مقارنة بين استخدام نموذج انحدار المربعات الصغرى الجزئية PLSR و انحدار المكونات الرئيسية PCR في العوامل المؤثرة على تمدد الاسمنت

الهام عبد الكريم حسين
قسم أنظمة الحاسبات
المعهد التقني / الموصل

الاستلام	القبول
٢٠ / ٠٦ / ٢٠١١	٠٢ / ١١ / ٢٠١١

Abstract

In this research, Two methods are used: Partial Least Squares Regression (PLSR) and Principal Components Regression (PCR) to build a model for autoclave cement on factors influnce on it. The comparison between these two methods is done by using two components for the PLSR and PCR, the plot of the fitted data shows that Partial Least Squares Regression represente the data better than Principal Components Regression, and R^2 insures this result which is showen by the figure. After that, 10 variables are used to compare these methods, this comparison indicates that the two methods represente the data in the same way. The goal is inreducing the number of components used in the two methods to avoid Over- Fitting, then it is depended on cross- validation method, this method indicates that Partial Least Squares Regression method is more economic than Principal Components Regression.

الملخص:

في هذا البحث تم استخدام طريقتي انحدار المربعات الصغرى الجزئية Partial Least Squares Regression (PLSR) وانحدار المكونات الرئيسية Principal Components Regression (PCR) في بناء نموذج تمدد الاسمنت على العوامل المؤثرة عليه، وقد تم مقارنة الطريقتين أولاً من خلال اخذ مكونين اثنين لكل من انحدار المربعات الصغرى الجزئية و انحدار

المكونات الرئيسية، بالاعتماد على رسم البيانات المطابقة فقد تبين ان انحدار المربعات الصغرى الجزئية يمثل هذه البيانات بطريقة أفضل من انحدار المكونات الرئيسية وان قيمة R^2 أكدت النتيجة التي تم التوصل اليها من خلال الرسم، ثم بعد ذلك تم استخدام عشرة متغيرات في مقارنة الطريقتين وتم التوصل إلى ان الطريقتين مثلتا البيانات تقريباً تمثيلاً متطابقاً، وبهدف اختزال عدد المكونات المستخدمة في الطريقتين وذلك لتجنب فرط المطابقة فقد أُعتمد على طريقة شرعية التقاطع Cross –Validation في هذا الاختزال الذي توضَحَ من خلال ذلك اقتصادية طريقة انحدار المربعات الصغرى الجزئية.

الهدف:

إجراء مقارنة بين طريقتي PLSR , PCR ومن خلال النتائج تُحدد الطريقة الأفضل في بناء نموذج انحدار.

الجانب النظري

١- المقدمة

ان تحليل الانحدار هو تقنية إحصائية لعرض العلاقة بين متغير معتمد يسمى متغير الاستجابة Response Variable ومتغير مستقل واحد او أكثر تسمى بالمتغيرات التوضيحية او التنبؤية Explanatory Variables or Predictors (Yan and Gang, 2009)، وفي حالة استخدام عدة متغيرات فان الب احث في هذه الحالة يواجه مشكلة التعدد في التحليل الاحصائي لذلك فان هناك طرائق متعددة أُكتشفت للتخلص من هذه المشكلة ، إحدى هذه الطرائق هو استبعاد بعض المتغيرات التنبؤية باستخدام طريقة تدريج الخطوة Stepwise او استخدام انحدار الحرف Ridge Regression وكذلك يمكن استخدام انحدار المكونات الرئيسية Principal Components Regression والتي يرمز لها بـ PCR للتخلص من هذه المشكلة وطريقة اخرى قريبة من طريقة PCR هي طريقة انحدار المربعات الصغرى الجزئية Partial Least Square Regression (Abdi, 2010) والذي يُرمز له بـ PLSR. ان PCR و PLSR هما طريقتان لنمذجة متغير الاستجابة عندما يكون لدينا عدد كبير من المتغيرات المستقلة وهذه المتغيرات تكون مرتبطة مع بعضها مكونة مشكلة الارتباط الخطي ، كلتا الطريقتين تعطي متغيرات مستقلة جديدة تعرف بالمكونات ، هذه المكونات تكون خطية متعامدة ومستقلة بعض ها عن البعض الآخر ، الاختلاف في الطريقتين هو ان طريقة PCR تكوّن المكونات الخطية لشرح التغيرات المشاهدة في المتغيرات المستقلة دون الأخذ بنظر الاعتبار

