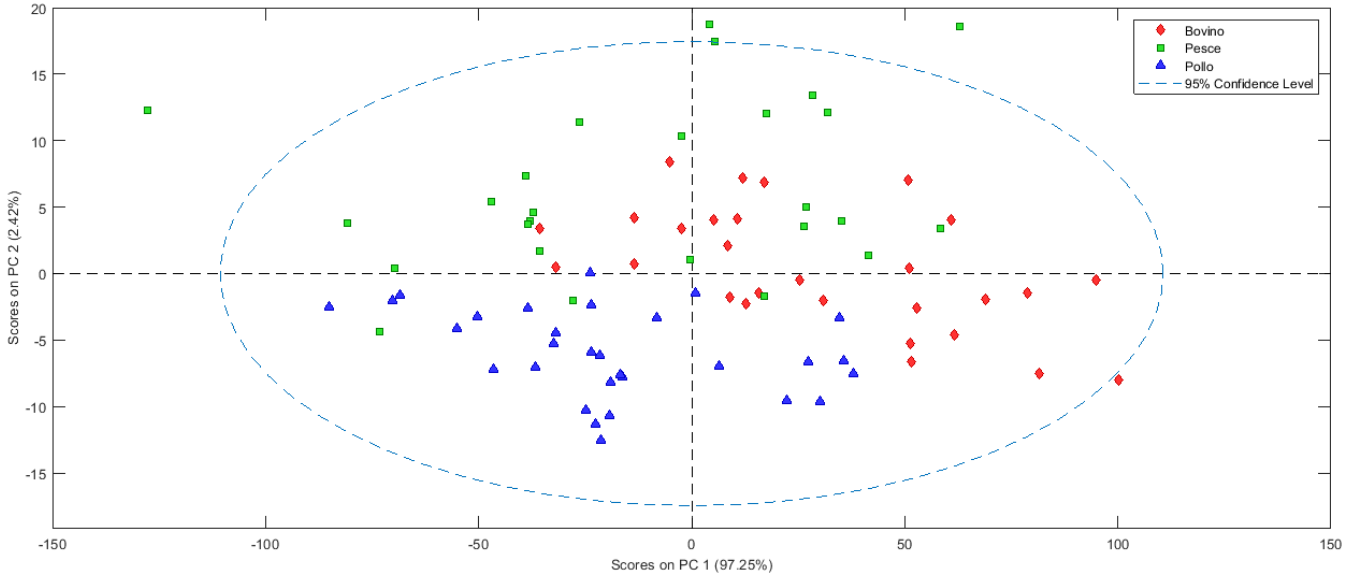


Invece questo che segue è l'insieme di tutti i dati, che mostra in generale una linea di base ma con degli shift, tranne per quello più in baso della categoria Pesce.



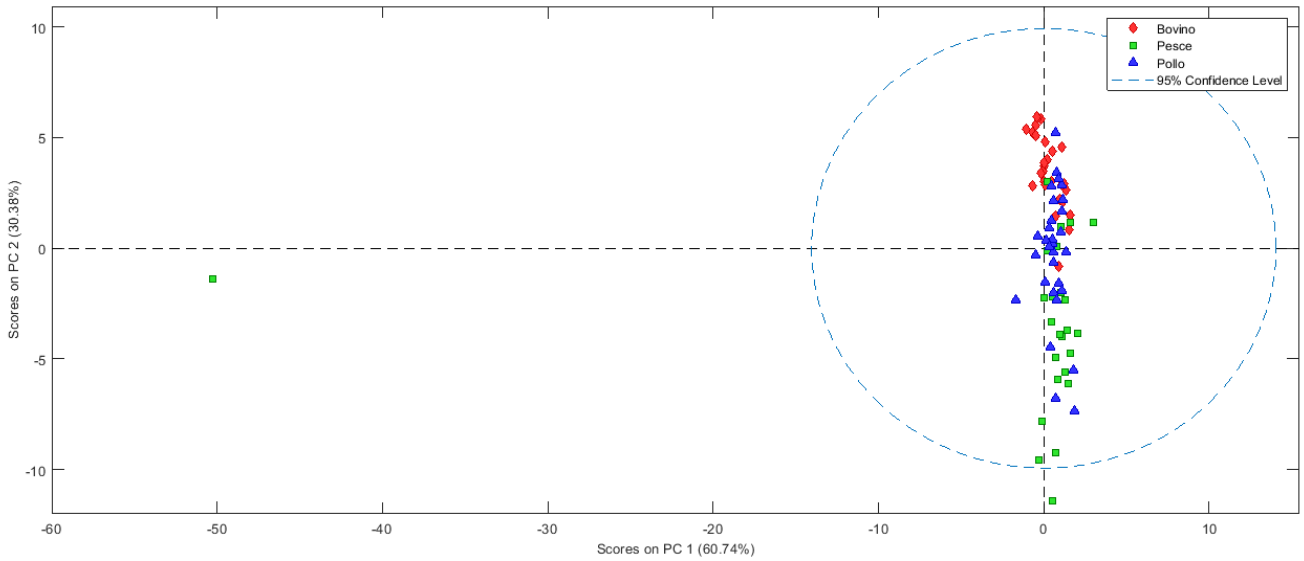
Sono stati fatti diversi tentativi per determinare il preprocessing migliore per questo scopo:

