Computers and Concrete, *Vol. 13, No. 2 (2014) 173-185*
DOI: http://dx.doi.org/10.12989/cac.2014.13.2.173

Use of partial least squares analysis in concrete technology

Bulent Tutmez*

School of Engineering, Inonu University, 44280 Malatya, Turkey

(*Received March 6, 2013, Revised September 3, 2013, Accepted September 16, 2013*)

Abstract. Multivariate analysis is a statistical technique that investigates relationship between multiple predictor variables and response variable and it is a very commonly used statistical approach in cement and concrete industry. During model building stage, however, many predictor variables are included in the model and possible collinearity problems between these predictors are generally ignored. In this study, use of partial least squares (PLS) analysis for evaluating the relationships among the cement and concrete properties is investigated. This regression method is known to decrease the model complexity by reducing the number of predictor variables as well as to result in accurate and reliable predictions. The experimental studies showed that the method can be used in the multivariate problems of cement and concrete industry effectively.

Keywords: concrete; PLS; regression; slump flow; blaine

1. Introduction

As a strongly demanded engineering material, concrete comprised of coarse granular material embedded in a hard matrix of material (cement or binder) that fills the space among the aggregate particles and glues them together. The main characteristics of major constituents of concrete mixtures such as aggregates, cementitious materials, water, and admixtures should be comprehended first to better learn the properties and performance of concrete (Li 2011). In addition to these characteristics, workability of a concrete can be recorded as one of the important parameters. A concrete having a high workability with good consistency can manifest the properties of high dimensional stability and high durability (Mehta and Monteiro 1993).

Investigation of the relationships among the cement and concrete characteristics such as chemical and mechanical properties is still a novel topic. In addition to modeling workability behavior of High Performance Concrete (HPC) from ingredients, predicting mechanical and chemical properties of cement and concrete is very important due to costs. Furthermore, using a smaller number of predictors, variables can also reduce the computation time considerably.

Recently, there have been many conventional works recorded on modeling the cement and concrete characteristics in literature ((Eswari *et al*. 2011, Ahangar-Asr *et al*. 2011, Miled *et al*. 2012). In addition to the conventional approaches, computational intelligence-based modeling algorithms have been extremely popular. By using Artificial Neural Networks (ANN) (Yeh IC

*Corresponding author, Professor, E-mail: bulent.tutmez@inonu.edu.tr

http://www.techno-press.org/?journal=cac&subpage=8 ISSN: 1598-8198 (Print), 1598-818X (Online)

1999, Bai *et al*. 2003, Rasa *et al*. 2009) and neural-fuzzy synergism (Zarandi *et al*. 2008, Tutmez 2009, Mohammadhassani *et al*. 2013) in modeling and calibration problems, many strong model structures, which were aimed to provide high performance capacity, were published. On the other hand, as discussed in different works (Taha 2012) even though neural network-based models produce some successful predictions, their disadvantages include its "black box", greater computational burden, proneness to over fitting, and the empirical nature of model development. Furthermore, in case of the training set consists of too little or too much data, they may be very sensitive and not perform well.

Multivariate analysis is a well-known statistical technique that investigates the nature and significance of the relationship between multiple predictor variables and response variable (Neter *et al*. 1996). It is a very commonly used statistical approach while dealing with complex problems in engineering since it provides power with high performance compared to the simple regression analysis (Musa 2013). However, in multiple regression analysis there exist some crucial statistical phenomenons that should be taken into account such as decreasing the model complexity by reducing number of predictor variables in the model and handling the collinearity problems between these predictors.

As a general shortcoming, the multicollinearity problem, which is the presence of highly inter-correlated predictor variables in multivariate regression models, is ignored in engineering modeling works. However, it has a crucial importance to obtain reliable and meaningful model outputs (Yeniay and Goktas 2002, Chen 2012, Chen 2012). It usually leads to unreliable estimates of the regression coefficients, which then have large variances and covariances. (Draper and Smith 1998). Since a notable relationship between concrete variables is expected, diagnosing multicollinearity between the independent variables and developing strong model structures by eliminating the collinearity effect is required. This study examines the use of Partial Least Squares (PLS) method (Wold *et al*. 2001), which solves collinearity problem in the data and reduces the number of predictor variables, in cement and concrete industry. Besides, in many engineering solutions, computation time is an important parameter that cannot be neglected, especially in cross-validation and variable selection steps (Martins *et al*. 2008). The computation time mainly depends on the matrix dimension (number of predictors) used in the solution. Therefore, a fast algorithm is needed for such cases, since the computation time can be reduced during model development works.