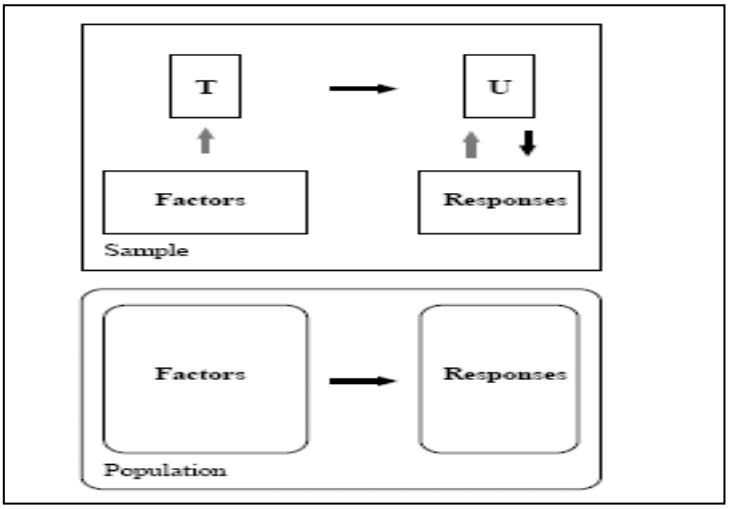
متغير الاستجابة، أما طريقة PLSR فإنها تأخذ بنظر الاعتبار متغير الاستجابة مما يؤدي الى الحصول على نموذج تنبؤي جيد من خلال عدة مكونات (Math Works. 2008).

٢- طريقة انحدار المربعات الصغرى الجزئية Partial Least Square Regression

ان طريقة المربعات الصغرى الجزئية Partial Least Square والتي يرمز لها بالرمز PLS هي تقنية عُمِمَت خصائصها من خلال تحليل المكونات الرئيسي ة Principal Components Analysis ويرمز لها بالرمز PCA والانحدار المتعدد Multiple Regression والذي يرمز له بـ MR هذا اذا كان لدينا متغير معتمد واحد ، كذلك اذا كان لدينا اكثر من متغير معتمد واحد فان طريقة انحدار المربعات الصغرى الجزئية مفيدة عندما نحتاج الى التنبؤ لمجموعة من المتغيرات المعتمدة عن طريق مجموعة كبيرة من المتغيرات المستقلة (Abdi, 2003)، وهذه الطريقة لا تستعمل التباين المشترك بين المتغيرات التنبؤية على عكس PCA لذلك فان هذا الأسلوب يعتبر منيع من مشكلة التعدد الخطي والتي تسمى في بعض المصادر بـ Curse of dimensionality (Lin and Tsay, 2004) ويمكن استخدام متغير معتمد واحد، عندها تسمى الطريقة PLS1 او متغيرين معتمدين فتسمى عندها الطريقة PLS2 (Rosipal, et al., 2006). اول من استخدم هذه التقنية هو العالم المتخصص بعلم الاقتصاد Herman Wold وذلك في العام ١٩٦٦ ثم أصبحت هذه الطريقة تستخدم في مجال الكيمياء عندما استخدمها الكيميائيون في تحليل تركيب عينة كيميائية ، كذلك استخدم PLS في مجال الكهرباء والحاسبات ولكنها في النهاية وُضِعت في اطار احصائي من قبل Friedman وذلك في العام ١٩٩٣ (Abdi, 200).

ان الانحدار المتعدد الخط ي (MLR) يمكن استخدامه مع مجموعة من العوامل والتي من خلالها يمكن بناء نموذج يكون مناسبا " لتمثيل البيانات لكنه في الوقت نفسه غير مناسب للتنبؤ، مثل هذه الظاهرة تسمى فرط المطابقة او المطابقة المفرطة Over – Fitting. في مثل هذه الحالة فعلى الرغم من وجود مجموعة م ن العوامل الظاهرة فان هناك مجموعة من العوامل المستترة التي تُحسب في متغير الاستجابة. الفكرة العامة في PLS هي مشابهة للفكرة السابقة أي انها محاولة انتزاع مجموعة من العوامل المستترة التي من خلالها يكون التفسير بقدر اكبر من العامل الظاهر ، بينما يتم نمذجة عامل الاستجابة جيداً ، ولهذا السبب فان مختصر PLS يعني " Projection to latent structure يعني الإسقاط إلى التركيب المستتر (Tobias, 2007)، من ثم نمذجة العلاقة بين هذه المجموعة من المتغيرات المستترة، والفرضية في كل طرق PLS هي ان البيانات المشاهدة تتولد بنظام أو عم لية تقاد بعدد قليل من المتغيرات

المستترة (Rosipal et al., 2006) والتي تعرف على انها تركيب خطي بين المتغيرات المفسرة ومتغيرات الاستجابة (Carrascal et al., 2009)، ويمكن توضيح طريقة PLS بالشكل التالي:



الشكل رقم (١): نموذج PLSR

الشكل يعطينا خلاصة تخطيطية للطريقة، حيث ان الهدف العام هو استعمال العوامل Factors لتتنبأ عن الاستجابات في المجتمع. ويتم هذا مباشرة عن طريق انتزاع متغيرات مستترة مثل T و U من عوامل العينة والاستجابات ، وهذه العوامل المنتزعة T (وأيضاً تشير إلى x القياسية) تستعمل للتنبؤ عن U او y القياسية وهذه التنبؤات y تستعمل لبناء التنبؤات عن الاستجابات . وفي السنوات الأخيرة كان الاهتمام من قبل الإحصائيين بإضافة خصائص إحصائية لـ PLS حيث رُبطت هذه الطريقة بطرق انحدار المكونات الرئيسية PCR وانحدار الحرف RR.

