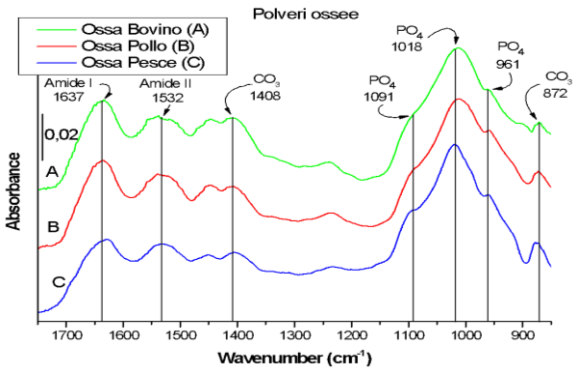
# Classification farine animali

Relazione di Francesco Malferrari

Il set di dati preso in analisi è composto da:

- Un set di calibrazione composto da 133 campioni;
- Un set di test composto da 37 campioni.




Questi dati rappresentano vari campioni di mangimi con fluoro composti da bovini, polli o pesci e sono rappresentati attraverso i valori campionati in 416 lunghezze d'onda (in $cm^{-1}$).

Gli spettri sono stati acquisiti dopo la calcinazione del campione.

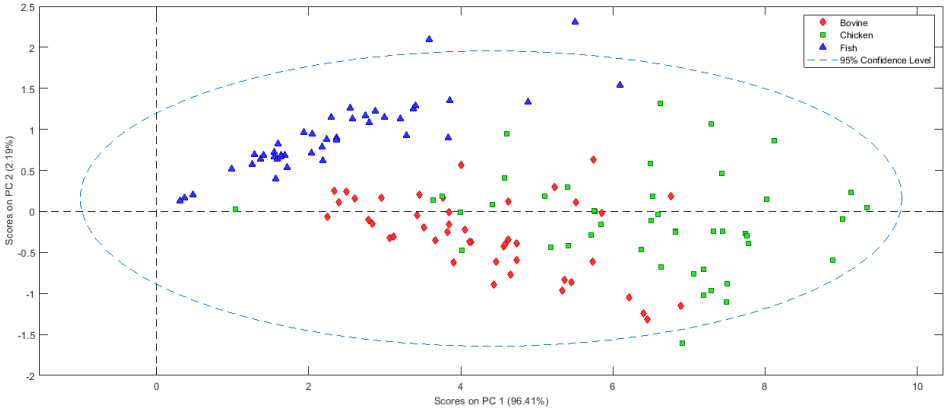
Riuscire a classificare correttamente i mangimi è di estrema importanza, dato che il bovino è vietato nell'alimentazione degli animali, il pesce è consentito mentre il pollo dipende.

Lo scopo è realizzare modelli adeguati di classificazione supervisionata e per fare ciò verranno usati due metodi:

- SIMCA (Soft Independent Modelling of Class Analogy);
- PLS/DA (Partial Least Squares Discriminant Analysis).

Ma prima di procedere con questi algoritmi, bisogna capire se è necessario eseguire un preprocessing sui dati e nel caso affermativo di trovare il giusto metodo che distingui per bene le tre classi nel miglior modo possibile.

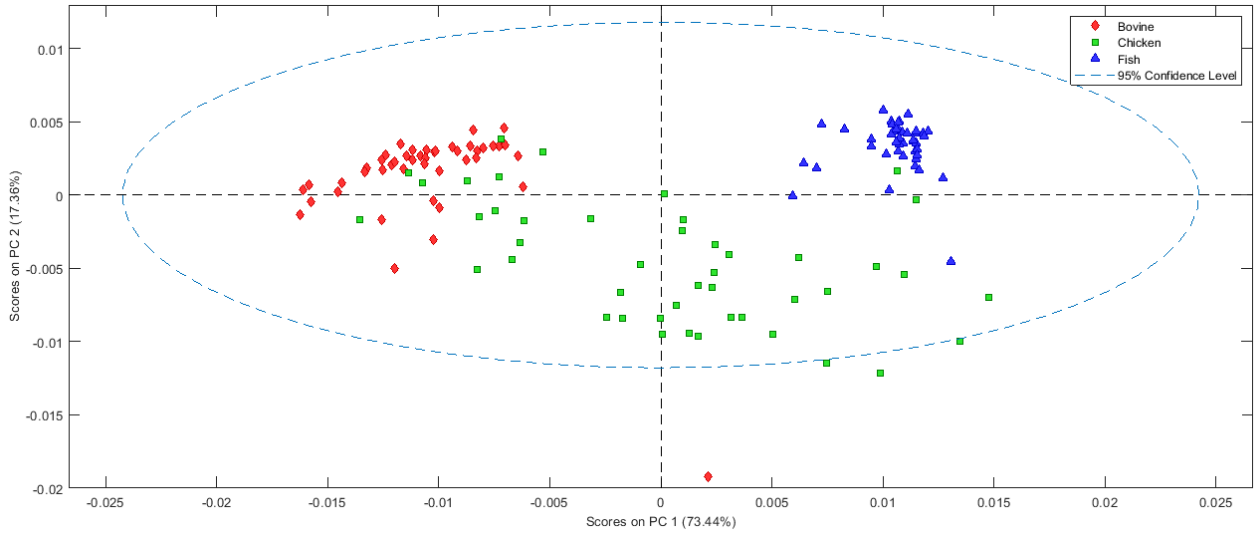
Il grafico degli score dei dati senza preprocessing non distingue bene le classi pollo e pesce.



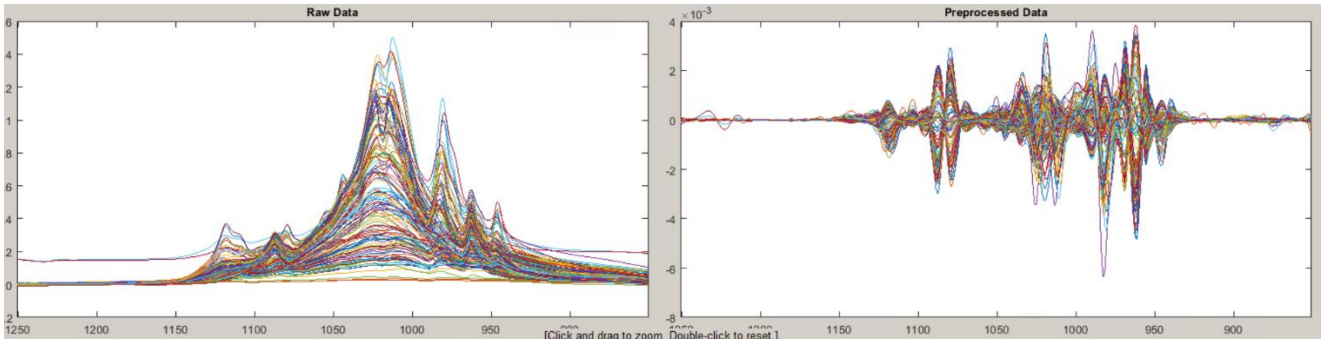
Le varie strategie provate sono le seguenti:

- Mean centering;
- Derivative e mean centering;
- MSC (median) e mean centering;
- 2nd derivative e mean centering;
- Weighted Baseline (order=3) e mean centering;
- Smoothing, MSC (median) e mean centering;
- Derivative, MSC (median) e mean centering;
- Weighted Baseline (order=3), derivative e mean centering;
- Weighted Baseline (order=3), 2nd derivative e mean centering;
- 2nd derivative, MSC (median) e mean centering;

Il risultato che nello score plot ha mostrato la divisione migliore è il 2nd derivative, MSC (median) e mean centering.



