Electronic Journal of Applied Statistical Analysis
EJASA (2011), Electron. J. App. Stat. Anal., Vol. 4, Issue 2, 203 – 214
e-ISSN 2070-5948, DOI 10.1285/i20705948v4n2p203

MULTIBLOCK REDUNDANCY ANALYSIS FROM A USER'S PERSPECTIVE. APPLICATION IN VETERINARY EPIDEMIOLOGY

Stéphanie Bougeard[(1)]*, El Mostafa Qannari[(2)], Coralie Lupo[(1)], Claire Chauvin[(1)]

[(1)]*Department of Epidemiology, Anses, France*
[(2)]*Department of Chemometrics and Sensometrics, Oniris, France*

Received 17 September 2010; Accepted 06 June 2011
Available online 14 October 2011

Abstract*: For the purpose of exploring and modelling the relationships between a dataset Y and several datasets (X_1, …, X_K) measured on the same individuals, multiblock Partial Least Squares is a regression technique which is widely used, particularly in chemometrics. In the same vein, an extension of Redundancy Analysis to the multiblock setting is proposed. It is designed to handle the specificity of complex veterinary epidemiological data. These data usually consist of a large number of explanatory variables organized in meaningful blocks and a dataset to be predicted, e.g. the expression of a complex animal disease described by several variables. Some appropriate indices are also proposed, associated with different interpretation levels, i.e. variable, block and component. These indices are linked to the criterion to be maximized and therefore are directly related to the solution of the maximization problem under consideration.*

Keywords*: Multiblock modelling, multiblock Redundancy Analysis, PLS Path Modelling, epidemiology*

1. Introduction

Research in veterinary epidemiology is often concerned with assessing risk factors for complex animal health issues described by several variables. The associated data are usually organized in (K+1) blocks of variables, consisting of a large number of explanatory variables organized in K

* Corresponding author. E-mail: stephanie.bougeard@anses.fr

meaningful blocks (X_1, …, X_K) and several variables to explain Y. For this purpose, multiblock Partial Least Squares [17] is a multiblock modelling technique which is widely used. Redundancy Analysis [8; 10] is another popular method for linking two datasets X and Y, which is more oriented towards the explanation of the dataset Y. In this paper, an extension of Redundancy Analysis to the multiblock setting is proposed. This method is introduced by maximizing a criterion that reflects the objectives to be addressed and the solution of this maximization problem is directly derived from the eigenanalysis of a matrix. Moreover, multiblock modelling gives valuable tools both for the explanation and the investigation of the relationships among variables and among datasets. For the various interpretation levels, *i.e.* variable, block and component, we propose relevant indices, related to the criterion to be maximized and therefore directly derived from the solution of the maximization problem under consideration.

2. Method

2.1 Multiblock Redundancy Analysis

Consider the multiblock setting where we have (K+1) datasets: a dataset Y to be predicted from K datasets X_k (k=1, …, K). The Y table contains Q variables and each table X_k contains P_k variables. The merged dataset X is defined as $[X_1|\ldots|X_K]$ and contains $P=\sum_k P_k$ explanatory variables. All these quantitative variables are measured on the same N individuals and supposed to be column centered. The main idea is that each dataset is summed up by latent variables which are linear combinations of the variables derived from each dataset, as illustrated in Figure 1.

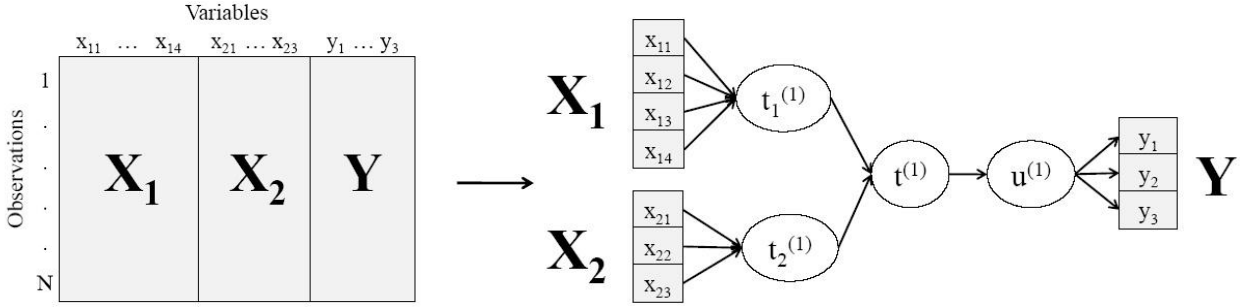

Figure 1. Conceptual scheme of the multiblock data structure, which highlights the relationships between the variable blocks (X_1, …, X_K, Y) and their associated latent variables (t_1, …, t_K, u) for the first dimension (h=1).