- Mean centering



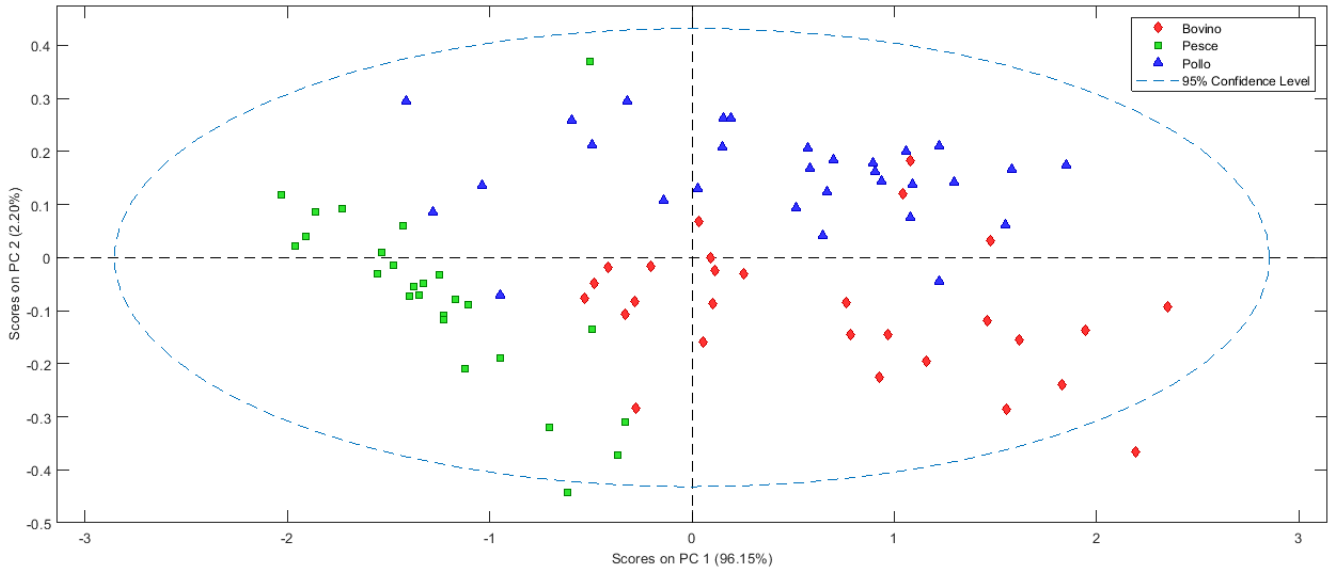
Le farine di Pollo sono abbastanza isolate, mentre le altre due categorie sono un po' mischiate ma comunque approssimativamente distinguibili;

- Normalizzato con l'algoritmo SNV + mean centering.