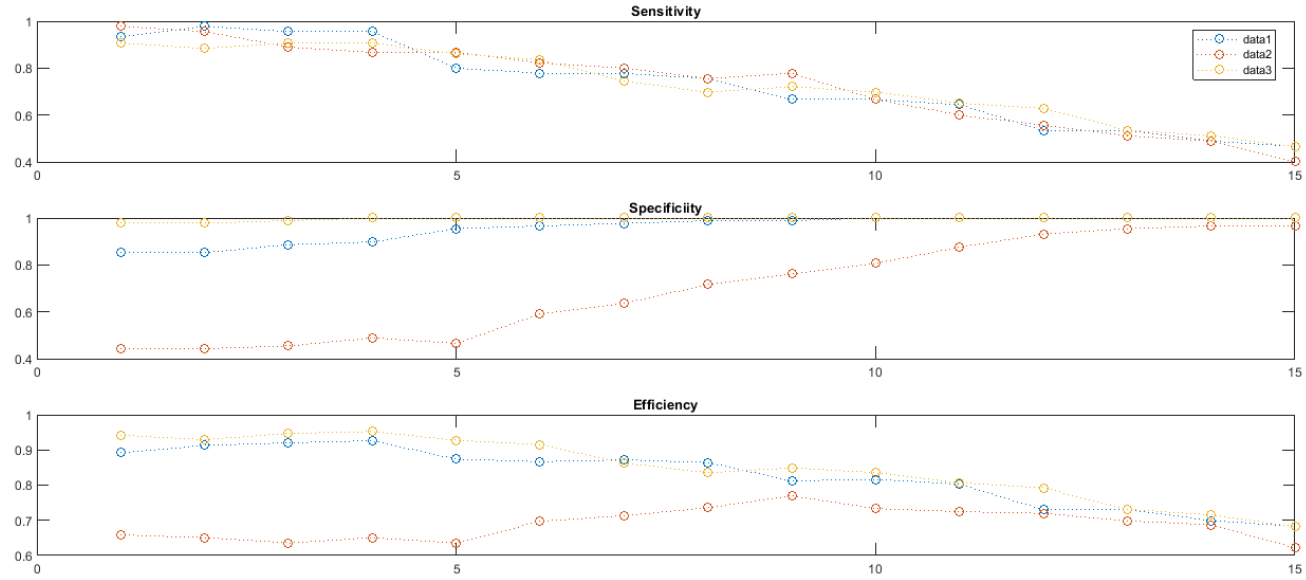

Di seguito a sinistra vi è lo spettro iniziale mentre a destra lo spettro dopo il preprocessing appena descritto.



Per quanto concerne la cross-validazione, per entrambi gli algoritmi di classificazione è stata scelta la tecnica “Venetian blinds”, che consiste in una preliminare divisione del dataset in splits e per ciascuno dei raggruppamenti si porta fuori un campione per volta. Ciò permette il miglioramento della capacità di generalizzazione del modello. In questo caso il numero scelto di splits è 7, quindi si fa la cross-validazione togliendo 19 (133/7) campioni alla volta per un totale di 114 elementi su cui si ricalcola per ogni ciclo fino a che ogni campione non è stato lasciato fuori una volta.

Il numero di componenti principali deve essere ancora deciso, ma per stare larghi si sono presi i primi 15 per SIMCA, mentre per PLS/DA 12 (si era già visto con il precedente che 15 erano troppi).

Il primo algoritmo di classificazione che viene usato è SIMCA che crea un modello PCA separato per ciascuna variabile; perciò, è necessario selezionare individualmente il numero di componenti principali per ciascun modello. Partendo dai tre grafici che mostrano sensibilità (percentuale dei campioni correttamente riconosciuti), specificità (percentuale dei campioni correttamente rifiutati) ed efficienza (media geometrica tra sensibilità e specificità) si è provato a scegliere il numero di componenti principali corrispondenti alla massima efficienza per classe, quindi 4 per la prima, 9 per la seconda e 4 per la terza.




La sensibilità per il training set è molto buona, invece per la specificità la classe 2 nel distinguere la classe 1 sbaglia la metà delle volte circa.

| | 1 Class1 | 2 Class2 | 3 Class3 |
|---|---|---|---|
| 1 Sensitivity | 0.9778 | 1 | 0.9302 |

| | 1 Class1 | 2 Class2 | 3 Class3 |
|---|---|---|---|
| 1 Specificity of C1vs. | 0 | 0.8000 | 1 |
| 2 Specificity of C2vs. | 0.5111 | 0 | 0.9767 |
| 3 Specificity of C3vs. | 1 | 0.9778 | 0 |

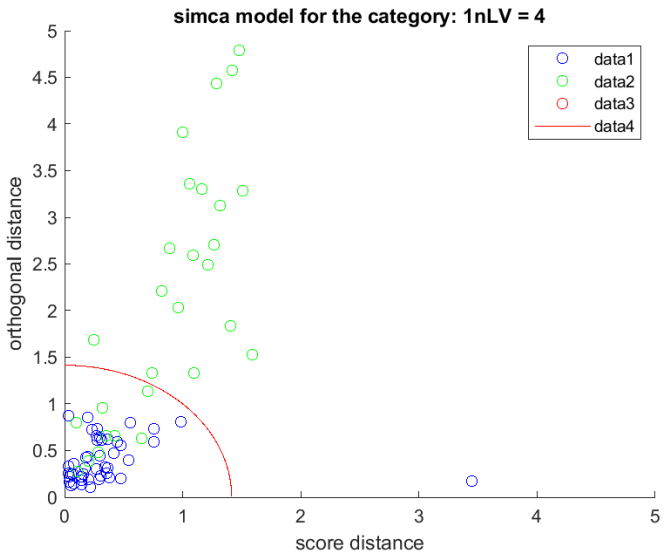
La sensibilità per il training set è buona solo per la prima classe e nelle altre è un po' bassa. Invece le specificità sono tutte molto buone

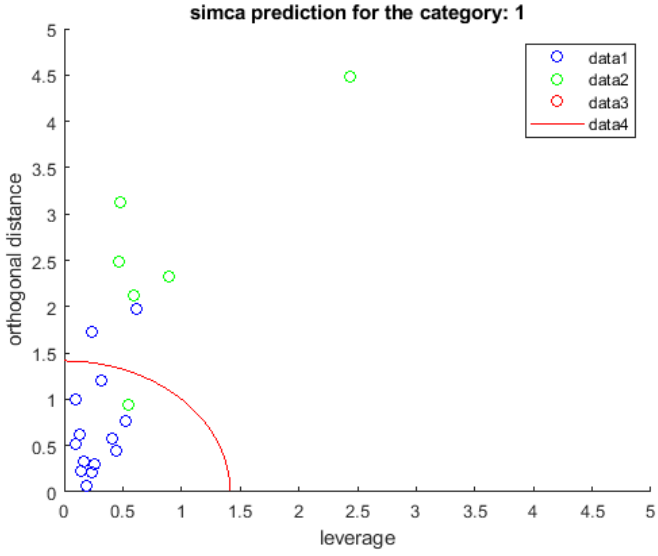
| | 1<br>Class1 | 2<br>Class2 | 3<br>Class3 |
|---|---|---|---|
| 1 Sensitivity | 0.8571 | 0.6923 | 0.6000 |

| | 1<br>Class1 | 2<br>Class2 | 3<br>Class3 |
|---|---|---|---|
| 1 Specificity of C1vs. | 0 | 0.9231 | 1 |
| 2 Specificity of C2vs. | 0.7857 | 0 | 0.9000 |
| 3 Specificity of C3vs. | 1 | 1 | 0 |

Rivedendo il grafico per ogni PC si può notare che è possibile ridurre il numero di componenti principali della terza classe da 4 a 1. I risultati si possono mostrare in grafici con la score distance (SD) e la orthogonal distance (OD), fattori molto analoghi a $T^2$ e Q.