ان خوارزمية انحدار المربعات الصغرى ا لجزئية (Rosipal et al.,2006)و (Abdi,2003) هي:

نفرض ان خوارزمية PLS خطية لتشكيل العلاقة بين مجموعتين من المتغيرات تعرف بالشكل التالي: YCR^M , XCR^N. ان PLS هي علاقة بين هاتين المجموعتين والتي تسمى متجهات قياسية score vectors، بعد اخذ عينة من المشاهدات من كل من هاتين المجموعتين من المتغيرات . PLS يتألف من مصفوفة قياسية مثل X ذات بعد (n x N) بوسط صفر ومصفوفة قياسية مثلY ذات بعد (n x M) ذات وسط صفر وحسب المعادلتين التاليتين:

$$\left.\begin{aligned} X &= T P^T + E \\ Y &= UQ^T + F \end{aligned}\right] \qquad (1)$$

حيث ان:

T , U مصفوفتين قياسيتين تتضمنان متجهات مستترة Latent Vectors

Q , P : مصفوفتا التحميلات Loading برتبة N*P و M*P على التوالي (والتي تسمى في المكونات الرئيسية الجذور المميزة)

F, E: مصفوفتان ذات رتبة n*M و n*N تمثلان البواقي.

وان التباين المشترك T,U يتطلب إيجاد متجهات الأوزان wight vectors مثل w,c حسب المعادلة:

$$[\ cov(t,u)]^2 = [cov(Xw\ ,\ Yc)]^2$$
$$= \max_{|r| = |s| = 1} [\ cov\ (X_r\ ,\ Y_s)\ ]^{\ 2} \qquad (2)$$

حيث ان:

$$cov(t,u) = t^T u / n \qquad (3)$$

هو التباين المشترك بين المتجهات القياسية t,u

وان:

$$\left.\begin{array}{l} w = X^T u / (u^T u)\ ,\ \|\ w\ \| \rightarrow 1\ \ ,\ \ t = Xw \\ c = Y^T\ t / (t^T\ t)\ ,\ \|\ c\ \|\ \rightarrow 1\ \ ,\ \ u = Yc \end{array}\right\} \qquad (4)$$

وان u = Y اذا كانت صفوف Y (والتي عُرفت بـ M) اذا كانت تساوي ١ هذا يعني ان Y تمثل متجه ذو بعد واحد.

ان متجه الاوزان w يمثل اول متجه مميز Eigen Vector للقيمة المميزة التالية:

$$X^T\ YY^T\ Xw = \lambda w \qquad (5)$$

من المعادلة (١) فكما ذُكِرفان P,Q تمثلان مصفوفات التحميلات (p,q اذا كانت متجهات) وتحسب على اساس انها معاملات انحدار X / t و Y / u وحسب المعادلات التالية:

$$\left.\begin{array}{l} P = X^T t / (t^T t) \\ q = Y^T u / (u^T u) \end{array}\right\} \qquad (6)$$

فاذا كان لدينا مجموع مربعات الانحدار الجزئية بمتغير واحد PLSR1 او بمتغيرين PLSR2 او عدة متغيرات والتي يطلق عليها متعدد الابعاد Multidimensional، فان العلاقة بين X و Y هي علاقة غير متناظرة Asymantic، فان هناك فرضيتان:

١ – المتجهات القياسية $\{\ t_i\ \}^{p}_{i=1}$ هي متغيرات مفسرة جيدة لـ Y، حيث ان p تُعرف على انها عدد المتجهات القياسية.

٢ – العلاقة الداخلية الخطية بين المتجهات القياسية t و u هي علاقة موجودة exist كما في المعادلة:

$$U = T\ D + H \qquad (7)$$

حيث ان:

D: هي مصفوفة قطرية ذات بعد p*p.

H: مصفوفة البواقي.

وبالتعويض عن U في المعادلة (١) ينتج:

$$Y = TDQ^T + (HQ^T + F)$$
$$Y = TC^T + F^* \quad (8)$$

حيث ان:

$$C^T = DQ^T \quad (9)$$

وهي مصفوفة ذات بعد p*M وتمثل معاملات الانحدار.

وان:

$$F^* = HQ^T + F \quad (10)$$

وتمثل مصفوفة البواقي.