The rest of the paper is structured as follows. Section 2 describes the methodology of PLS calibration and regression (PLSR). Section 3 presents the case studies. Results, performance comparisons and discussion are given in Section 4. Section 5 concludes the paper.

2. Methodology

2.1 Multiple regression and multicollinearity

A multivariate regression model deals with several *x*-variables $x_1, x_2, \dots, x_p$, on the same individuals. This results in the values x_{ij} where $i = 1, \dots, n$ is the index for the objects and $j = 1, \dots, m$ is the index for the variables. In this problem, the target variable can be observed on the same objects with the corresponding values $y_1, y_2, \dots, y_n$ and a linear model can be established to relate all *x*-variables with *y*.

In a general multi-linear regression structure, y is related to a linear combination of the *x*-

variables, plus an additive error term *e*. Based on input matrix *X* and output vectors *y* and *e*, the multi-linear model is formulated as follows

$$y = Xb + e \tag{1}$$

where b denotes the regression coefficients $b = (b_0, b_1, \dots, b_m)^T$. The least squares solution for the coefficients can be given by

$$b = (X^T X)^{-1} X^T y \tag{2}$$

If the term $X^T X$ is singular, an inversion cannot be made and the normal equations do not have a unique solution. It seems from the fact that there is at least one linear combination of the columns of the *X* matrix that is zero. In other words, at least one column of *X* is linearly dependent on the other columns. Thus, collinearity among the columns of *X* arises (Draper and Smith 1998).

Diagnosing and eliminating collinearity has critical importance in modelling problems. It usually leads to unreliable estimates of the regression coefficients, which then have extremely large variances and covariances (Neter *et al*. 1996). One of the tools for reducing the number of predictor variables and removing multicollinearity is using the variable selection methods such as stepwise selection and best-subset regression. Although the variable selection approach may lead to a regression model with a good interpretability, the price to pay requires a high computational effort, especially for a large number of independent variables (Dobrska *et al*. 2012). As an alternative approach, Principal Component Regression (PCR) solves the problem of data collinearity and reduces the number of predictor variables, but the predictor variables are no longer the original observed independent predictor variables but linear combinations thereof (Martens and Naes 1998).

2.2 Partial least squares method

In its general form PLS produces orthogonal score vectors (also called latent vectors or components) by maximizing the covariance between different sets of variables. It extracts the score vectors which serve as a new predictor representation and regresses the target variables on these new predictors (Rosipal and Kramer 2006).

In handling numerous and collinear *x*-variables, and response profile (*y*), PLS regression gives an opportunity to handle more complex problems, and analyze available data in a more realistic way (Wold *et al*. 2001). Essentially, the model structures of PLS and PCR are very similar: the data are first transformed into a set of a few 'intermediate linear latent variables' and the dependent variable *y* is regressed on these new variables. The criterion for the intermediate latent variables that is mostly applied in PLS is maximum covariance between scores and *y* (Liebmann *et al*. 2010).

The PLS algorithm aims to find a linear relation considering the linear structure in Eq. (1) with $n \times m$ matrix X of independents and $n \times p$ matrix Y of dependent variables. One can approximate

$$X \approx TP^T$$

$$Y \approx UQ^T \tag{3}$$

where *T* and *U* represent the respective score matrices that comprise of linear combinations of the

variables and P and Q represent the loadings matrices of X and Y , respectively (Varmuza and Filzmoser 2009). There is an inner relation between T and U as follows:

$$U = TD + H \tag{4}$$

where H is matrix of the residuals and D covers the regression coefficients on diagonals, $D = diag(d_1, d_2, \ldots d_a)$.

The estimator for D presents an estimate for $\widehat{U} = T\widehat{D}$ and in turn, that for $\hat{Y} = T\widehat{D}Q^T$. Based on the general linear relationship $\hat{Y} = X\hat{B}$, the estimator of PLS can be provided as follows ((Liebmann *et al*. 2010):

$$\hat{B}_{PLS} = P\widehat{D}Q^T \tag{5}$$

Due to technical reasons, orthogonal weight vectors w and c, and loadings $t = Xw$ and $u = Yc$ are used. Finally, the maximization of the objective function of PLS can be expressed as follows

$$cov(Xw, Yc) \rightarrow \max \ \ \|t\| = \|Xw\| = 1 \ \text{ and } \ \|u\| = \|Yc\| = 1 \tag{6}$$

where 'cov' represents the sample covariance. It should be noticed that the constraints have to be either length of score vectors equals 1 or length of weight vectors w and c equals 1 (Varmuza and Filzmoser 2009).