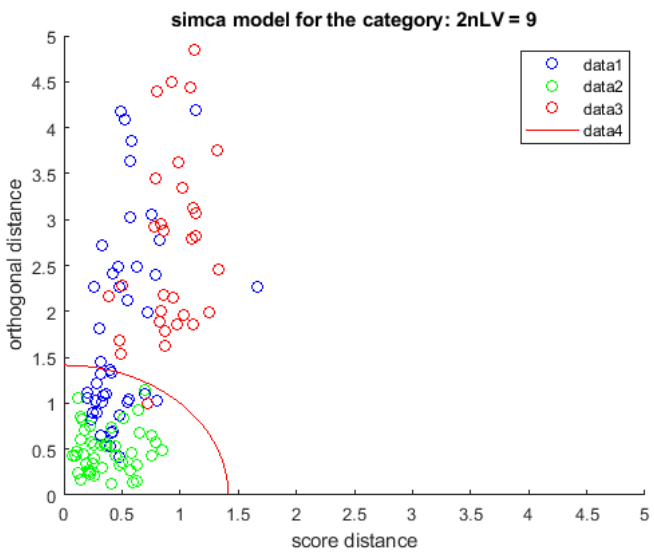
Nei primi di questi si possono vedere a sinistra il training set per la classe 1, mentre a destra il test set per la stessa classe.

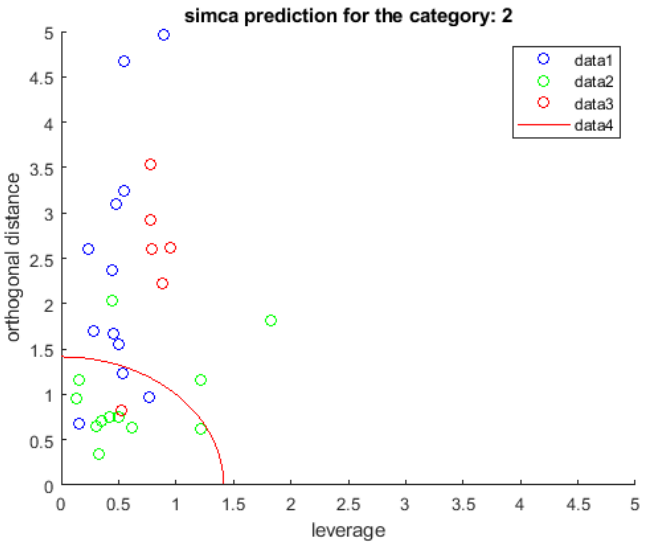


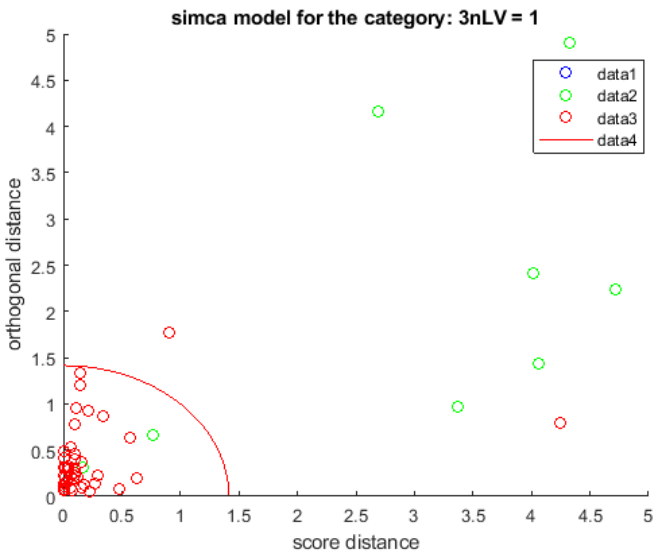

Nel complesso la sensibilità sembra andare bene perché salvo un punto nel training set e due punti del test set il resto dei punti sono accettati. Per la specificità invece nel grafico a sinistra ci sono diversi punti della classe due nella regione d'accettazione e in quella a destra un punto solo.

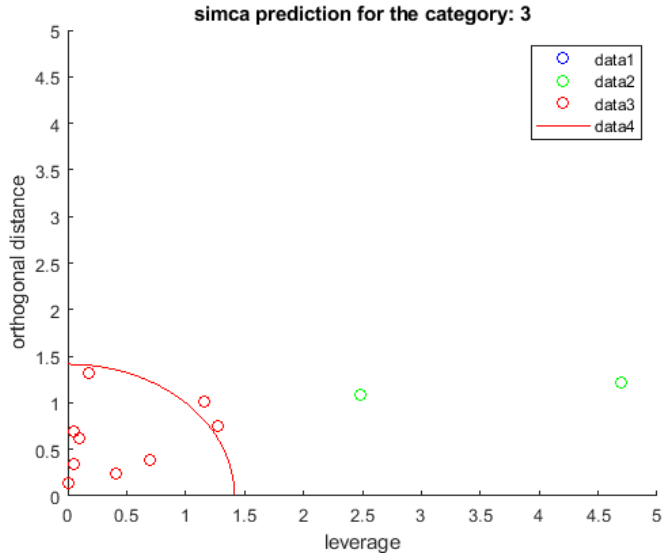
Per quanto riguarda la classe 2 il grafico del training set (a sinistra) presenta diversi punti della prima classe all'interno della regione d'accettazione e uno della terza ma tutti i punti propri visibili sono accettati. Invece il grafico del test set ha tre propri punti non accettati e tre punti della classe 1 e un punto della classe 3 tra gli accettati. Quindi bene per la sensibilità, un po' peggio per la specificità.




Infine per la classe 3 il grafico del training set (sinistra) accetta tutti i campioni suoi tranne due, ma ci sono due elementi della classe 2 all'interno della regione d'accettazione. Nel grafico del test set (destra) invece non vengono accettati campioni di altre classi ma anche due della propria (anche se sono molto vicini).




Per mostrare in numeri quanto descritto in questi grafici si può notare, che rispetto alla scelta precedente delle componenti principali, per il training set è leggermente peggiorata la specificità della classe tre contro la classe due in maniera praticamente indifferente ed è migliorata la sensibilità della classe 3.

| | 1 Class1 | 2 Class2 | 3 Class3 |
|---|---|---|---|
| 1 Sensitivity | 0.9778 | 1 | 0.9535 |

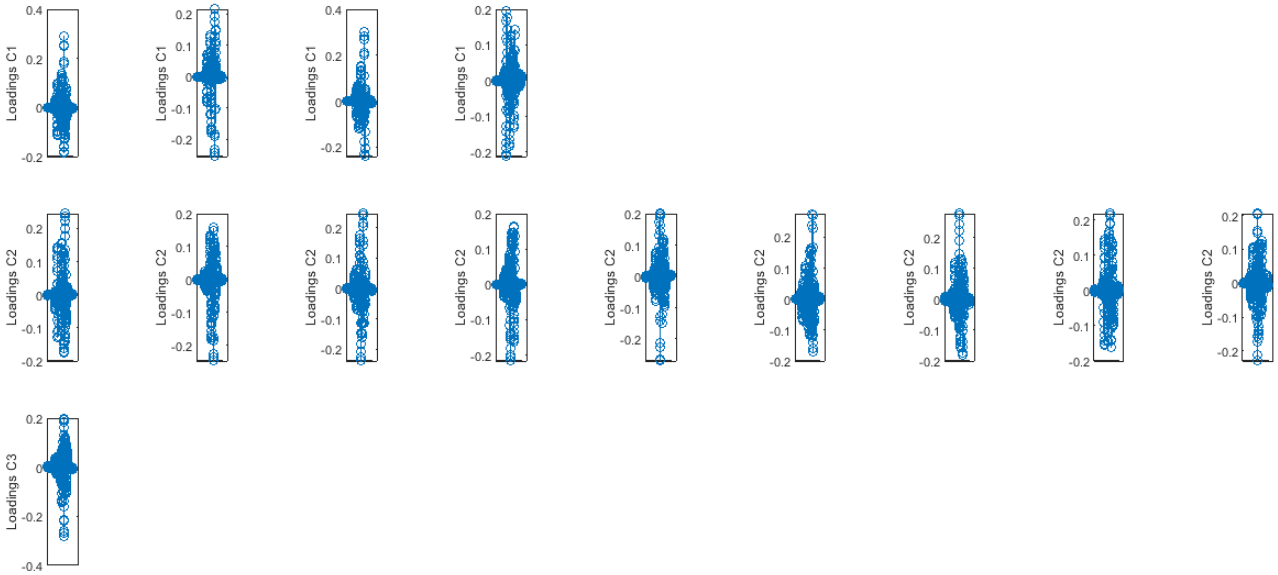
| | 1 Class1 | 2 Class2 | 3 Class3 |
|---|---|---|---|
| 1 Specificity of C1vs. | 0 | 0.8000 | 1 |
| 2 Specificity of C2vs. | 0.5111 | 0 | 0.9767 |
| 3 Specificity of C3vs. | 1 | 0.9556 | 0 |

Mentre per il test set è migliorata di molto la sensibilità della classe 3, passando da 0.6 a 0.8.