المعادلة (٨) تمثل معادلة مبسطة لـ Y، تستخدم انحدار مجموع المربعات الصغرى الجزئية مع المتغيرات المتعامدة المتمثلة بالمصفوفة T واتي تضم المتجهات القياسية t بحيث ان:

$$T^TT = I \quad (11)$$

وان:

$$C = Y^TT \quad (12)$$

مصفوفة ليست قياسية والتي يمكن الاستفادة منها في إعادة تعريف المعادلة (٨) بالاستفادة من المتغيرات الأصلية X وكما يلي:

$$T = XW(P^TW)^{-1} \quad (13)$$

حيث ان P مصفوفة متجهات التحميل كما معرفة في المعادلة (١)، عندها تصبح المعادلة (٨) كما يلي:

$$Y = X B + F^* \quad (14)$$

حيث ان B تمثل مصفوفة معاملات الانحدار

وبالتعويض عن T:

$$B = W (P^TW)^{-1}C^T \quad (15)$$

وبالتعويض عن المعادلة (4) وعن P^T في المعادلة (١) وعن C^T في المعادلة (٨) ينتج:

$$B = X^TU(T^T XX^T U)^{-1} T^T Y \quad (16)$$

إذن تصبح معادلة انحدار مجموع المربعات الصغرى الجزئية PLSR لتقدير القيمة $\hat{Y}$:

$$\hat{Y} = XB$$
$$= T T^T Y \quad (17)$$

وبالتعويض عن T^TY التي نحصل عليها من المعادلة (٨) ينتج:

$$\hat{Y} = T C^T \quad (18)$$

ولاختبار البيانات:

$$\hat{Y}_t = X_t B$$
$$= T_t T^T Y$$
$$\hat{Y} = T_t C^T \quad (19)$$

حيث ان:

X_t

و

$$T_t = X^T U (T^T X X^T U)^{-1} \quad (20)$$

تمثلان مصفوفتان لاختبار البيانات والمتجهات القياسية على التوالي.

شرعية التقاطع Cross – Validation

يمكن تعريف شرعية التقاطع على انها وسيلة إحصائية لتقييم ومقارنة خوارزميات التعلم، وذلك بتقسيم البيانات الى قسمين ، القسم الأول يستخدم في تعليم learn او تدريب train النموذج والقسم الثاني يستخدم لشرعية ا لنموذج validate the model (Refaeilzadeh et al., 2008).

نفرض ان y_i متغير استجابة معين وان $(\chi_{i1}, \chi_{i2}, \ldots, \chi_{ip})$ هو متجه مشاهدات انحدار، عليه يمكن كتابة هذه المشاهدات كما في المصفوفة التالية:

$$(y_i, X) = \begin{pmatrix} y_1 & \chi_{1,1} & \chi_{2,2} \cdots \chi_{1,p} \\ y_2 & \chi_{2,1} & \chi_{2,2} \cdots \chi_{2,p} \\ \ldots\ldots\ldots & & \\ \ldots\ldots\ldots & & \\ y_{n1} & \chi_{n1,1} & \chi_{n1,2} \cdots \chi_{n1,p} \end{pmatrix} \quad (21)$$

حيث ان $1 < n_1 < n$. ان طريقة شرعية التقاطع هي اختيار n_1 من البيانات المكونة من n من المشاهدات لبناء نموذج انحدار ، وتستخدم البيانات المتبقية $n\text{-}n_1$ لاختبار النموذج، حيث ان البيانات n1 تعتبر كعينة تعلم ، والبيانات المتبقية $n\text{-}n_1$ تمثل عينة اختبار . لأختبار النموذج، تستخدم البيانات المتبقية $n - n_1$، ويتم حساب مجموع مربعات اخطاء التنبؤ sum of square prediction errors كمعيار للاختبار، وحسب المعادلة:

$$\sum_{i=n_1+1}^{n}(y_i - \hat{y}_i)^2 \qquad (22)$$

حيث ان $\hat{y}_i$ تمثل القيمة المتنبأ بها من نموذج الانحدار من المجموعة الاولى من البيانات وهي n_1. الفكرة من استخدام هذه الطريقة لاختبار النموذج هي اذا كان نموذج المعلمات يلائم البيانات ككل فلن ذلك سيؤدي الى الحصول على قيمة صغيرة من مجموع مربعات خطأ التنبؤ للبيانات المتبقية $n-n_1$.