More precisely, the method derives a global component $t^{(1)}$ oriented towards the explanation of Y, that sums up partial components ($t_1^{(1)}$, …, $t_K^{(1)}$) respectively associated with the blocks (X_1, …, X_K). These components are computed following the maximization problem (1).

$$\operatorname{cov}^2(u^{(1)},t^{(1)}) \quad \text{with} \quad t^{(1)}=\sum_{k=1}^{K} a_k^{(1)} t_k^{(1)},\ \sum_{k=1}^{K} a_k^{(1)2}=1, \qquad (1)$$

$$u^{(1)}=Yv^{(1)}, t_k^{(1)}=X_k w_k^{(1)} \quad \text{and} \quad \left\|t_k^{(1)}\right\|=\left\|v^{(1)}\right\|=1$$

Other equivalent maximization problems may also be given which highlight other perspectives of the method [2]. We prove that the solution is given by $v^{(1)}$ the eigenvector of the matrix (2).

$$\sum_k Y' X_k (X_k X_k')^{-1} X_k' Y \text{ associated with the largest eigenvalue } \lambda^{(1)} \quad (2)$$

The partial components ($t_1^{(1)}$, ..., $t_K^{(1)}$) are given by the normalized projection of $u^{(1)}$ onto each subspace spanned by variables in blocks X_1, …, and X_k respectively, as detailed in Equation (3).

$$t_k^{(1)} == \frac{P_{X_k} u^{(1)}}{\left\| P_{X_k} u^{(1)} \right\|} \text{ for } k = (1,...,K) \quad (3)$$

where $P_{Xk}=X_k(X_k'X_k)^{-1}X_k'$ is the projector onto the subspace spanned by the variables from the X_k dataset. The coefficients $a_k^{(1)}$ are given by Equation (4).

$$a_k^{(1)} = \frac{\text{cov}(u^{(1)}, t_k^{(1)})}{\sqrt{\sum_l \text{cov}^2 (u^{(1)}, t_j^{(1)})}} = \frac{\left\| P_{X_k} u^{(1)} \right\|}{\sqrt{\sum_l \left\| P_{Xl} u^{(1)} \right\|^2}} \text{ for } k = (1,...,K) \quad (4)$$

This implies that the global component is given by Equation (5).

$$t^{(1)} = \sum_k a_k^{(1)} t_k^{(1)} = \frac{\sum_k P_{X_k} u^{(1)}}{\sqrt{\sum_l \left\| P_{Xl} u^{(1)} \right\|^2}} \quad (5)$$

In order to obtain the second order solution, we regress the variables in the datasets (X_1, …, X_K) onto the first global component $t^{(1)}$ and consider the residuals of these regressions. The same maximization is then performed by replacing the datasets by their associated residuals. Subsequent components can be found by reiterating this process.

2.2 Associated interpretation tools at the variable level

A main difference of multiblock Redundancy Analysis with most existing (K+1) methods lies in the fact that it focuses on the global components rather than on the partial components, as reflected by the adopted deflation procedure. We believe that the global components give more insight into the problem under study and give valuable tools both for the investigation of the relationships among datasets and the prediction [14]. The global components being orthogonal with each others, it is possible to depict overall graphical displays which highlight the relationships among all the variables, and to get an associated overall interpretation of similarities and differences between individuals. Moreover, the global component orthogonality

allows orthogonalised regression which takes into account all the explanatory variables and leads to the model in Equation (6).

$$Y = \hat{Y}^{(h)} + Y^{(h)} = X[w^{*(1)}c^{(1)\prime} + ... + w^{*(h)}c^{(h)\prime}] + Y^{(h)} \text{ for } h=(1, ..., H) \quad (6)$$

$\hat{Y}^{(h)}$ being the predicted values and $Y^{(h)}$ the residual matrix of the model based on h components. As a remark, the weights (7) and the loadings (8)

$$w^{*(h)} = \prod_{l=1}^{h-1}[I - w^{(l)}(t^{(l)\prime}t^{(l)})^{-1}t^{(l)\prime}X]w^{(h)} \quad (7)$$

$$c^{(h)} = (t^{(h)\prime}t^{(h)})^{-1}Y't^{(h)} \quad (8)$$

are defined as in PLS framework [16]. From a practical point of view, the final model may be obtained by selecting the appropriate number h of components to be introduced in the model by a validation technique such as cross-validation [11]. Moreover, bootstrapping simulations are performed to provide standard deviations and tolerance intervals, associated with the regression coefficient matrix [4; 9].

2.3 Associated interpretation tools at the block level

Multiblock methods are devoted to large datasets and lead to a high number of outcomes. Therefore, the user needs to get overall results. Firstly, he or she needs statistical indices to sort the explanatory variables by order of priority when the number of variables in Y is large. The Variable Importance for the Projection (VIP) proposed in the PLS framework [3; 18] is a relevant tool. It summarizes the overall contribution of each explanatory variable to the prediction of the Y block, summed over all components and weighted according to the amount of Y variance accounted for by each component. Associated standard deviations and tolerance intervals, computed using bootstrapped results, are also given. Additionally, it is of paramount relevance for the user to assess the contributions of the blocks (X_1,…, X_K) to the modelling task. The Block Importance in the Prediction, BIP is a relevant indice [12]. It is computed in the same vein as the VIP calculation, while using the weighted average values of the coefficients ($a_1^{(h)2}$, …, $a_K^{(h)2}$) which respectively reflect the link between Y and (X_1, …, X_K). An appealing feature is that these coefficients are included in the criterion to be maximized and are therefore directly derived from the maximization problem under consideration. Associated standard deviations and tolerance intervals, computed using bootstrapped samples, can also be given.