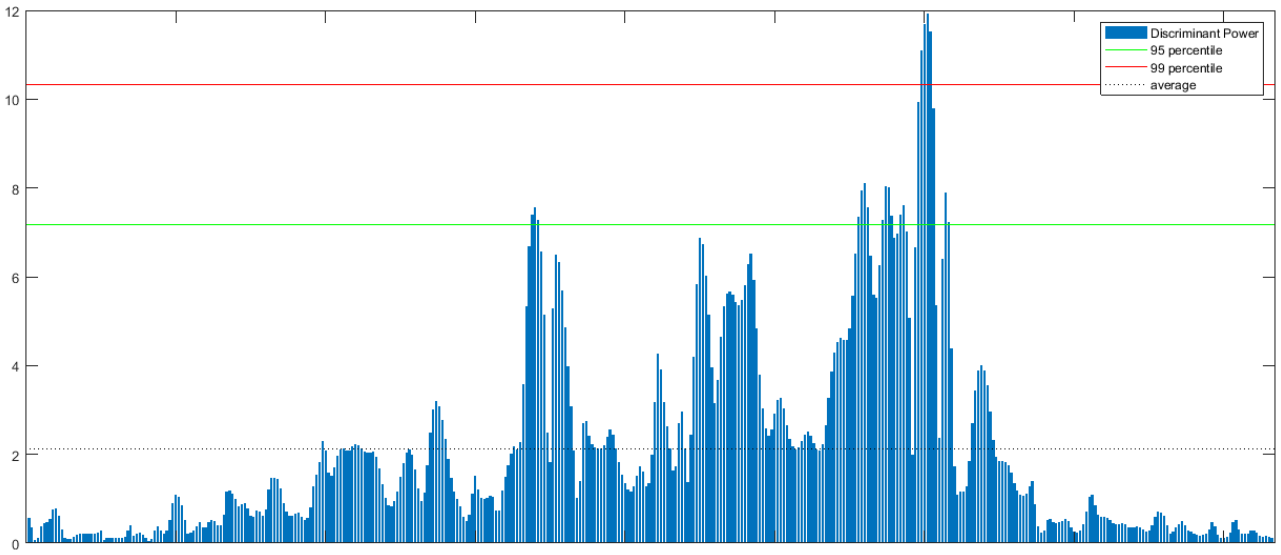
| | 1<br>Class1 | 2<br>Class2 | 3<br>Class3 |
|---|---|---|---|
| 1 Sensitivity | 0.8571 | 0.6923 | 0.8000 |

| | 1<br>Class1 | 2<br>Class2 | 3<br>Class3 |
|---|---|---|---|
| 1 Specificity of C1vs. | 0 | 0.9231 | 1 |
| 2 Specificity of C2vs. | 0.7857 | 0 | 0.9000 |
| 3 Specificity of C3vs. | 1 | 1 | 0 |

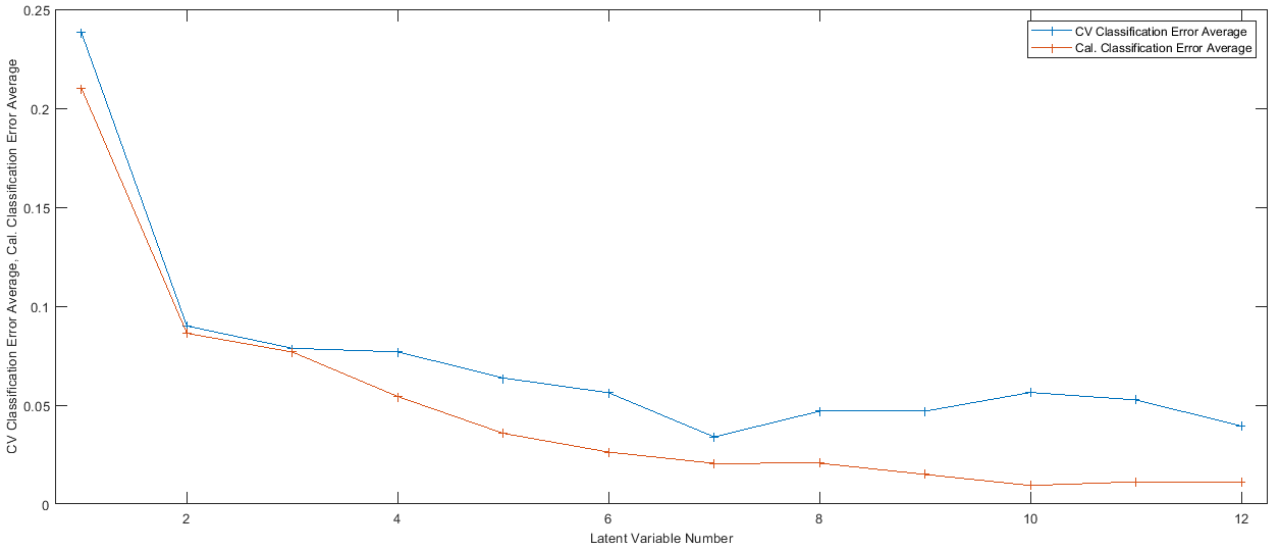
Questi sono i grafici dei loadings che permettono di vedere le varie correlazioni tra le variabili.




Un altro grafico importante a questo scopo è del potere discriminante. Questo parametro è calcolato solo sulla base dei residui. Se una variabile è importante per una classe questa è spiegata bene dalla sua pca, quindi i residui saranno abbastanza piccoli e viceversa se una variabile non è importante avrà residui abbastanza alti perché partecipa poco al modello.

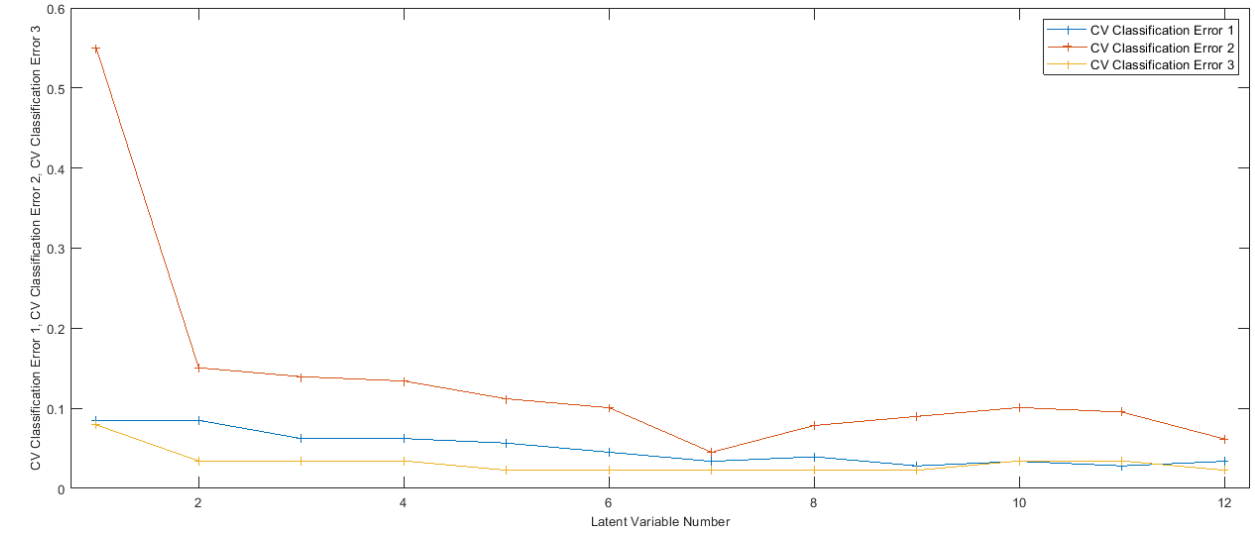


Finito con SIMCA, si continua usando PLS/DA dove bisogna selezionare un unico numero di componenti principali per tutte le classi.