ويمكن استخدام مجموع مربعات الخطأ sum of square error أيضاً كمعيار للاختبار. بالإمكان تعديل الطريقة السابقة وذلك بالاختيار العشوائي لمجموعة جزئية subset من البيانات التي لدينا حجمها v لملائمة نموذج الانحدار، ثم تستخدم البيانات المتبقية n-v لحساب خطأ التنبؤ، ويمكن إجراء هذه العملية من الاختيار العشوائي لـ k من المرات وفي كل مرة يتم حساب مقدر estimate النموذج β باقل خطأ ممكن مثل $\beta_1, \beta_2, \beta_3, \ldots \beta_k$، بعد تقدير هذه المقدرات، نستخدم معدل هذه التقديرات وكما يلي:

$$\beta = \frac{1}{k}\sum_{i=1}^{k}\beta_i \qquad (23)$$

هذا المعدل يمكن اعتباره كمقدر لمعلمات الانحدار، ولاختبار النموذج يُستخدم معدل أخطاء التنبؤ، وهذا المعدل يعتبر كمعيار لتشخيص نموذج الانحدار . هذا الأسلوب يسمى شرعية تقاطع الطبقات v-fold cross – validation v وهذه الطريقة تعتبر أفضل من طريقة شرعية التقاطع لطبقة واحدة كما في عينة التعلم . تستخدم شرعية التقاطع عندما تتحقق المعادلة n>=2P+20، حيث ان p عدد المعلمات في نموذج الانحدار (Xin , et al., 2009).

الجانب العملي:

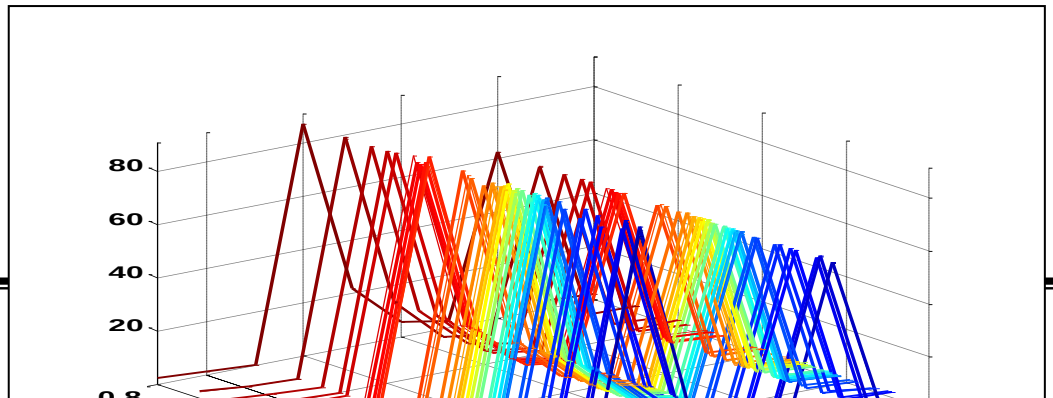
للوصول الى الهدف من البحث من حيث إيجاد الطريقة الأفضل بين PLSR و PCR في بناء نموذج انحدار بالاعتماد على عدة معايير، فقد أُخذت بيانات من معمل سمنت بادوش للفترة من كانون الاول / ٢٠٠٣ ولغاية حزيران / ٢٠٠٤ والمبينة بالملحق A، مع الاخذ بنظر الاعتبار شهر الصيانة الذي صادف في شهر كانون الثاني / ٢٠٠٤ والذي توقف فيه انتاج الإسمنت.

ان الاسمنت يتكون من بعض المواد الأساسية المتوافرة بصورة طبيعية من الحجر والرمل والجص وبعض الإضافات الأخرى اثناء عملية التصنيع التي تتضمن مواد تعمل على التغلب

على بعض المشاكل الفنية ومواد لزيادة بياض الاسمنت . يمتلك الاسمنت خواص كيميائية وفيزيائية ، حيث ان الخواص الكيميائية تتمثل بالمواد الكيميائية المتوفرة في المواد الاولية للاسمنت من الحجر و الرمل والجص مثل ثالث اوكسيد الكبريت So_3 واوكسيد المغنيسيا Mgo وثنائي اوكسيد السيليكون Sio_2 واوكسيد الكالسيوم Cao ...الخ وهي تمثل المدخلات Inputs، اما الخواص الفيزيائية فـتشمل النعومة والتصلب والتمدد والمتانة بالنسبة للاسمنت وهي تمثل المخرجات Outputs (نيفيل . ١٩٨٥). ويجب أن تحقق هذه المواد النسب المسموح بها والموضوعة من قبل بعض الدول الصناعية وتعتبر مواصفات عالمية والتي تؤثر على جودة الاسمنت . وكل مادة من المواد الكيميائية تؤثر بشكل مباشر او غير مباشر على الخواص الفيزيائية للاسمنت من النعومة والتصلب والتمدد والم تانة. بعد الانتهاء من إنتاج الاسمنت، يتم قبل تعبئته في الأكياس يأخذ عينات منه لإجراء عدد من الفحوصات للتأكد من جودته، من هذه الفحوصات على سبيل المثال: فحص نعومة الاسمنت، فحص مقدار التمدد الذي يجب أن يكون قليلاً في كل الحالات. والمسؤول عن تمدد الإسمنت هو أكسيد الكلسيوم الحر (الكلس الحر) أو أكسيد المغنزيوم، ولذا يجب أن تكون نسبة هذين الأكسيدين قليلة لا تؤثر في تمدد الإسمنت. وحددت المواصفة البريطانية نسبة أكسيد المغنزيوم بـ ٤% حداً أعظمياً، أما نسبة أكسيد الكلسيوم الحر يجب ألا تزيد على ١.٥%، ان أي زيادة في التمدد عن الحد المسموح به سيؤدي ذلك الى حدوث تشققا ت في الخرسانة المستخدمة في البناء مع مرور الزمن ، ويقاس التمدد بجهاز «لوشاتولييه» Le Chatelier باستعمال ملاط إسمنتي قياسي . ويجب ألا تزيد ثخانته على ١٠مم، كذلك فحص التصلب للاسمنت الذي يُعرف على انه الفترة التي تمر مابين إضافة المياه الى الاسمنت الجاف الى ال لحظة التي تفقد فيها الخلطة ليونتها . إضافة إلى فحوصات أخرى تُجرى على الاسمنت ليست موضوع بحثنا (الموسوعة العربية السورية. ٢٠١١)