Among the novel PLS algorithms, the SIMPLS algorithm is mostly preferred (de Jong 1993). This algorithm directly maximizes the objective function of PLS under the constraint of orthogonality of the *t*-scores for different components (Mevik and Wehrens 2007). One advantage of the SIMPLS method is that it is not necessary to deflate X or Y, which may result in faster computation and less memory requirements (Martins *et al*. 2008).

In SIMPLS algorithm, the deflation is conducted for the cross-product matrix, $S = X^TY$ between the data, not for the centered data matrices, X and Y. The pseudocode for the SIMPLS algorithm is given in Appendix A (Varmuza and Filzmoser 2009).

As a result of the SIMPLS implementation, the resulting weights w and scores t are stored as columns in the weight matrix W and score matrix T, respectively. Thus, the final regression coefficients are provided by the following expression:

$$B = WT^TY \tag{7}$$

3. Case studies

To illustrate the PLS analysis on concrete properties, two case studies are considered. Performance of the SIMPLS algorithm in cement and concrete industries are assessed through these data sets. The major parts of the analyses were carried out using the *pls* (Mevik and Wehrens, 2007) and *chemometrics* (Varmuza and Filzmoser 2009) packages in *R* (R Development Core Team 2008).

3.1 Case study 1

Predicting workability behaviour of HPC from concrete ingredients can provide many practical advantages. In this context, to evaluate the prediction capacities of two models which are second order regression and Neural Network models, Yeh (2007) investigated the relationship between slump flow (cm) of the fresh concrete and seven ingredients: cement (kg/m^3), blast furnace slag (kg/m^3), water (kg/m^3), fly ash (kg/m^3), coarse aggregate (kg/m^3), fine aggregate (kg/m^3), and super-plasticizer (kg/m^3). In our first application, based on the data and model structures given in Yeh (2007), a PLS model is developed. The data set used in this study comprises of 78 measurements.

3.1.1 Determining multicollinearity

It is known that when the predictor variables in a multiple regression model are uncorrelated, the relationship between a predictor variable and the response variable is the same as their relationship in a simple regression model. Any deviations may be indicative of multicollinearity. A way of exploring the presence of the multicollinearity is computing the variance inflation factor (VIF) statistics for each predictor variable. This statistics is obtained when one of the predictor variables regressed on the remaining predictor variables. Simple regression, multiple regression analysis and collinearity statistics are depicted in Table 1.

Table 1 Regression results and collinearity statistics for case study 1

	Simple regression			Multiple regression			Collinearity statistics
	Coefficient	Std. error	*t*	Coefficient	Std. error	*t*	VIF
Intercept				459.836	3985.116	0.115	
Cement	0.029	0.023	1.295	-0.165	1.268	-0.130	4309.951
Slag	-0.044	0.030	-1.489	-0.301	1.794	-0.168	4977.699
Water	0.454	0.084	5.408	-0.016	3.976	-0.004	2295.287
Flyash	-0.005	0.023	-0.197	-0.180	1.397	-0.129	5320.860
Caggr	-0.047	0.022	-2.125	-0.204	1.567	-0.130	6773.331
Faggr	6.598e-04	3.179e-02	0.021	-0.184	1.567	-0.118	3447.220
SP	-1.749	0.594	-2.945	-0.473	3.388	-0.140	41.483

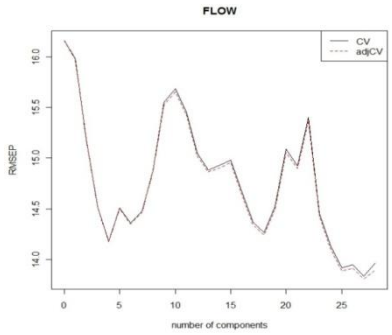


Fig. 1 Determining number of components

3.1.2 Model development

In the first stage, four sample testing sets were sampled from the original data set. The numbers of the observations in these sets were defined as 20, 20, 20, and 18, respectively, similar to the selection process of Yeh (2007). The remaining values of the data were employed for building the regression models. To be accordant with the model structures of the referred study, interactions of the independent variables were also taken into account in this application.

In the second stage, the optimum number of components was determined as four by a cross-validation (CV) method (Fig. 1). Because the correlations between independent variables are not small, a dimension reduction is possible without loss of variance (Fig. 2). In addition, Fig. 3 indicates the loading values via spectral peaks of the components.