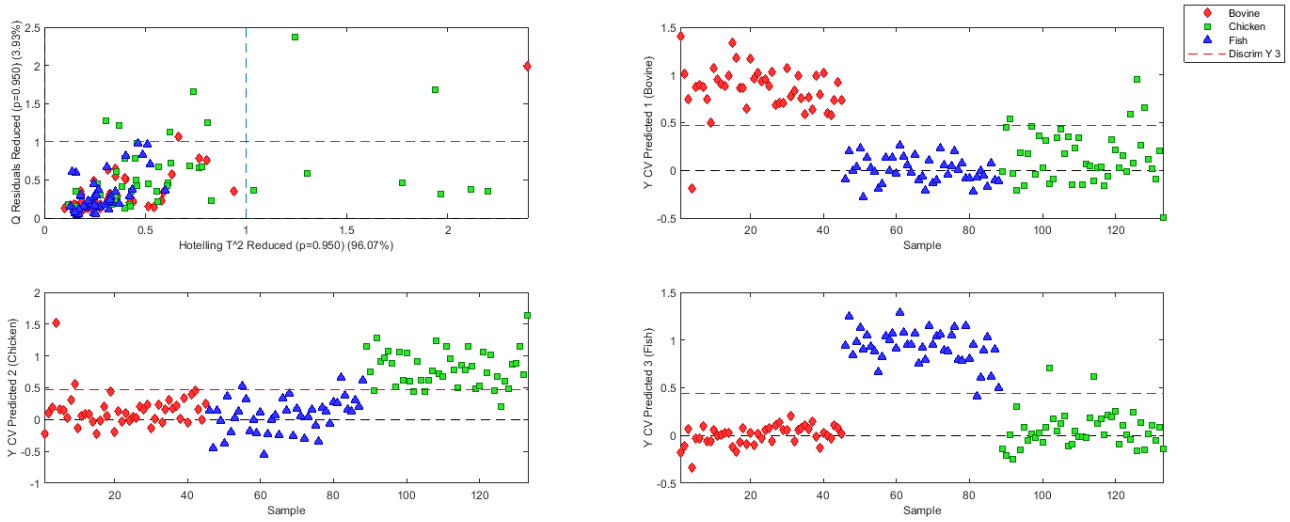

Dal grafico con la media degli errori medi in calibrazione e in cross-validazione si può vedere come questi ultimi si abbassano con 7 componenti per poi tornare a salire.

Per maggiore precisione si possono anche vedere gli errori singolarmente per classe.

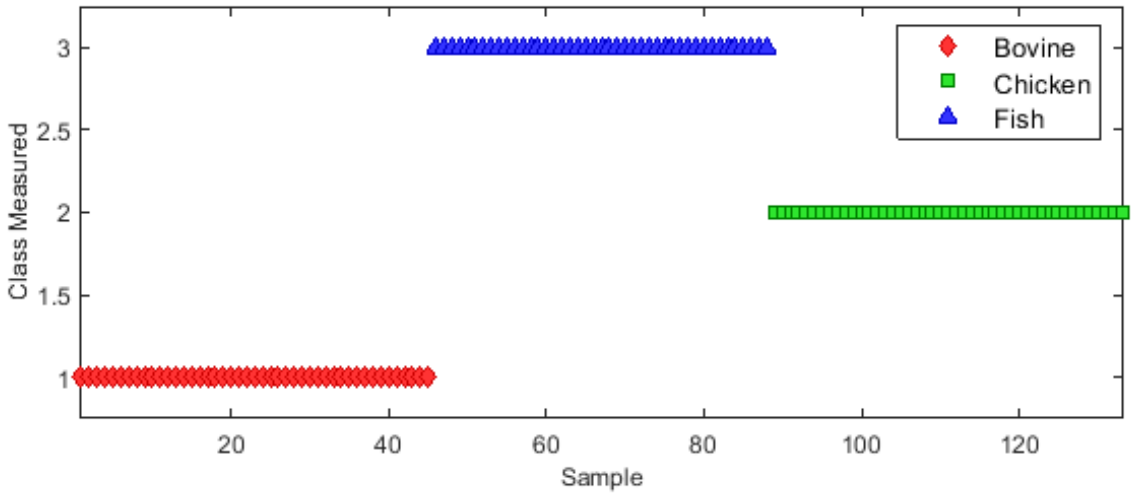



Si conferma quanto affermato dato che con sette componenti si raggiunge un punto basso per la seconda classe, mentre le altre due sono abbastanza stabili.

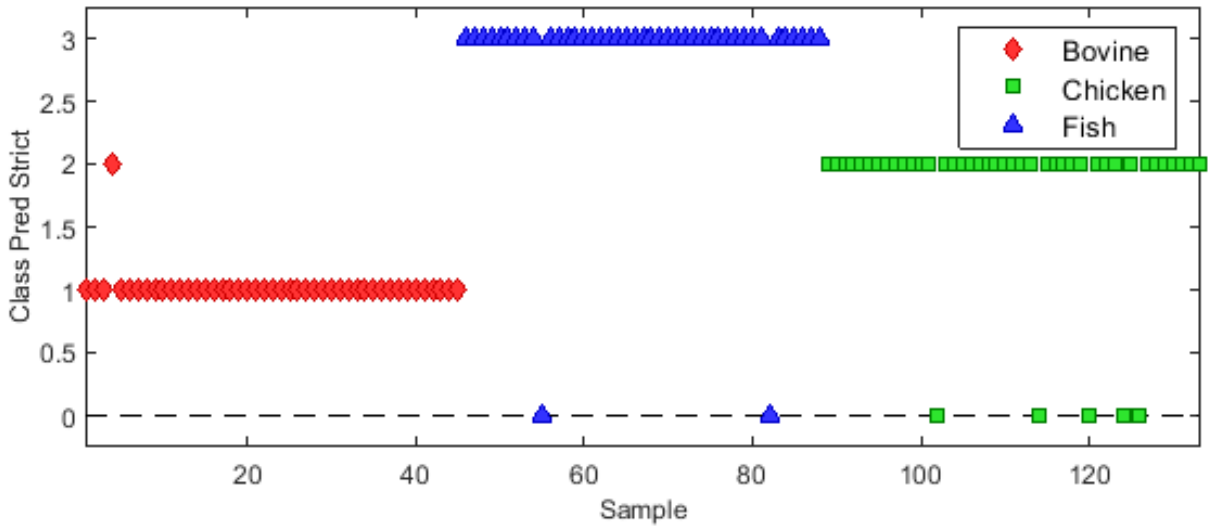
Il grafico di seguito in alto a sinistra rappresenta $T^2$ contro Q con tutte le classi, che essendo tutte diverse risulta una misura inaffidabile. Gli altri tre mostrano per ogni classe le y predette in cross-validazione con il threshold rappresentato dalle linee rosse tratteggiate e da questi si può affermare che si riescono a dividere le tre classi correttamente, salvo qualche punto di troppo accettato.