تتضمن البيانات قيد البحث عشرة متغيرات حيث تمثل المدخلات الكيميائية وهي كما يلي مع رموزها : ثالث اوكسيد الكبريت SO_3، الفقدان بالحرق L.O.I، مواد غير قابلة للذوبان In.R، معامل الاشباع الجيري L.S.F، ثنائي اوكسيد السيليكون SiO2، اوكسيد الالمنيوم $AL2O_3$، اوكسيد الحديديك $Fe2O_3$، اوكسيد الكالسيوم CaO، اوكسيد المغنيسيوم MgO، والكلس الحر FrL، اما المخرجات الفيزيائية فتمثل تمدد الاسمنت Autoclave.

البيانات المذكورة يمكن تمثيلها بالشكل التالي:



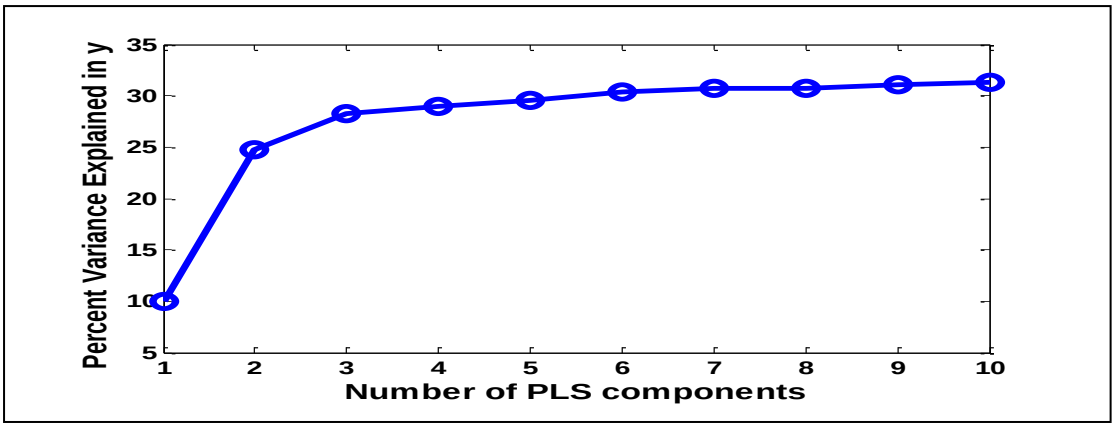
الشكل رقم (٢): بيانات الاسمنت من المتغيرات المفسرة ومتغير الاستجابة

حيث ان محور (independent) يمثل تسلسل المتغيرات التنبؤية وعددها ١٠ متغيرات ومحور (dependent) يمثل قيمة تمدد الاسمنت ، بينما يمثل المحور العمودي الثالث قيمة المتغيرات التنبؤية، حيث ان كل قيمة من قيم تمدد الاسمنت يقابلها (١٠) قيم، كل قيمة منها تمثل قيمة واحدة من قيم المتغيرات التنبؤية. وكما في المثال التالي:

y Autoclave	Frl	Mgo	cao	fe2o3	ai2o3	sio2	L.S.F.	In.R	L.O.I.	SO3
0.26	1.45	2.83	62.66	2.49	5.4	21.7	87.18	0.21	1.19	2.64

وقد تم الاعتماد على برنامج Matlab 2011 في رسم الأشكال المستخدمة في البحث وإيجاد النتائج.