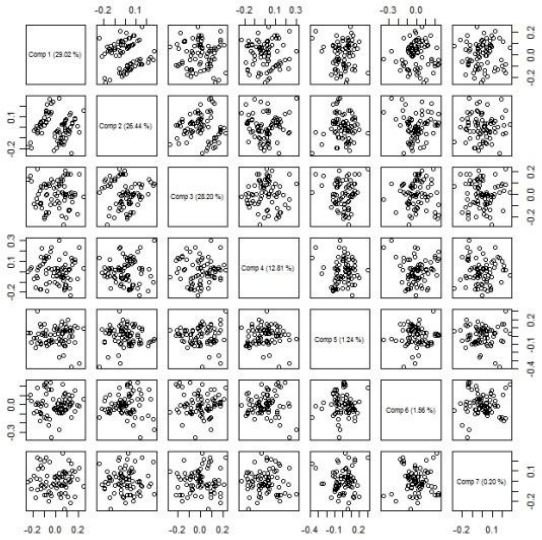


Fig. 2 Cross correlations between the variables

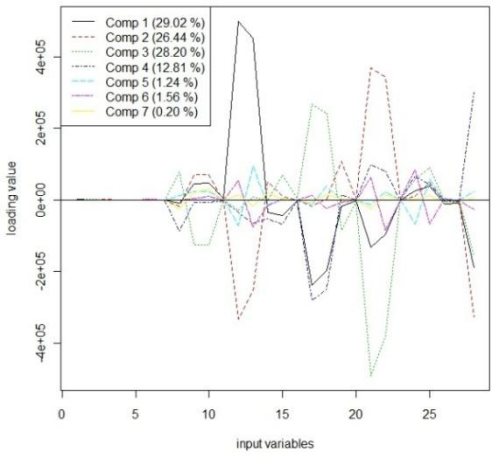


Fig. 3 Changes of loading values for components

3.2 Case study 2

As a concrete ingredient, cement is a crucial industrial material used in construction and building works. Few construction projects can take place without utilizing cement somewhere in the design. The fineness of cement particles is an important property and it can be determined by the value Blaine (kg/m^2). Although Blaine value addresses a surface parameter, some relationships between Blaine and chemical components such as CaO, SiO_2, Al_2O_3, Fe_2O_3, Na_2O, K_2O, SO_3, SCaO, C_3S are expected.

To investigate the relationships between Blaine and chemical properties (9 parameters), a real data set obtained from Adana Cement Factory (Sahin 2009, Tutmez and Dag 2012) was considered. The data set covers 40 laboratory measurements for each parameter.

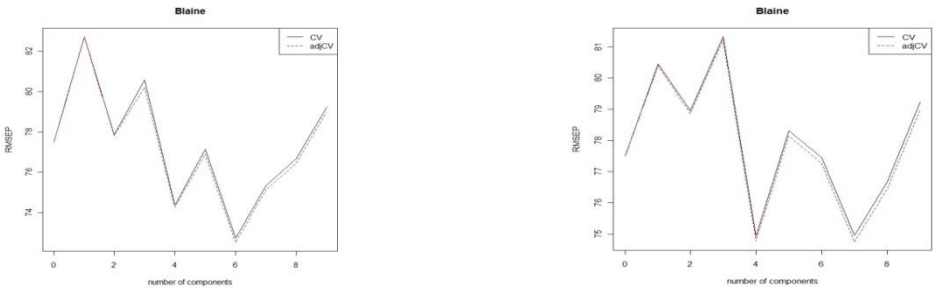
3.2.1 Determining multicollinearity

Similar to the Case Study 1, a simple regression, multiple regression analysis and collinearity statistics are computed and reported in Table 2. It especially shows that multiple regression analysis results in estimates and estimated variance, which are very large in magnitude. In fact, VIF statistics other than Na2O, K2O and SO3 are fairly large than 20, which indicate these predictors are strongly correlated with each other.

3.2.2 Model development

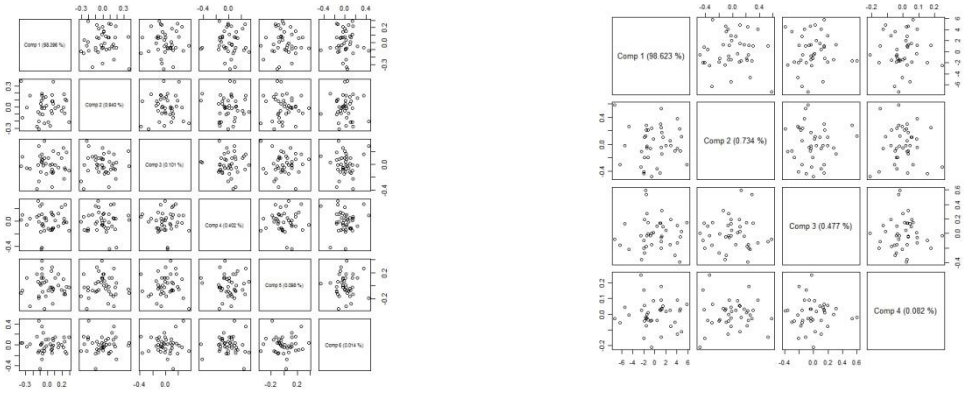
To assess the model capacities by different numbers of data, two minor applications have been carried out. For the first application, the data set was randomly divided into two subsets: the training set (50%: 20 records) and the testing set (50%: 20 records), respectively. Similarly for the second application, the training set (75%: 30 records) and the testing set (25%:10 records) were sampled, respectively. To make a comparative assessment, the PCR method was performed for each application as well.