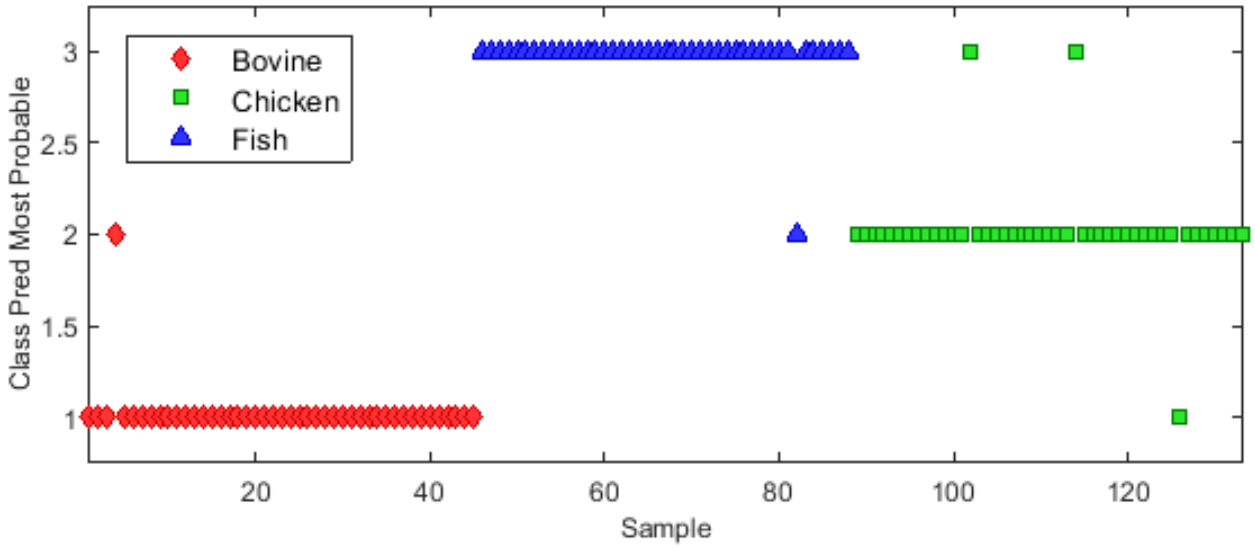

Vedendo il grafico delle class measured possiamo notare come le classi sono abbastanza bilanciate, il ché lì rende ideali per i metodi discriminanti.



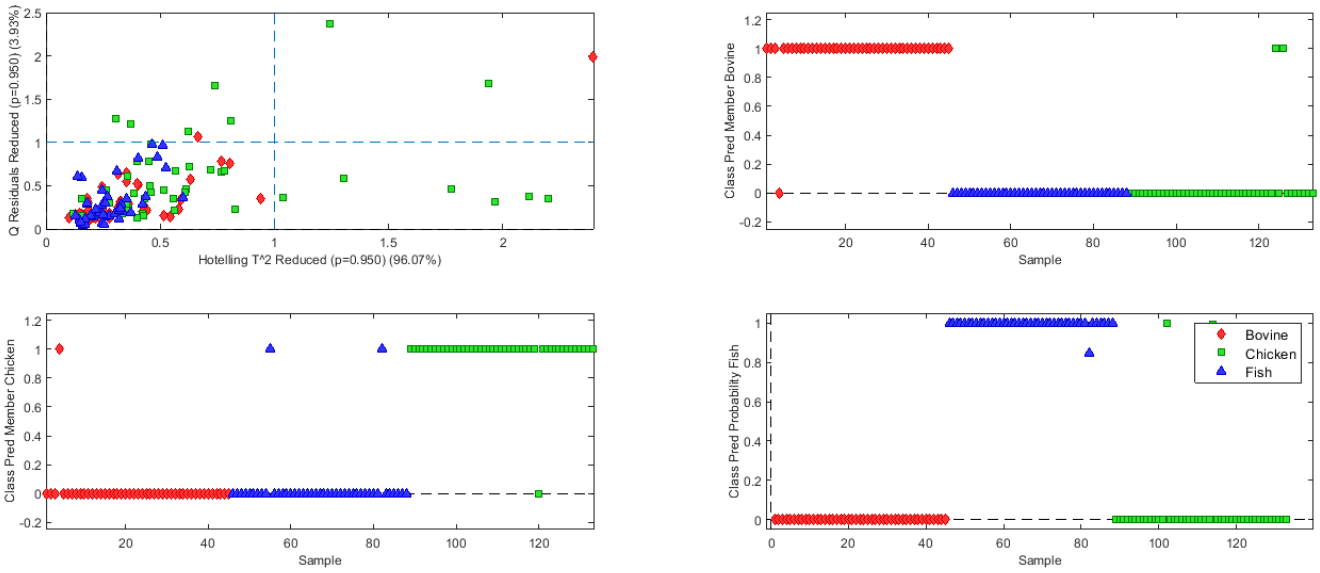

Applicando ai campioni il vincolo di stare in un'unica categoria (pred strict) si può riscontrare che la maggior parte dei gruppi è raggruppata correttamente, con l'eccezione di un campione bovino raggruppato in pollo e due campioni di pesce e cinque campioni di pollo non classificati.



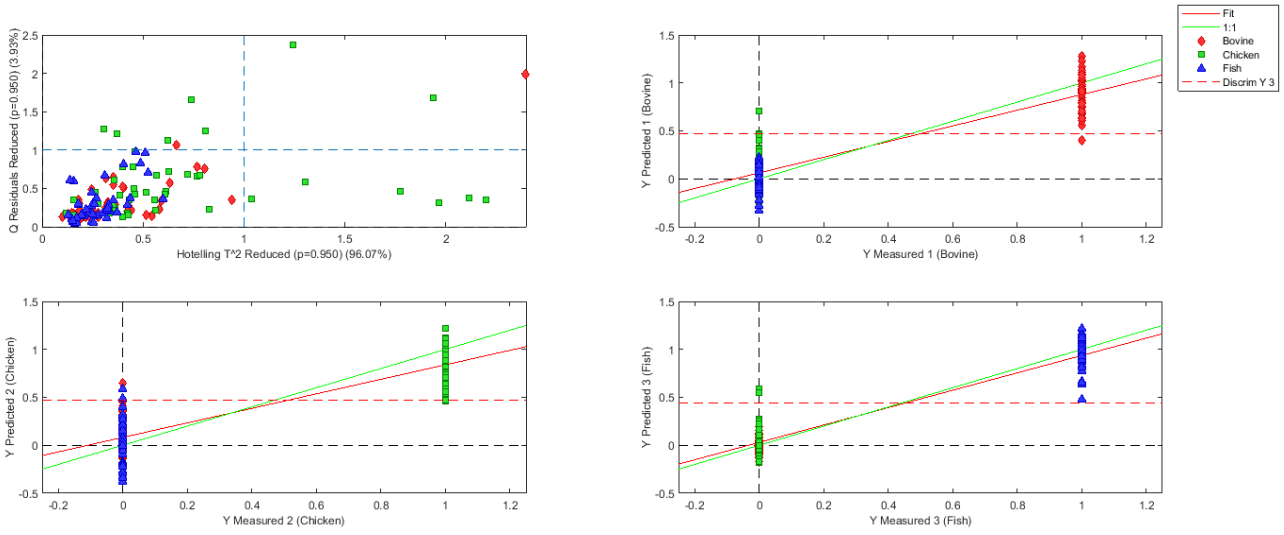

Usando l'altro criterio della probabilità (most probable), cioè un campione viene classificato nella categoria a cui è più probabile appartenga anche senza aver superato il threshold, tutti i campioni sono classificati con un errore nei bovini, due nei polli e due nei pesci.