نحتاج إلى عدد من المكونات لملائمة البيانات بشكل كافي وواحدة من هذه الطرق السريعة لاختيار عدد المكونات هو رسم نسبة التباين المفسرة من متغير الاستجابة كدالة لعدد المكونات وكما في الرسم أدناه:



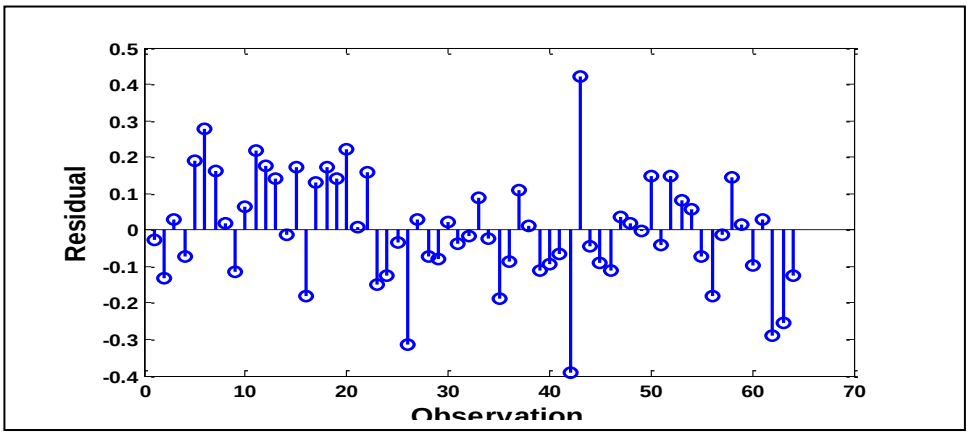
الشكل رقم (٣): نسبة التباين المفسرة لمشاهدات y لعشرة مكونات من مكونات PLS

حيث يتبين من الشكل ان المكون الثاني من مكونات PLS يشرح معظم التباين في مشاهدات y والمبينة قيمته ونسبته في الجدول رقم(١) إضافة إلى قيمة R^2 وقيمة MSE لـ PLSR:

الجدول رقم (١): قيمة معلمات نموذج PLSR وقيمة R^2 و MSE

betaPLSR10	MSE PLSR	R^2PLSR	explained variance in y for PLS2	Percentage of explained variance in y for PLS2
٠.٢٩٢٦	0.025842	0.312653	14.779	47 %

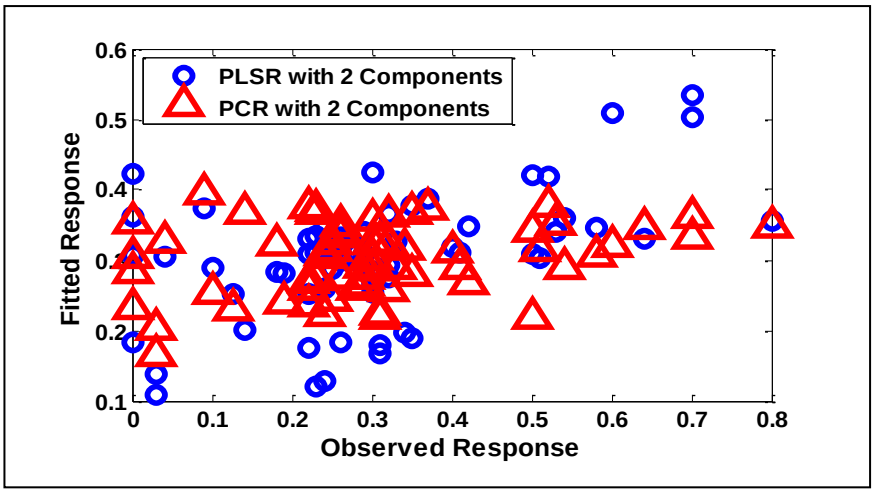
والشكل التالي يبين البواقي لمتغير y



الشكل رقم (٤): البواقي لمتغير y

عملياً فان اختيار عشرة متغيرات في بناء نموذج يعتبر عدد كبير نسبياً لذلك سوف يُلجأ إلى مكونين رئيسيين بعد ان نعمل على تحويل بيانات المتغيرات التنبؤية إلى الحالة القياسية في بناء PCR وكذلك PLSR وإجراء المقارنة بينهما.

لجعل نتائج PCR سهلة التفسير من ناحية البيانات الأصلية ، تحول إلى معاملات انحدار للبيانات الأصلية . ثم نعمل على مقارنة البيانات المطابقة لمتغير الاستجابة لكل من PLSR, PCR الذي يتوضح ذلك من خلال الشكل التالي:



الشكل رقم (٥): مقارنة بين المشاهدات الأصلية مع البيانات المطابقة لكل من PCR وPLSR بمكونين اثنين