2.4 Associated interpretation tools at the component level

In effect, the user aims at achieving a synthesis of variables through components. As an example, the epidemiologist is not really interested in explaining each element of the disease (*i.e.* the Y block), but the disease, as a synthesis of these various elements. This can be achieved by linking their corresponding latent variables, $u^{(h)}$ on the one hand and ($t_1^{(h)}$, …, $t_K^{(h)}$) on the other hand. We propose a model that directly highlights the link between $u^{(h)}$ (*i.e.* the latent variable from the Y space) and $t_1^{(h)}$, …, $t_K^{(h)}$ (*i.e.* the latent variables from the explanatory blocks) for a given

dimension h. By regressing the latent variable $u^{(h)}$ on the basis of the partial latent variables ($t_1^{(h)}$, …, $t_K^{(h)}$), we get the following model (9):

$$u^{(h)} = \sum_k t_k^{(h)}[a_k^{(h)} c^{(h)'} v^{(h)}] + \varepsilon^{(h)} \text{ for } h = (1,...,H) \quad (9)$$

where $\varepsilon^{(h)}$ is the residual matrix of the model based on h components. The vector of coefficients $\beta_k^{(h)} = a_k^{(h)} c^{(h)'} v^{(h)}$ gives the direct link between the latent variables $u^{(h)}$ and ($t_1^{(h)}$, …, $t_K^{(h)}$) for a given dimension h. The percentage of variance of Y explained by $u^{(h)}$ is an additional interesting information. The main advantage of the proposed model in Equation (9) in comparison with the inner model available from PLS Path Modeling [15], is that the regression coefficients are not computed within an iterative algorithm, but are directly derived from the eigensolution. It follows that multiblock Redundancy Analysis may be a competitive alternative method to PLS Path Modelling in case of variables organized in (K+1) datasets. Associated standard deviations and tolerance intervals are provided from bootstrapped samples.

2.5 Alternative methods

It is worth mentioning that other methods are proposed in order to investigate the relationships among (K+1) datasets. Among all these techniques of analysis, we single out those methods which are based on a more or less similar maximization criterion than the one we have adopted herein. Some of these methods are proposed within the framework of generalized canonical analysis (*e.g.*, Generalized Canonical Analysis with a Reference Table [5]), others are proposed within the context of PLS regression (*e.g.*, multiblock PLS [13; 17]). Unlike Generalized Canonical Analysis with a Reference Table, we advocate using a deflation procedure which leads to orthogonal global latent variables. This enhances the outcome interpretation and improves the prediction. With regards to multiblock PLS, our method of analysis is more oriented towards the Y prediction but is likely to be less stable in case of multicollinearity within explanatory blocks. It is worth mentioning that multiblock Redundancy Analysis also bears high similarities to a method of analysis called Generalized Constrained Principal Component Analysis (GCPCA) [1]. Following the same notations as above, GCPCA is based on the maximization of the criterion $\sum_k \sum_{k'} \text{cov}^2(Y' t_k^{(1)}, Y' t_{k'}^{(1)})$ under the constraints $\sum_k \left\| t_k^{(1)} \right\|^2 = 1$ whereas, as stated above, mbRA is based on the maximization of the criterion $\sum_k \text{cov}^2(u^{(1)}, t_k^{(1)})$ with $u^{(1)} = Y v^{(1)}$, $\left\| t_k^{(1)} \right\| = \left\| v^{(1)} \right\| = 1$. We can show that the solutions of these two optimization problems are tightly linked. However, although we believe that further investigations are needed to compare more precisely mbRA and GCPCA, it is clear that the advantage of our strategy of analysis is to exhibit latent variables associated with the dependent dataset Y, namely $u^{(1)}$, $u^{(2)}$, etc. As shown in the case study, the latent variables are useful for quantifying the relationships between the explanatory blocks X_k and the block to be predicted Y.

3. Case study in veterinary epidemiology

3.1 Multiblock data and objectives

The dataset consists of a cohort of (N=351) broiler chicken flocks slaughtered in France. The aim is to assess the overall risk factors for losses in broiler chicken flocks (Y) described by four variables, *i.e.* the first-week mortality (Mort7), the on-farm mortality (Mort), the mortality during the transport to the slaughterhouse (Doa) and the condemnation rate at slaughterhouse (Condemn). For a complete description on the whole survey, the data collection and the exhaustive results refer to [6; 7]. Twenty three potential explanatory variables, organized in four meaningful blocks are selected: 5 variables pertain to the farm structure (X_1), 5 are selected from the flock's characteristics and the on-farm history of chicks at placement (X_2), 6 come from the flock's characteristics during the rearing period (X_3) and 7 from the catching, transport, lairage conditions, slaughterhouse and inspection features (X_4). Indicator (dummy) variables are considered for the categorical variables. As all the variables are expressed in different units, they are column centered and scaled to unit variance. We want to jointly assess the relative impact of the different production stages on the whole losses and the specific risk factors for each element of losses.

3.2 Interpretation at the variable level

The relationships between the 23 explanatory variables X and the 4 variables in Y are investigated using the graphical display in Figure 2, which depicts the loadings associated with the first three global components.