In the first application, by using the CV approach, the numbers of components have been determined for both PLSR and PCR methods (Fig. 4). As seen in the Fig. 4, these numbers have been recorded as 6 and 4 for PLSR and PCR models, respectively. Fig. 5 summarizes the cross-correlations between the variables for optimal numbers of components. The loading values of the models can be followed in Fig 6. The similar procedures were applied to take the results for second application.



(a) (b)

Fig. 4 Determining number of components (a) PLSR model (b) PCR model



(a) (b)

Fig. 5 Correlations between the variables (a) PLSR model (b) PCR model

Table 2 Regression results and collinearity statistics for case study 2

	Simple regression			Multiple regression			Collinearity statistics
	Coefficient	Std. error	t	Coefficient	Std. error	t	VIF
Intercept				23571.6	12767.0	1.846	
CaO	-104.02	60.22	-1.727	-1940.3	19098.3	-0.102	121857.6e+05
SiO2	-27.56	30.80	-0.895	3126.2	35752.3	0.087	1724232e+06
Al2O3	-89.49	164.88	-0.543	2474.6	31644.8	0.078	47761.63e+04
Fe2O3	-57.05	146.07	-0.391	465.1	6727.2	0.069	2760.563e+03
Na2O	220.2	810.2	0.272	-721.4	1126.9	-0.640	2.523
K2O	-163.1	396.2	-0.412	-855.0	454.2	-1.883	1.709
SO3	280.2	119.1	2.352	178.6	161.0	1.109	2.084
SCaO	71.05	63.23	1.124	1838.3	19152.3	0.096	116022.0e+05
C3S	1.071	4.088	0.262	412.1	4702.6	0.088	1726104e+06

4. Results and discussion

The results of the models were appraised by some performance measures from a comparative perspective. Like Yeh (2007), in the first case study the well-known indicators, coefficient of determination (r^2) and roots mean squared error (RMSE) were utilized. In the second case study, three performance indicators: mean absolute error (MAE), coefficient of correlation (CoC) and standard deviation (Std) were employed.

The slump flow models developed in the first case study provided some reliable results as in Fig. 7. The performances of the models with the results given in Yeh (2007) are summarized in Table 3.

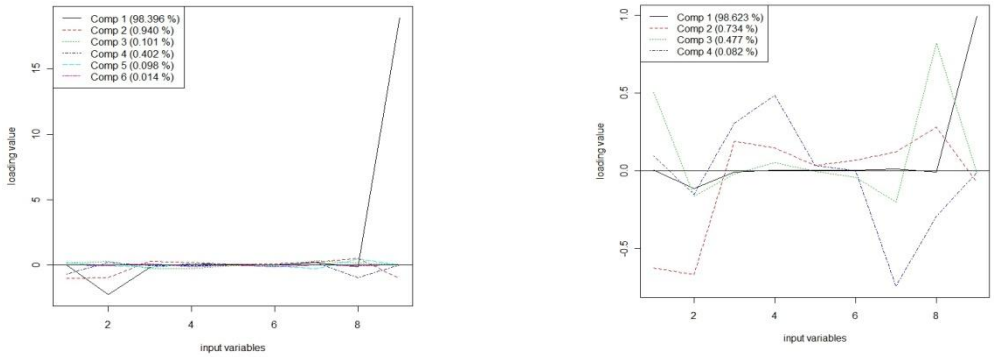
The PLS algorithm also produced the reliable results for Blaine prediction from the chemical ingredients. Table 4 and Fig. 8 indicated better performances for PLS algorithm compared to PCR algorithm via different performance indicators. In addition, it should be stressed that the PLS model reproduces the variability of the measured data in the value of predicted values. This result

Table 3 Performances of different models

Testing data set		Second-order regression (Yeh 2007)	Neural network (Yeh 2007)	PLS analysis
Data set 1	r^2	0.131	0.696	0.709
	RMSE	15.15	9.93	10.11
Data set 2	r^2	0.466	0.812	0.770
	RMSE	10.11	9.10	11.57
Data set 3	r^2	0.304	0.775	0.702
	RMSE	22.29	7.51	18.38
Data set 4	r^2	0.446	0.803	0.801
	RMSE	10.81	8.14	8.97
Integral testing set	r^2	0.323	0.724	0.707
	RMSE	15.57	8.51	10.25