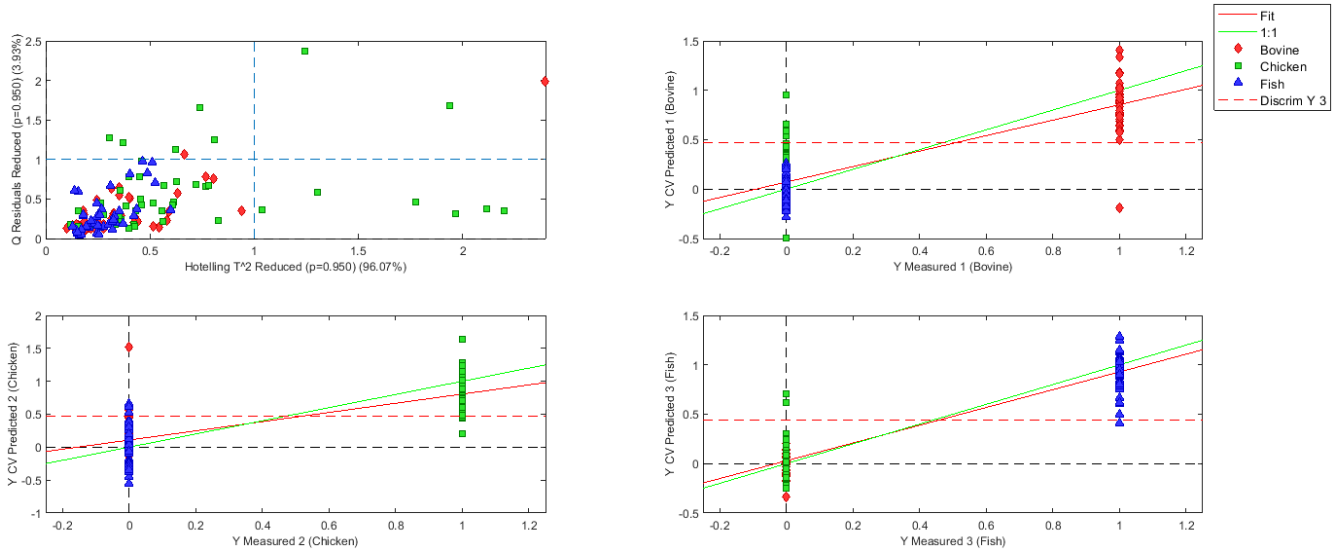

Anche guardando le probabilità per ciascuna classe, dove la distribuzione va da 0 a 1, la situazione non cambia. Per i bovini ci sono due errori, nei polli tre e nei pesci due e tutti i valori sono agli estremi del range, tranne un campione dei pesci.




Di seguito vi sono i grafici per classe delle y predette e y misurate dove tanto è migliore la classificazione tanto più il gruppo di campioni a destra sta a cavallo della linea verde (linea di regressione 1:1) e in generale sono stretti. Questo succede a tutti e tre, un po' meno i polli e il migliore è quello dei pesci.




Opinioni analoghe si possono riscontrare in cross-validazione dove però il gruppo dei bovini funziona peggio, soprattutto per la distribuzione.




Guardando la matrice di confusione, in particolare le tabelle con le classificazioni, ci si accorge come i valori sono molto simili tra tutte le tabelle.

Queste sono le generate con la regola del most probable; a sinistra vi è quella sul training set e a destra quella in cross-validazione.

Confusion Table:

| | Actual Class | | |
|---|---|---|---|
| | Bovine | Chicken | Fish |
| Predicted as Bovine | 44 | 1 | 0 |
| Predicted as Chicken | 1 | 42 | 1 |
| Predicted as Fish | 0 | 2 | 42 |
| Predicted as Unassigned | 0 | 0 | 0 |

Confusion Table (CV):

| | Actual Class | | |
|---|---|---|---|
| | Bovine | Chicken | Fish |
| Predicted as Bovine | 43 | 4 | 0 |
| Predicted as Chicken | 2 | 39 | 1 |
| Predicted as Fish | 0 | 2 | 42 |
| Predicted as Unassigned | 0 | 0 | 0 |

Queste invece sono le generate con la regola strict; a sinistra vi è quella sul training set e a destra quella in cross-validazione.

Confusion Table:

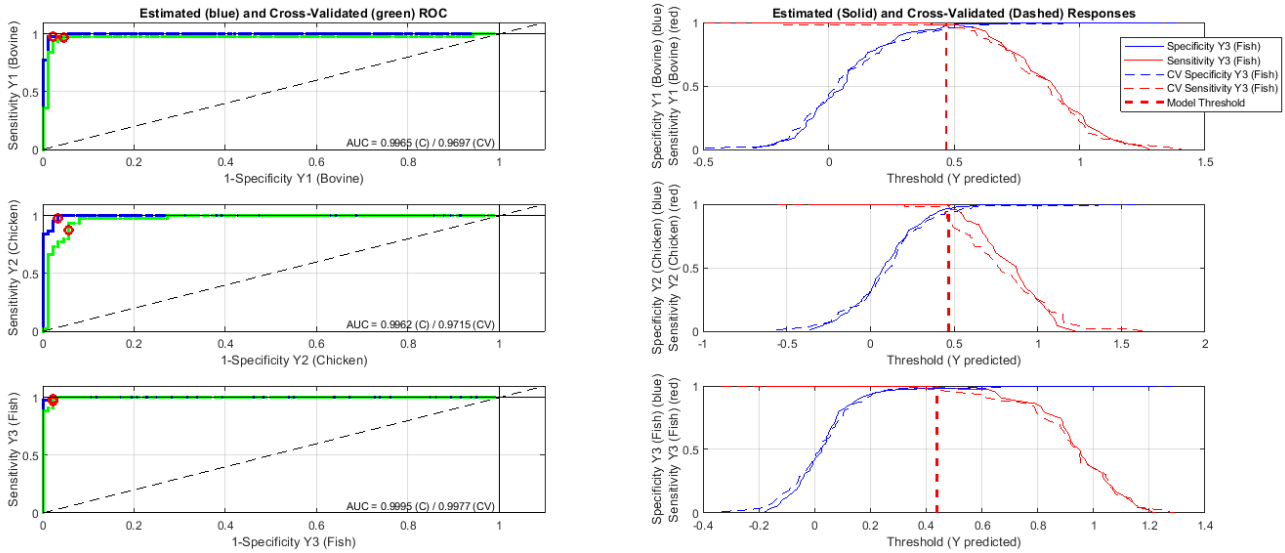
| | Actual Class | | |
|---|---|---|---|
| | Bovine | Chicken | Fish |
| Predicted as Bovine | 44 | 0 | 0 |
| Predicted as Chicken | 1 | 40 | 0 |
| Predicted as Fish | 0 | 0 | 41 |
| Predicted as Unassigned | 0 | 5 | 2 |

Confusion Table (CV):