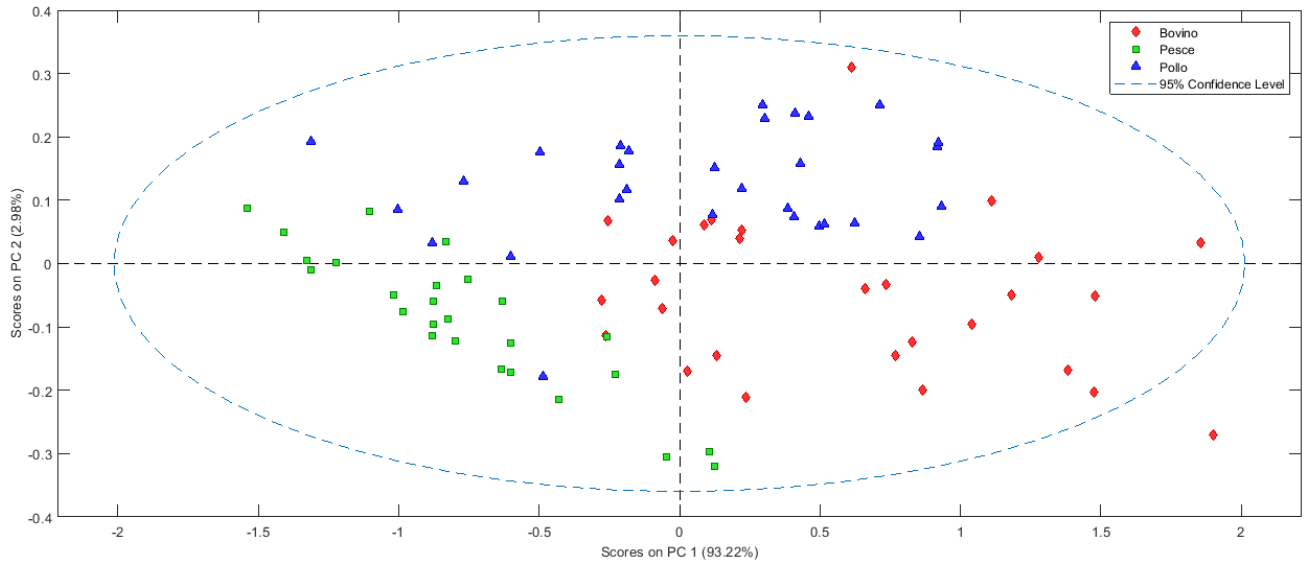
Le componenti principali non sono in grado di separare le farine, tranne per quelli più in basso di Pesce e per quelli più in alto di Bovino;

- Baseline di ordine 5 + mean centering.



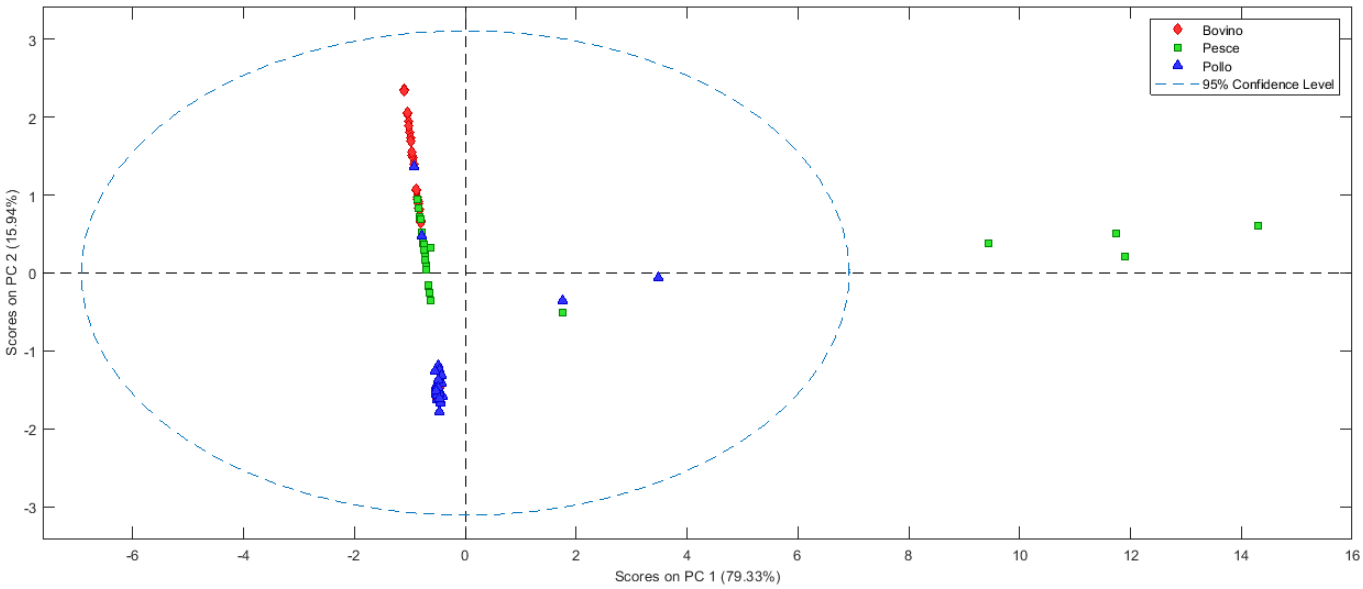
Le componenti principali sono in grado di separare, salvo qualche singolo punto sparso, le varie tipologie di farina;