It turns out that the risk factors for the overall losses are associated with positive values on the global component $t^{(1)}$, whereas protective factors are associated with negative values. More precisely, the positive values on the component $t^{(2)}$ are associated with high values of mortality during the transport to the slaughterhouse (Doa, Y), whereas negative values are associated with high values of early mortality (Mort7, Y). The positive values on the component $t^{(3)}$ are associated with high values of condemnation rate at slaughterhouse (Condemn, Y), whereas negative values are associated with high values of on-farm mortality (Mort, Y). As an example, the mortality during the rearing period (Mort, Y) is in particular associated with the stress occurrence during rearing (Stress, X_3), the locomotor disorder observed (Locpb, X_3) and the genetic strain (Strain, X_3), among others.

In addition, a prediction model is set up by regressing the Y variables on the basis of the optimal number of components. This number is determined while using (m_{cv}=500) cross-validated samples. It is a compromise which optimizes both the fitting and prediction abilities and leads to a model with (h=4) components. The regression coefficients and the associated tolerance intervals of the 23 explanatory variables are performed for each dependent variable, while using (m_{bt}=500) bootstrapped samples. It appears that each element of losses is related with a specific set of risk and protective factors. As an example, Figure 3 shows the explanatory variables which are significantly associated with the first-week mortality (Mort7, Y).

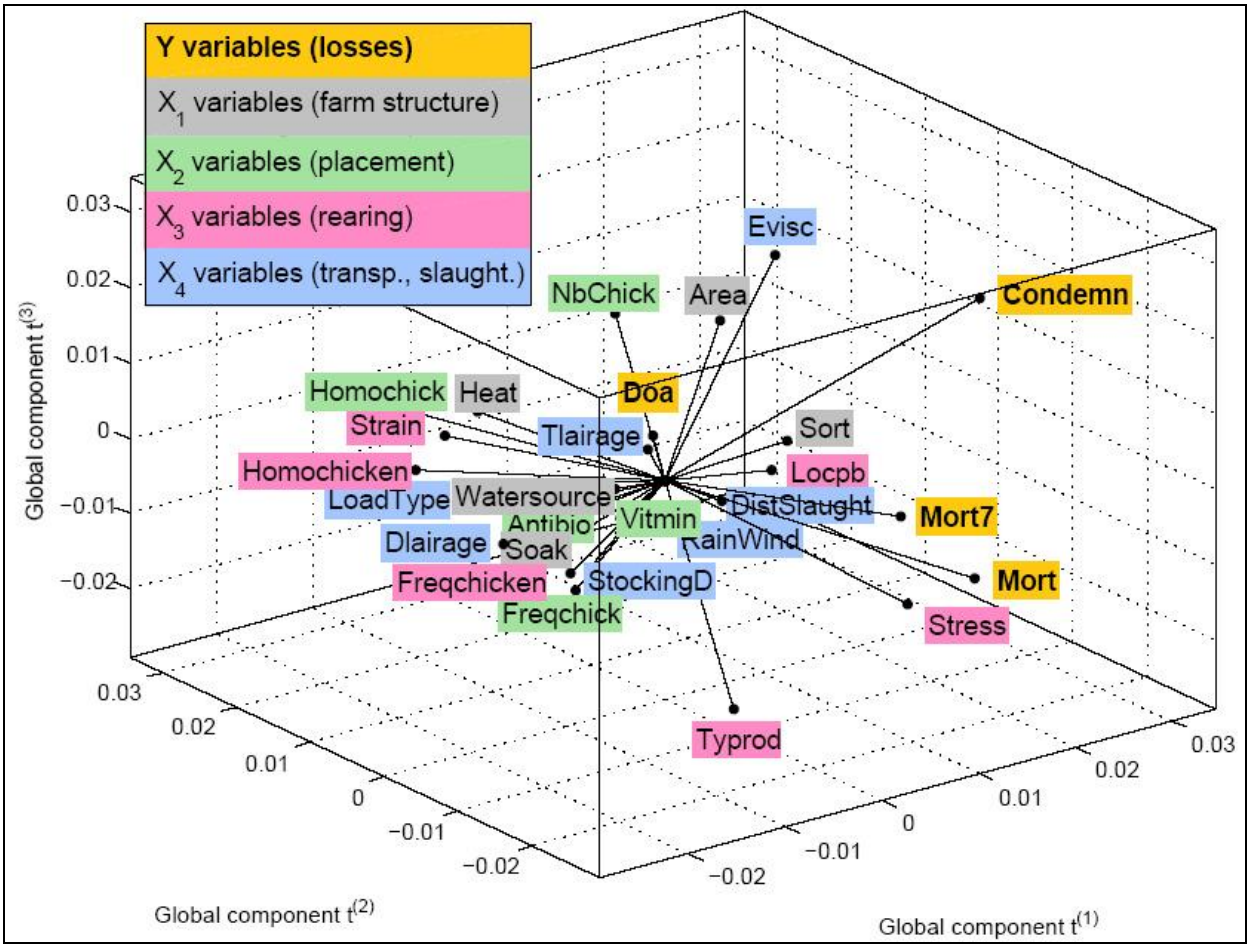

Figure 2. Interpretation at the variable level - Plot of the variable loadings associated with the first three global components ($t^{(1)}$, $t^{(2)}$, $t^{(3)}$).