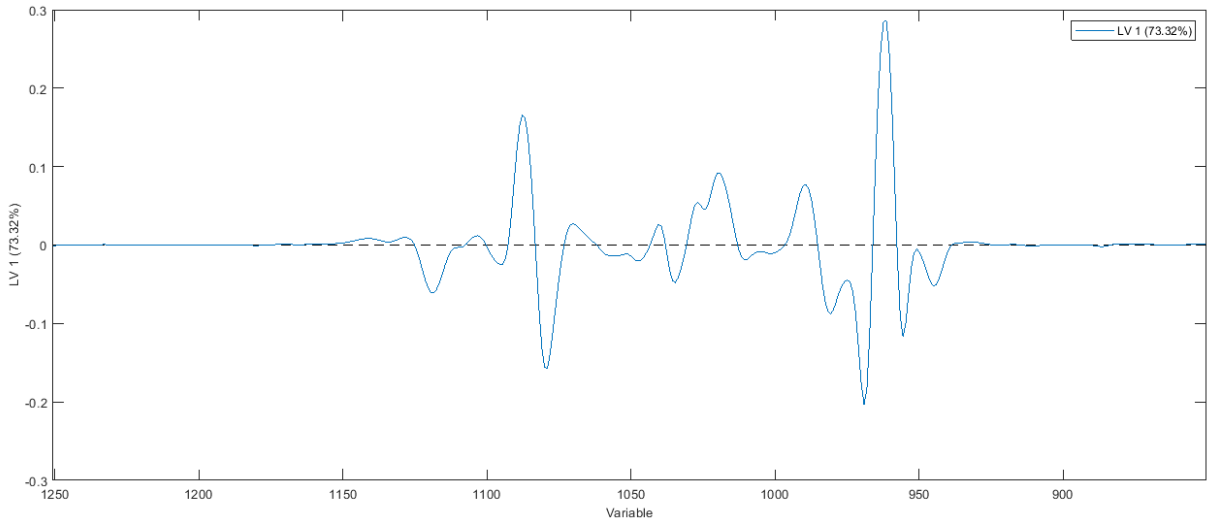
| | Actual Class | | |
|---|---|---|---|
| | Bovine | Chicken | Fish |
| Predicted as Bovine | 43 | 3 | 0 |
| Predicted as Chicken | 1 | 38 | 1 |
| Predicted as Fish | 0 | 1 | 40 |
| Predicted as Unassigned | 1 | 3 | 2 |

Dal grafico che segue si può vedere come viene deciso il threshold, cioè quando sensibilità e specificità sono vicine a 0.95. Le linee in fit e in cross-validazione hanno andamenti molto simili.

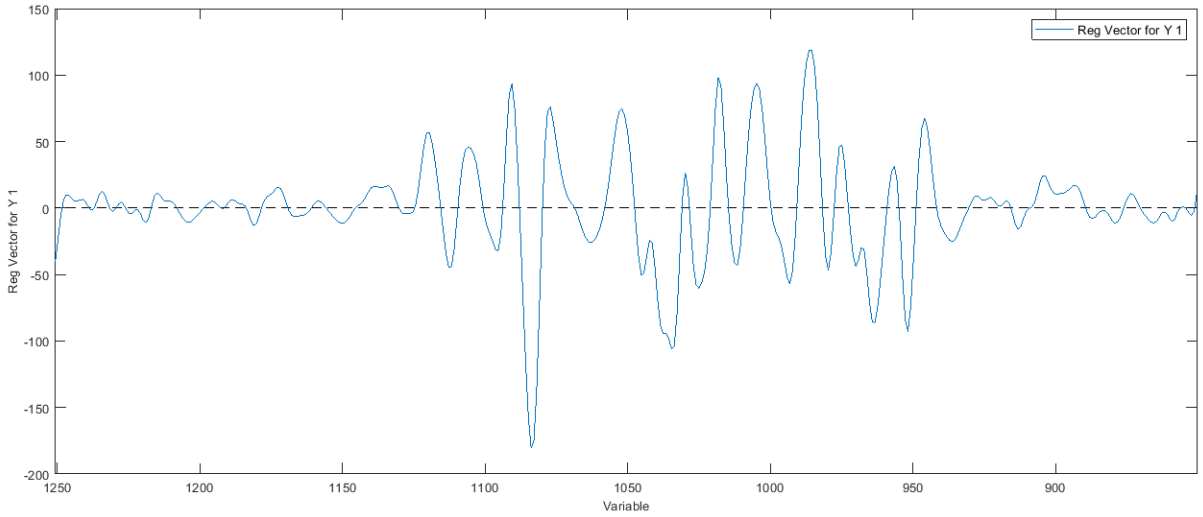
Nei grafici a sinistra, che rappresentano il rapporto tra sensibilità e 1-specificità, il punto migliore è 1 e per le tre classi i valori ci si avvicinano anche se la classe due meno rispetto alle altre due.



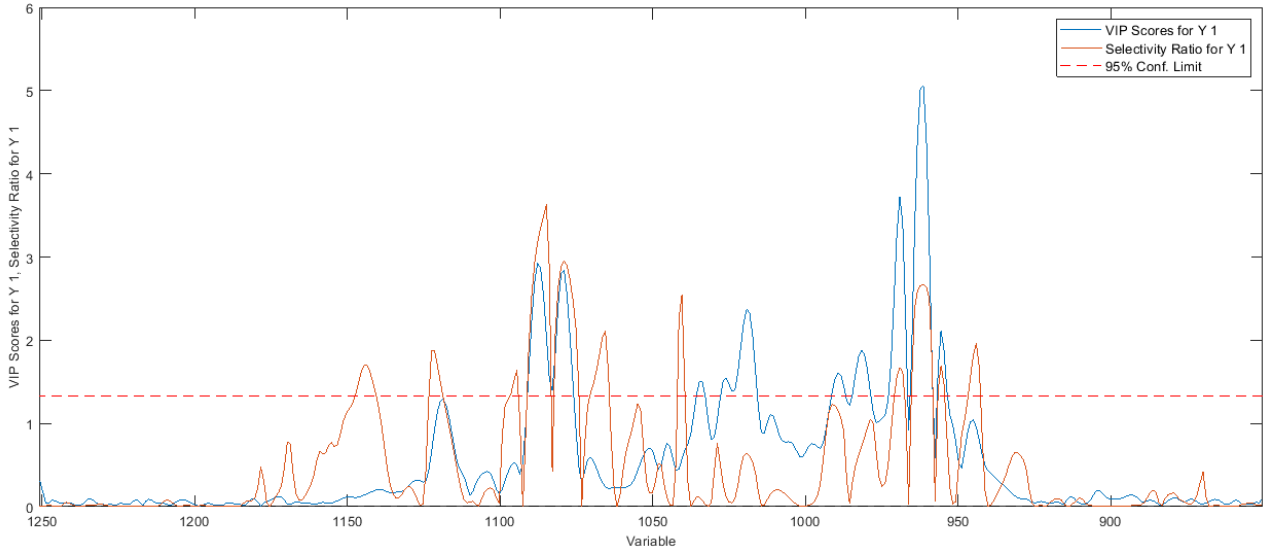

Un altro controllo che si può fare è verificare che non ci siano regioni rumorose poco significative, per fare ciò si può vedere il grafico dei loadings, che non mostra nulla che meriti una rimozione.




Anche i coefficienti di regressione sono poco rumorosi.




Nemmeno nel grafico dei VIP (Variable Influence in Projection) con la selectivity ratio si notano regioni rumorose che conviene rimuovere.




Infine, calcolando le predizioni, si può evidenziare il fatto che la sensibilità e la specificità tendono a calare in predizione rispetto a fit e a cross-validazione ma non in maniera drammatica.

```
Statistics for each y-block column:
Sensitivity (Cal):  0.978  0.978  1.000
Specificity (Cal):  0.977  0.966  0.978
Sensitivity (CV):  0.978  0.978  0.977
Specificity (CV):  0.955  0.932  0.978
Sensitivity (Pred):  0.929  0.846  1.000
Specificity (Pred):  0.913  0.917  0.963
```

I risultati fanno pensare che in questo caso il PLS/DA sia stato più efficace dello SIMCA, soprattutto per la sensibilità.

Concludendo, si elencano i punti cardine di questa relazione:

- Il preprocessing che è stato usato è 2$^{nd}$ derivative, MSC (median) e mean centering;
- Il tipo di cross-validazione è “Venetian blinds” con numero di splits 7;
- Con SIMCA le componenti principali sono 4 per la prima, 9 per la seconda, 1 per la terza;
- Con PLS/DA sono state usate 7 componenti principali;
- La sensibilità con PLS/DA è migliore rispetto a quella con SIMCA.

Gli strumenti di lavoro sono stati il PLS-toolbox per PLS/DA, mentre per i codici Matlab di SIMCA si ringraziano il prof. Federico Marini dell’Università di Roma La Sapienza e la prof. Marina Cocchi dell’Università di Modena e Reggio Emilia.