- Baseline di ordine 6 + mean centering.



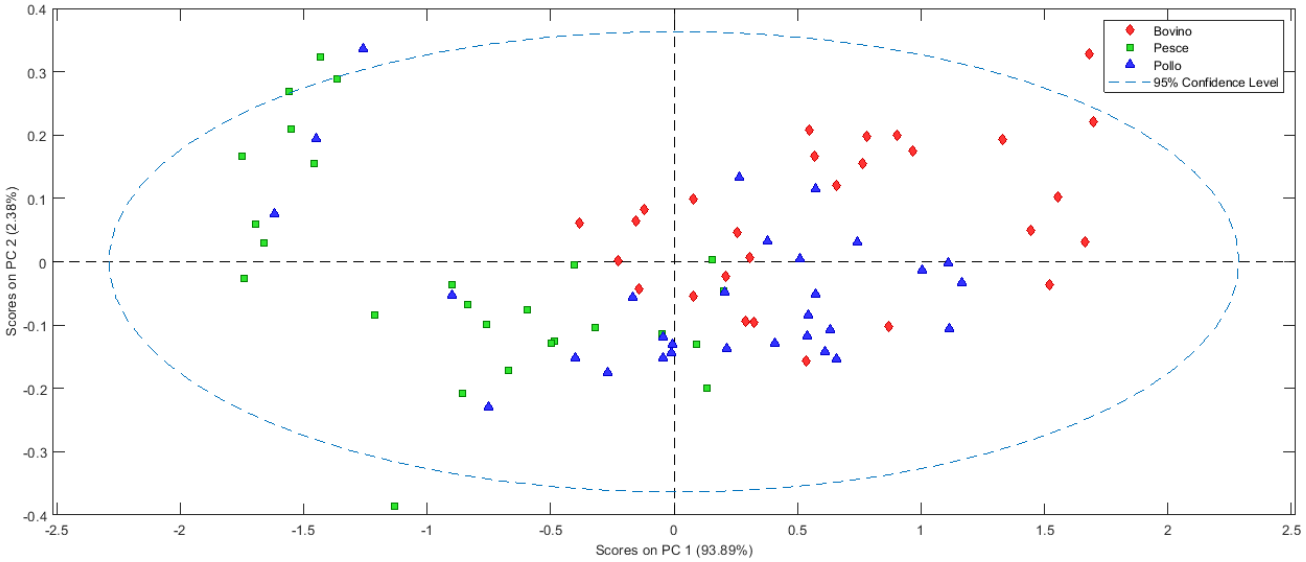
Simile al caso precedente, forse leggermente più impreciso nella separazione di qualche punto;

- Baseline di ordine 7 + mean centering.