Table 4 Performance measure for PCR and PLSR

Performance Measure	Measured data		PCR		PLSR	
	20 observations	10 observations	20 observations	10 observations	20 observations	10 observations
CoC (*r*)	-	-	0.678	0.806	0.794	0.843
MAE	-	-	83.14	64.64	36.99	60.44
Std	70.740	95.922	46.068	26.725	54.611	34.669



(a) (b)

Fig. 6 Changes of loading values for number of components

showed that the PLS algorithms can remove the smoothing problem substantially.

Although both neural network and the proposed model produced similar r^2 and RMSE values for the slump flow models, it should be kept in mind that neural network models work under the assumption of no collinearity. Table 1 indicates the presence of the multicollinearity in the data and it is justified through VIF statistics. For that reason, although PLS gives relatively lower performance compared to neural network, an inference made upon this model is more reliable compared to the Neural Network since it takes multicollinearity into account. In addition, as presented in introduction section, neural network-based models are "black box" models and the

knowledge of their internal working is never known nature. However, PLS model provides some transparency and a clear mathematical ground for understanding the internal working.

When two methods that consider multicollinearity are compared with each other through Blaine data, it is seen that PLS yields the lowest MAE's compared to PCR. This study reveals that the performance of PLS is comparable to other methods in case of multicollinearity. However, existence of outliers in the data, which is the beyond the scope of this study, can be considered as the achilles' heel of SIMPLS and use of alternative methods would be preferred.

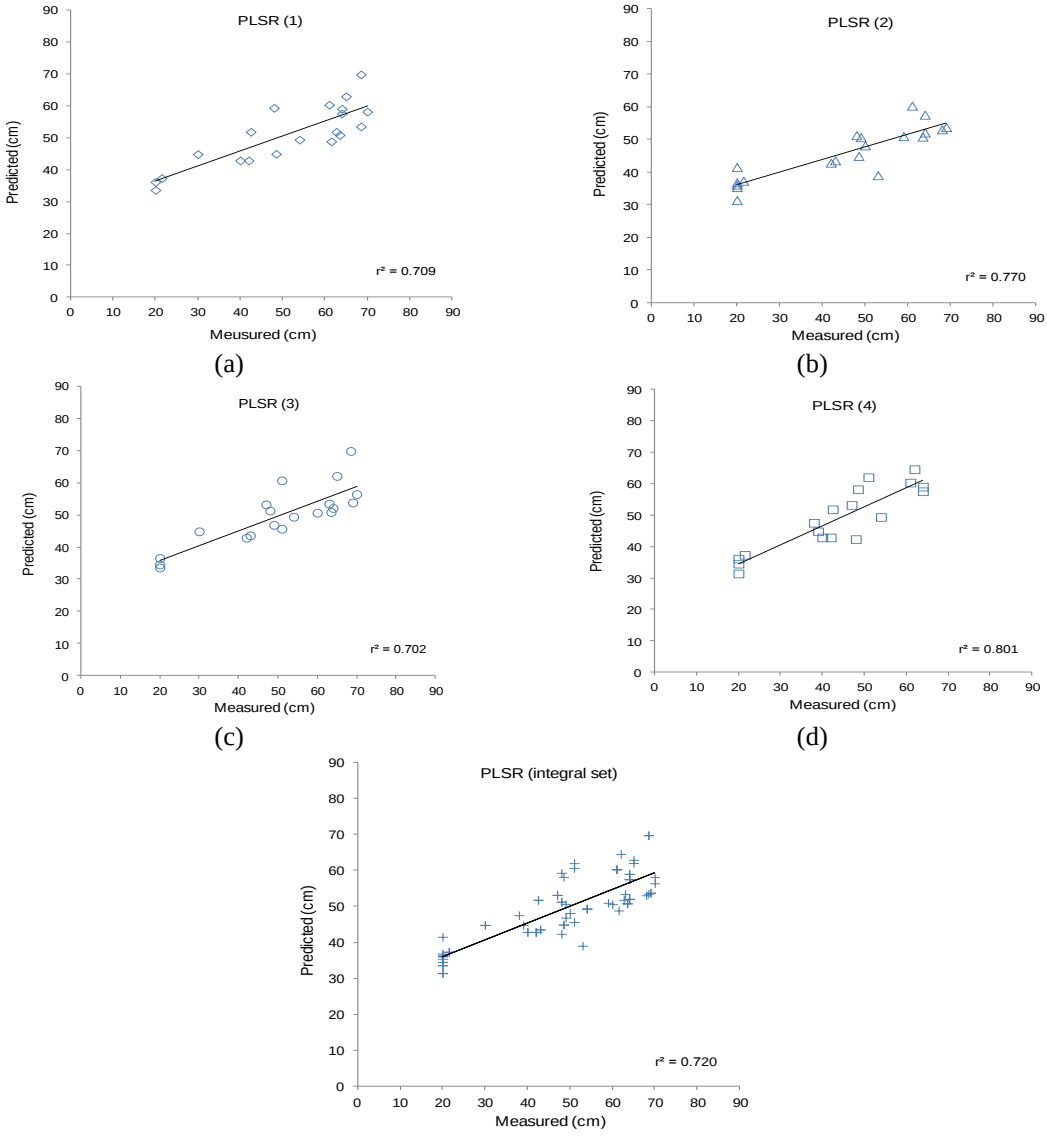
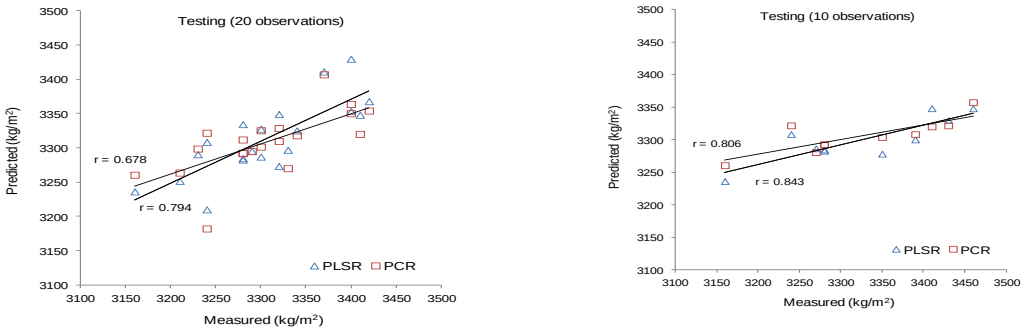