The antibiotic distribution during the starting period ($\beta_{Antibio}$=-0.06 with a tolerance interval $[-0.12;-0.00]_{95\%}$, X_2), the heating system with gas heaters (β_{Heat}=-0.12 $[-0.19;-0.05]_{95\%}$, X_1) and above all the chick homogeneity at placement ($\beta_{Homochick}$=-0.27 $[-0.36;-0.17]_{95\%}$, X_2) are highlighted as significant protective factors. This means that particular care with respect to these variables should be taken in order to reduce the first-week mortality. The sorting practice of the animals (β_{Sort}=0.10 $[0.02;0.18]_{95\%}$, X_1) appears to be more a consequence of the early mortality than a risk indicator. As a remark, the variable Dlairage, relative to a further stage associated with the lairage at slaughterhouse, is not interpreted.

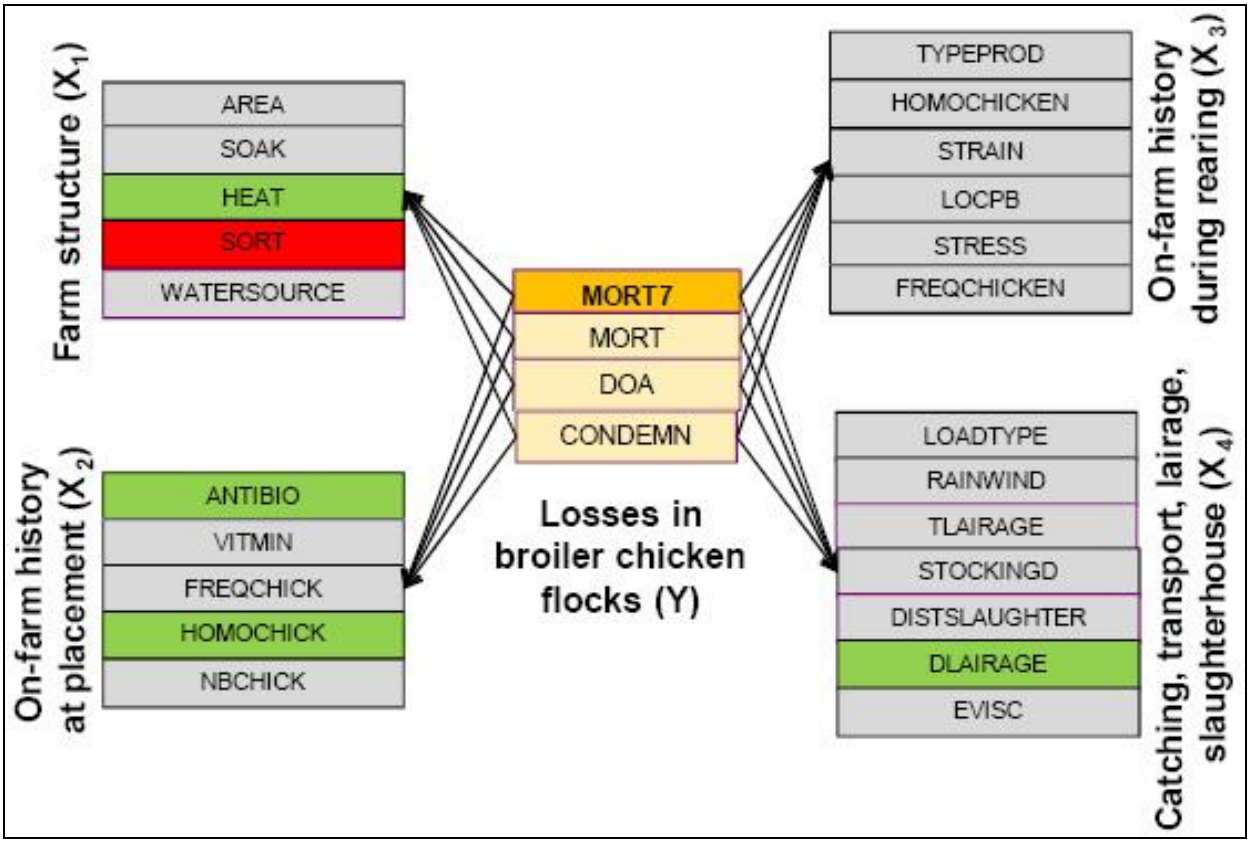

Figure 3. Interpretation at the variable level - Regression coefficients of the X variables for one variable in Y (Mort7). Significant risk factors are highlighted in red and protective factors in green.

3.3 Interpretation at the block level

As a consequence of the large number of variable to explain, the user needs statistical indices to sort the explanatory variables by order of priority. Figure 4 gives the Variable Importance for the Projection of the most important explanatory variables on the whole losses Y (VIP>0.8). Standard deviations and tolerance intervals associated with each VIP coefficient are computed using the results from the (m_{bt}=500) bootstrapped samples.

Four explanatory variables have a significant impact on the whole losses (Y): the stress occurrence during rearing (VIP_{Stress}=1.63 $[1.13;2.13]_{95\%}$, X_3), the homogeneity of chicks at placement ($VIP_{Homochick}$=1.45 $[0.97;1.93]_{95\%}$, X_2), the type of loading ($VIP_{LoadType}$=1.37 $[0.89;1.84]_{95\%}$, X_4) and the withdrawal of carcasses at the evisceration line (VIP_{Evisc}=1.19 $[0.80;1.58]_{95\%}$, X_4). The genetic strain has an important but not significant impact on losses (VIP_{Strain}=1.26 $[0.64;1.89]_{95\%}$, X_3). Each of these variables is especially related with one or two outcomes. In addition, Figure 5 gives the relative importance of each explanatory block X_k in the explanation of Y.