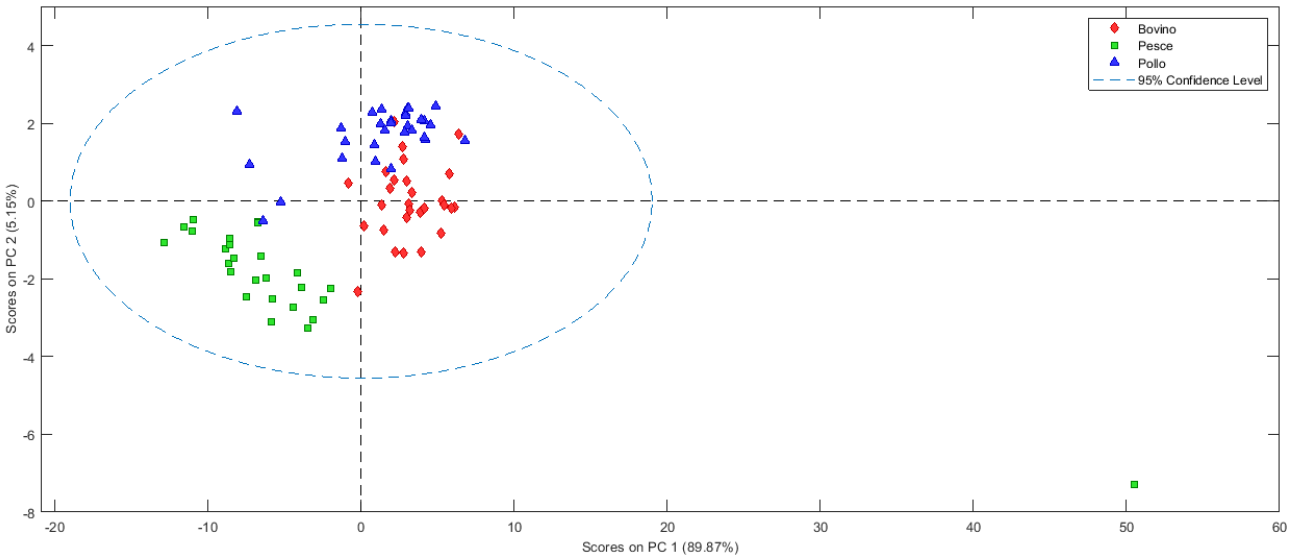
I punti sono troppo sovrapposti per poterli separare accuratamente, salvo quelli del Pollo;

- Baseline di ordine 8 + mean centering.



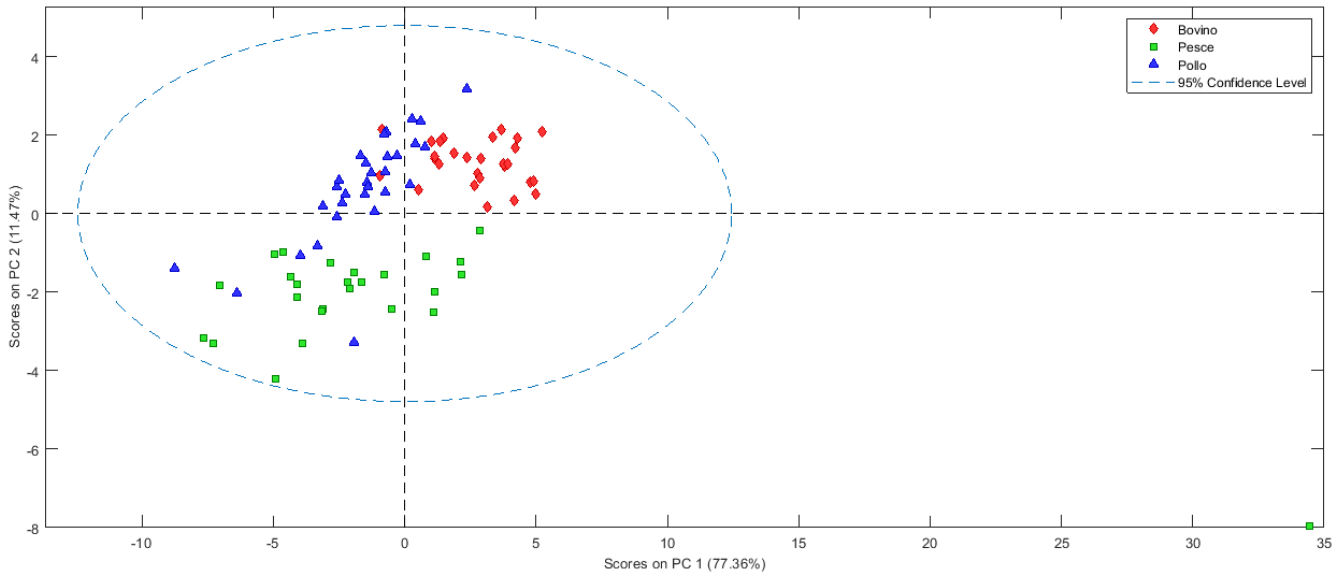
I punti sono troppo sparsi per isolare le categorie in maniera accettabile;

- SNV + Baseline di ordine 5 + mean centering.