Fig. 7 PLSR performances for case study1



(a) (b)

Fig. 8 Performance plots for PLSR and PCR

5. Conclusions

Although the multiple regression analysis is a very important tool for cement and concrete industry, the multi-collinearity problem which describes the dependency between the independent variables handled in the system modeling works is not seriously considered. In addition, the model complexity has a practical importance and it is also disregarded in general. By this paper, use of the Partial Least Squares (PLS) method, which solves the problem of data collinearity and reduces the number of predictor variables observed in the problems of cement and concrete industry, was examined.

The prediction capacity of the method and its superiority was appraised by using two different type real problems. The outcomes and performance comparisons showed that the proposed algorithm produced better results comparing with the traditional regression models. In addition, the PLS algorithm can reduce the computation time and it also partly remove the smoothing problem of the prediction. As a consequence, it can be stressed that the PLS regression can be used for multivariate modeling problems in cement and concrete industry efficaciously.

Acknowledgements

The author would like to extend his appreciation to Gül Inan (Department of Statistics, METU) for the discussions on statistical points. He also would like to thank Adana (Oyak) Cement Factory and Ahmet Dag (Cukurova University) for the data set.

References

Ahangar-Asr, A., Faramarzi, A. and Javadi, A.A. (2011), "Modelling mechanical behavior of rubber concrete using evolutionary polynominal regression", *Eng. Comput.*, **28**(3-4), 492-507.

Chen, G.J. (2012), "A simple way to deal with multicollinearity", *J. Appl. Statistics*, 39(9), 1893-1909.

De Jong, S. (1993), "SIMPLS: an alternative approach to partial least squares regression", *Chemom. Intell. Lab. Syst.*, **18**(3), 251-253.

Dobrska, M., Wang, H. and Blackburn, W. (2012), "Ordinal regression with continuous pairwise