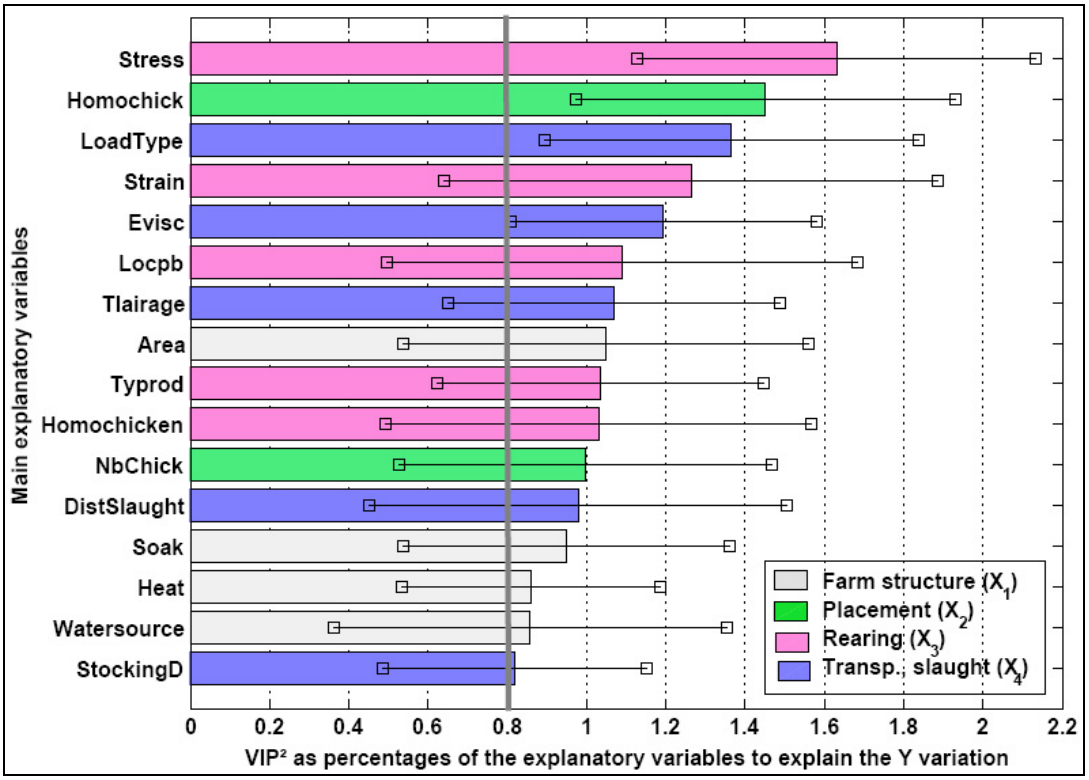

Figure 4. Interpretation at the block level - Variable Importance for the Projection (VIP) of the most important explanatory variables, associated with their 95% tolerance interval.

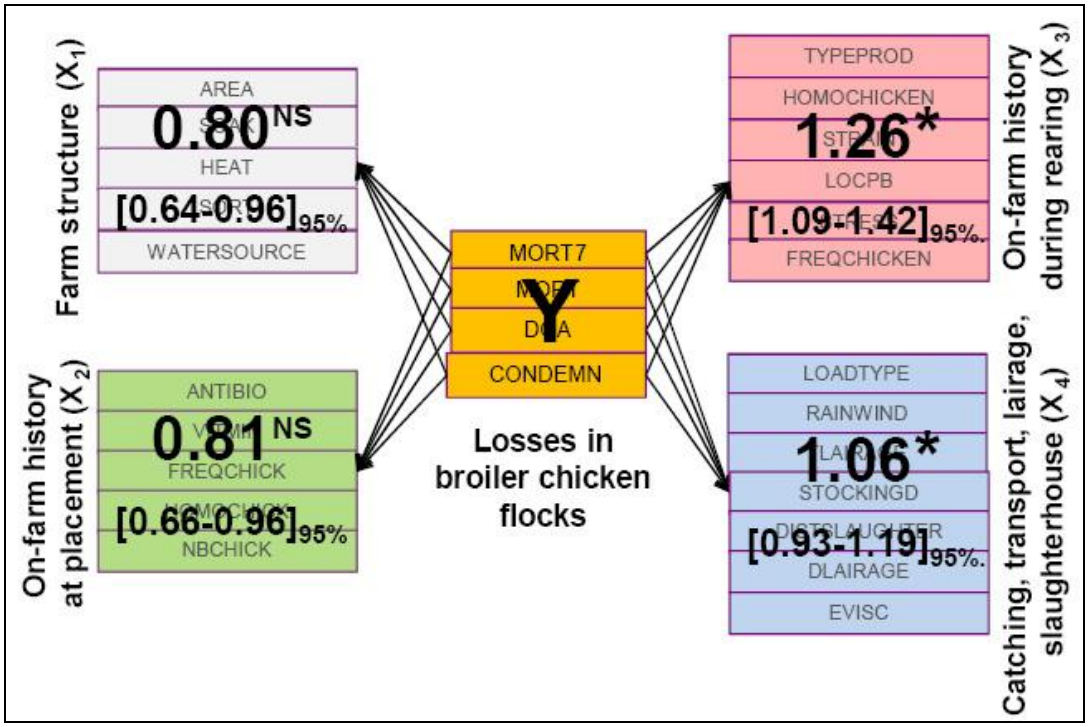

Figure 5. Interpretation at the block level - Block Importance in the Prediction (BIP) of the explanatory blocks X_k in the Y explanation, associated with their 95% tolerance interval.