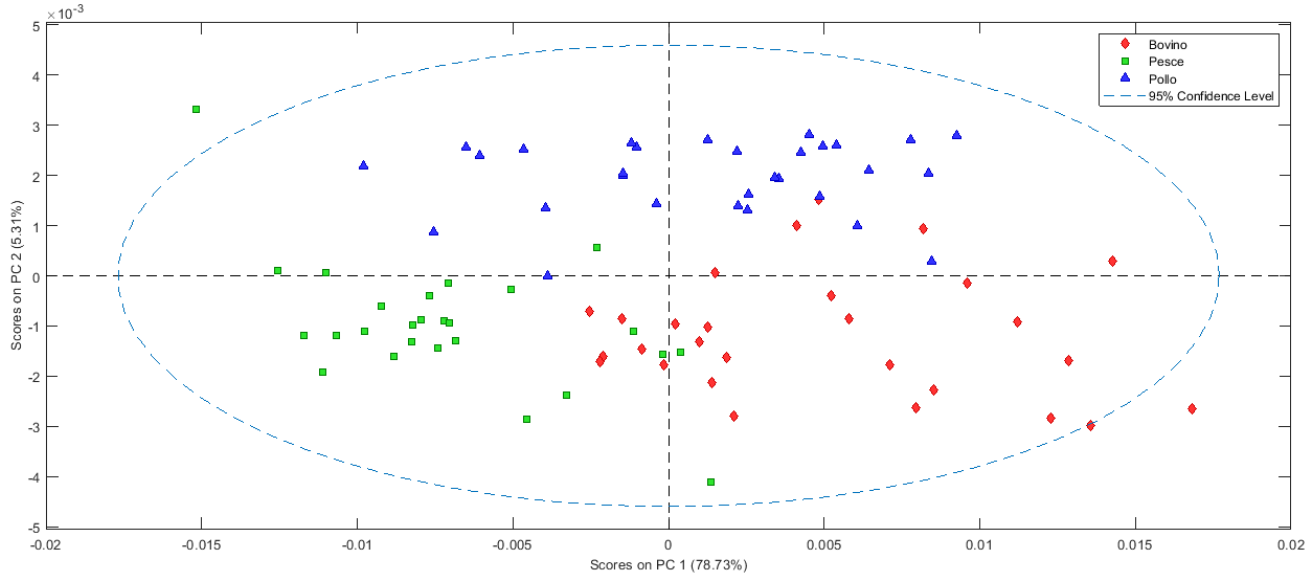
I punti sono separabili abbastanza bene, tranne qualche punto rosso nel blu. Inoltre c'è un punto lontano dagli altri della categoria Pesce, che comunque sembra non influire sulle componenti principali;

- SNV + Baseline di ordine 6 + mean centering.



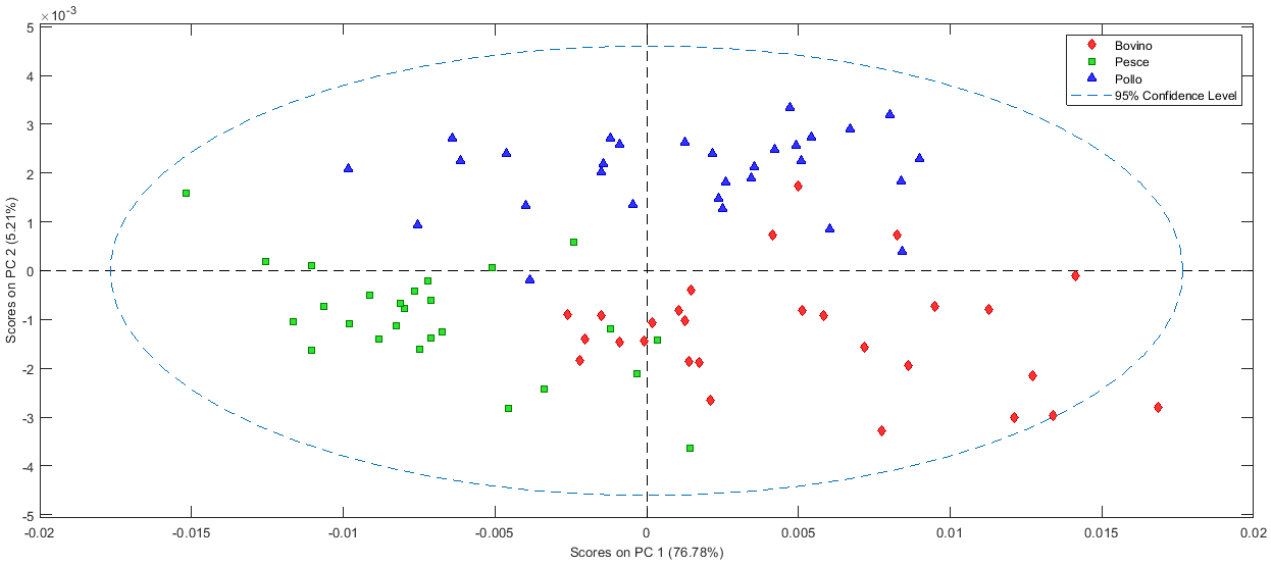
Simile al caso precedente ma un po' meno precisa la separazione;