preferences", *Int. J. Mach. Learning Cybernetics*, **3**(1), 59-70.
Draper, N.R. and Smith, H. (1998), Applied Regression Analysis, John Wiley and Sons, USA.
Eswari, S., Raghunath, P.N. and Kothandaraman, S. (2011), "Regression modeling for strength and thoughness evaluation of hybrid fibre reinforced concrete", *ARPN J. Eng. Appl. Sci*., **6**(5), 1-8.
Li, Z. (2011), Advanced Concrete Technology, John Wiley & Sons, 528.
Liebmann, B., Filzmoser, P. and Varmuza, K. (2010), "Robust and classical PLS regression compared", *J. Chem.*, **24**, 111-120.
Martens, H. and Naes, T. (1989), Multivariate Calibration, Wiley, Chichester, UK.
Martins, J.P.A., Teofilo, R.F. and Ferreira, M.M.C. (2010), "Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets", *J. Chem*., **24**, 320-332.
Mehta, P.K. and Monteiro, P.K. (1993), Concrete. Structure, Properties and Materials, Prentice-Hall, New York, 496p.
Mevik, B.H. and Wehrens, R. (2007), "The pls package: principal component and partial least squares regression in R", *J. Statistical Softw.*, **18**(2), 1-24.
Miled, K., Limam, O. and Sab, K. (2012), "A probabilistic mechanical model for prediction of aggregates' size distribution effect on concrete compressive strength", *PHYSICA A – Statistical Mechanics and Its Applications*, **391**(12), 3366-3378.
Mohammadhassani, M., Nezamabadi-Pour, H., Jumaat, MZ., Jameel, M., Hakim, S.J.S. and Zargar, M. (2013), "Application of the ANFIS model in deflection prediction of concrete deep beam", *Struct. Eng. Mech.*, **45**(3), 319-332.
Musa, A.B. (2013), "Comparative study on classification performance between support vector machine and logistic regression", *Int. J. Machine Learn. Cybernet.*, **4**(1), 13-24.
Neter, J., Kutner, H.M., Nachtshein, C.J. and Wasserman, W. (1996), Applied Linear Statistical Models, McGraw-Hill, Boston, Mass.
Rasa, E., Ketabchi, H. and Afshar, M.H. (2009), "Predicting density and compressive strength of concrete cement paste containing silica fume using artificial neural networks", *Sci. Iranica Transact. A - Civil Eng.*, **16**(1), 33-42.
R Development Core Team (2008), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0
Rosipal, R. and Kramer, N. (2006), "Overview and recent advances in partial least squares", in Subspace, Latent Structure and Feature Selection, Statistical and Optimization, Perspectives Workshop, SLSFS 2005 ,Eds. C. Sauners, Lecture Notes in Computer Science, Springer.
Sahin, F. (2009), "Using of soft computing techniques in raw material and cement production processes", Msc Thesis, Cukurova University, Adana, Turkey (in Turkish).
Taha, R.O.H. (2012), "The possibility of using artificial neural networks in auditing-theoretical analytical paper", *European J. Economics, Finance and Administrative Sciences*, 47, 43-56.
Tutmez, B. (2009), "Clustering-based identification for the prediction of splitting tensile strength of concrete", *Comput. Concr.*, **6**(2), 155-165.
Tutmez, B. and Dag, A. (2012), "Regression-based algorithms for exploring the relationships in a cement raw material quarry", *Comput. Concr.*, **10**(5), 459-469.
Varmuza, K. and Filzmoser, P. (2009), Introduction to multivariate statistical analysis in chemometrics, CRC Press, Boca Raton.
Wold, S., Sjöströmi, M. and Eriksson, L. (2001), "PLS-regression: a basic tool of chemometrics", *Chem. Intell. laboratory Syst.*, **58**, 109-130.
Yeniay, O. and Göktaş, A. (2002). "A comparison of partial least squares regression with other prediction methods", *Hacettepe J. Math. Statistics*, **31**, 99-111.
Yeh, I.C. (2007), "Modeling slump flow of concrete using second-order regressions and artificial neural networks", *Cement Concrete Compos.*, **29**, 474-480.
Zarandi, M.H.F. (2008), "Fuzzy polynominal neural networks for approximation of the compressive strength of concrete", *Appl. Soft Comput.*, **8**(1), 488-498.

APPENDIX A

The pseudocode for the SIMPLS algorithm (Varmuza and Filzmoser 2009)

1. initialize $\boldsymbol{S_0} = \boldsymbol{X^T Y}$ and iterate steps 2 to 6 for *j=1,…,a*
2. if j=1, $\boldsymbol{S_j} = \boldsymbol{S_0}$; if j>1, $\boldsymbol{S_j} = \boldsymbol{S_{j-1}} - \boldsymbol{P_{j-1}}\left(\boldsymbol{P_{j-1}^T P_{j-1}}\right)^{-1} P_{j-1}^T S_{j-1}$
3. compute w_{ij} as the first (left) singular vector of S_j
4. $w_j = w_j / \|w_j\|$
5. $t_j = X w_j$
6. $t_j = t_j / \|t_j\|$
7. $p_j = X_j^T t_j$
8. $P_j = [p_1, p_2, \ldots, p_{j-1}]$

CC