Two blocks have a significant influence on the whole losses (Y): the flock's characteristics and the on-farm history of chicks during the rearing period (BIP_{X3}=1.26 $[1.09;1.42]_{95\%}$) and the catching, transport, lairage conditions, slaughterhouse and inspection features (BIP_{X4}=1.06 $[0.93;1.19]_{95\%}$). The two other blocks have an influence on Y (BIP>0.8) but it does not seem to be significant.

3.4 Interpretation at the component level

Finally, a model which directly reflects the links between $u^{(h)}$ and ($t_1^{(h)}$, …, $t_K^{(h)}$) for a given dimension h, gives an overview of the multiblock analysis. Results in Figure 6 give a synthesis of the first two models (h=1, 2), directly derived from the eigensolution, between the synthesis of the losses (Y), *i.e.* the component $u^{(h)}$, and the synthesis of the explanatory blocks (X_1, …, X_K), *i.e.* the components ($t_1^{(h)}$, …, $t_K^{(h)}$) respectively.

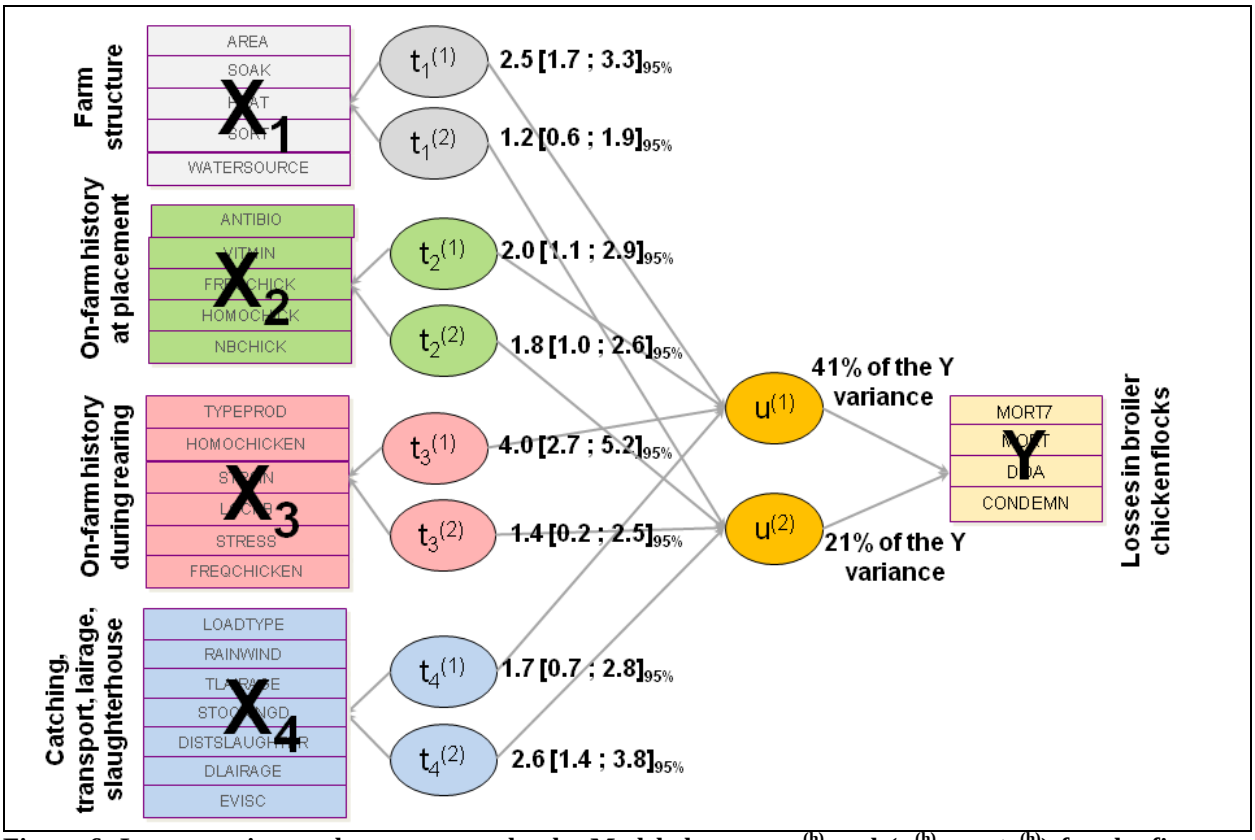

Figure 6. Interpretation at the component level - Models between $u^{(h)}$ and ($t_1^{(h)}$, …, $t_K^{(h)}$) for the first two dimensions h=(1, 2).