- Smoothing (grado 1, window 61) + mean centering.



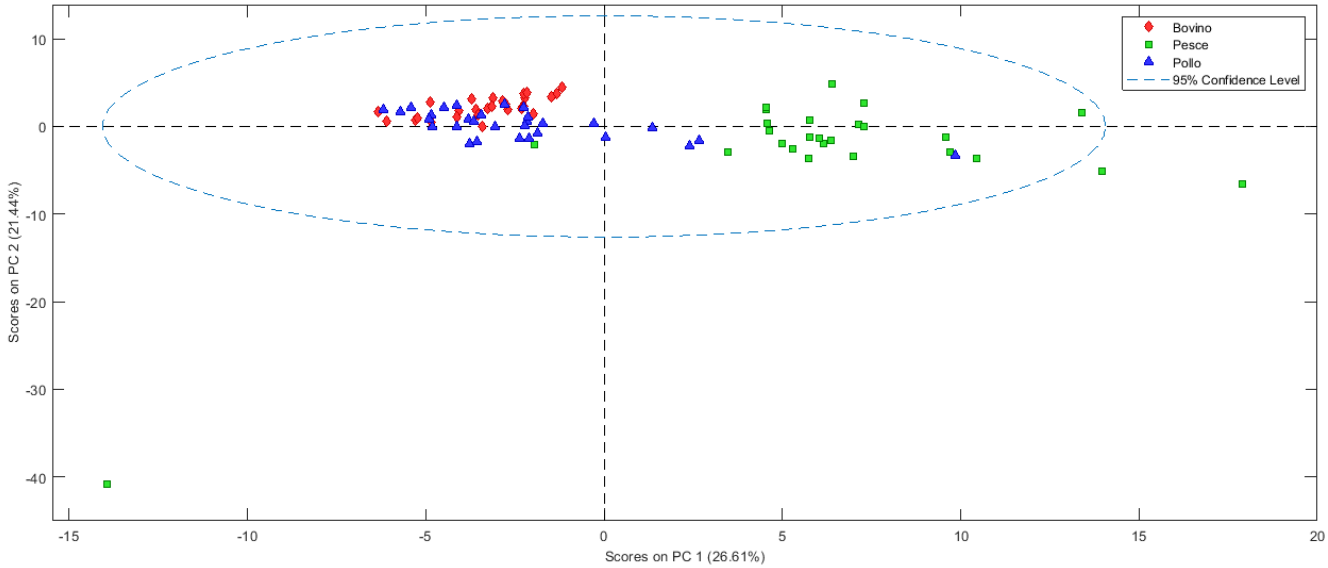
Le componenti principali sono in grado di separare i punti in maniera abbastanza accurata, salvo qualche punto "di frontiera";

- Smoothing (grado 2, window 61) + mean centering.



Molto simile al caso precedente;

- Smoothing (grado 2, window 61) + SNV + mean centering.



Le categorie Bovino e Pollo sono indistinguibili, Pesce invece sì anche se ha diversi punti lontani.

In base ai risultati appena elencati, si ritiene che il preprocessing più adatto a questo contesto il Baseline di ordine 5 + mean centering.