The first component $u^{(1)}$ explains 43% of the losses, is significantly related to all the partial components and is mainly linked to the flock's characteristics during the rearing period (β_{X3}=8.0 $[5.8;10.3]_{95\%}$). As an additional information, the component $u^{(1)}$ is mainly linked to the on-farm mortality (Mort, Y) and the condemnation rate at slaughterhouse (Condemn, Y). The second component $u^{(2)}$ explains 21% of the losses, is also significantly related to all the partial components and is mainly linked to the catching, transport, lairage conditions, slaughterhouse

and inspection features (β_{X4}=5.4 $[3.2;7.5]_{95\%}$). Moreover, we can notice that the component $u^{(2)}$ is mainly linked to the mortality rate during the transport to the slaughterhouse (Doa, Y).

4. Conclusion

Multiblock modelling handles the specificity of complex data and provides tools which can be helpful for the user to unveil hidden patterns and relationships among variables. For the purpose of exploring and modelling the relationships between one block of variables with several blocks of explanatory variables, we propose an extension of Redundancy Analysis, called multiblock Redundancy Analysis, as an alternative to multiblock PLS. The rationale behind this method is easy to understand because it is based on a global criterion to be maximized and, moreover, the solutions are directly derived from an eigenanalysis of a matrix. An appealing feature of the method lies in the introduction of global components on the one hand and partial components on the other hand, which highlight the relationships among the various datasets. Multiblock Redundancy Analysis is a relevant and useful tool to handle the specificity of complex data, as it enhances their interpretation and unveils new information for the user. As a general principle, it allows many possibilities of graphical displays and combines tools from factor analysis with tools pertaining to regression methods. Moreover, we propose some appropriate indices, related to the criterion to be maximized, which are therefore directly derived from the maximization problem under consideration. Multiblock Redundancy Analysis and associated interpretation tools are easy to grasp and implement. They are implemented using code programs developed in Matlab and are also made available in R. These code programs are available upon request from the first author.

References

[1]. Amenta, P., D'Ambra, L. (2001). Generalized Constrained Principal Component Analysis. *Advances in Classification and Data Analysis*. S. Borra, R. Rocci, M. Vichi & M. Schader (Eds.), Springer.

[2]. Bougeard, S., Hanafi, M., Qannari, E.M. (2007). ACPVI multibloc. Application à des données d'épidémiologie animale. *Journal de la Société Française de Statistique*, 148, 77–94.

[3]. Chong, I. G. and Jun, C. H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78, 103–112.

[4]. Efron, B., Tibshirani, R. (1993). *An introduction to the bootstrap*, New York: Chapman & Hall.

[5]. Kissita, G. (2003). *Les analyses canoniques généralisées avec tableau de référence généralisé : éléments théoriques et appliqués*. PhD Thesis, University of Paris Dauphine.

[6]. Lupo, C., Chauvin, C., Balaine, L., Petetin, I., Péraste, J., Colin, P. and Le Bouquin, S. (2008). Postmortem condemnations of processed broiler chickens in western France. *Veterinary Record*, 162, 709–713.

[7]. Lupo, C., Le Bouquin, S., Balaine, L., Michel, V., Peraste, J., Petetin, I., Colin, P. and Chauvin, C. (2009). Feasibility of screening broiler chicken flocks for risk markers as an aid for meat inspection. *Epidemiology and Infection,* 137, 1086–1098.
[8]. Rao, C.R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya, A.*, 26, 329–358.
[9]. Rebafka, T., Clémençon, S. and Feinberg, M. (2007). Bootstrap-based tolerance intervals for application to method validation. *Chemometrics and Intelligent Laboratory Systems,* 89, 69-81.
[10]. Sabatier R. (1987). Analyse factorielle de données structurées et métriques. *Statistique et Analyse des Données*, 12, 75–96.
[11]. Stone M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36, 111–147.
[12]. Vivien, M., Verron T. Sabatier, R. (2005). Comparing and predicting sensory profiles by NIRS: Use of the GOMCIA and GOMCIA-PLS multi-block methods. *Journal of Chemometrics,* 19, 162–170.
[13]. Wangen, L.E. Kowalski, B.R. (1988). A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of Chemometrics*, 3, 3–20.
[14]. Westerhuis, J.A., Smilde, A.K. (2001). Deflation in multiblock PLS *Journal of Chemometrics*, 15, 485–493.
[15]. Wold, H. (1982), Soft Modelling: The basic design and some extensions, in *Systems Under Indirect Observations*, part 2, eds. K.G. Jöreskog and H. Wold, Amsterdam: North-Holland, 1–54.
[16]. Wold, S., Martens, H., Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method. In *Conference on Matrix Pencils*, March 1982, 286–293.
[17]. Wold, S. (1984). Three PLS algorithms according to SW. In *MULDAST 1984: MULtivariate Analysis in Science and Technology*, Umea, Sweden, 4 - 8 June 1984, 26–30.
[18]. Wold, S. (1994). PLS for multivariate linear modeling. In *QSAR: Chemometric methods in molecular design. Methods and principles in medicinal chemistry*, eds H. van der Waterbeemd, Weinheim, Germany: Verlag-Chemie.