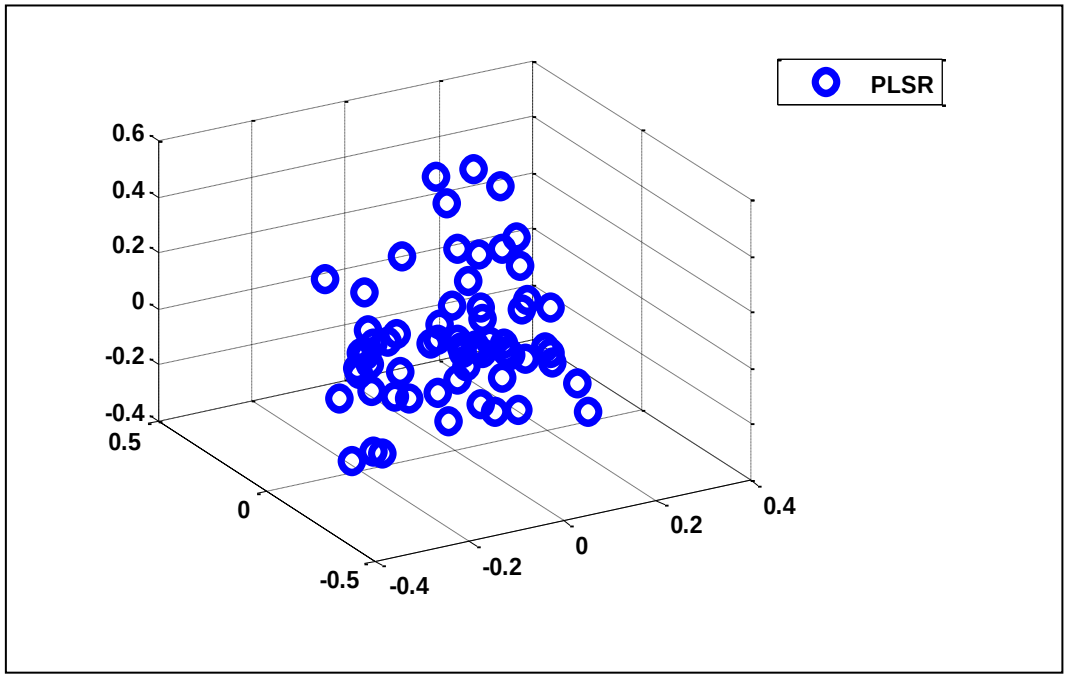
من الشكل يتبين ان الانتشار لقيم المطابقة لكل من PLSR, PCR تقريباً مختلفة، حيث ان الانتشار بالنسبة للمشاهدات حسب PLSR أفضل من PCR لمكونين اثنين ويتبين ذلك من خلال احتساب قيمة R^2 لكل من PLSR, PCR كما في الجدول:

الجدول رقم (٢): قيم R^2 لكل من PLSR و PCR

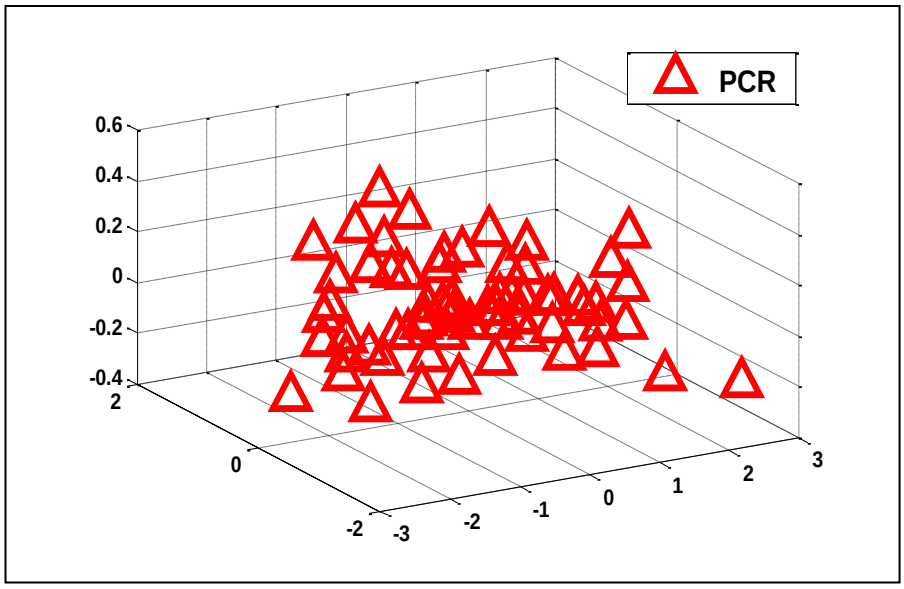
Method	R^2
PLSR	٠.٢٤٧٧
PCR	0.0832

من الجدول رقم (٢) يتبين ان قيمة R^2 لـ PLSR هي اكبر وأفضل من نفس القيمة بالنسبة لـ PCR، كمقياس عام لقيمة R^2 فان هذه القيمة بالنسبة للطريقتين غير مقبولة حيث ان نسبة ٢٤% قليلة جدا لتمثيل نموذج معين، لكن الهدف من البحث هو لتبيان ان PLSR أفضل من PCR ويوجد فرق بينهما.

و يتوضح من الشكليين التاليين بصورة اكبر ان انتشار المشاهدات المطابقة بالنسبة لطريقة PLSR أفضل بكثير من انتشار المشاهدات نسبة الى طريقة PCR:



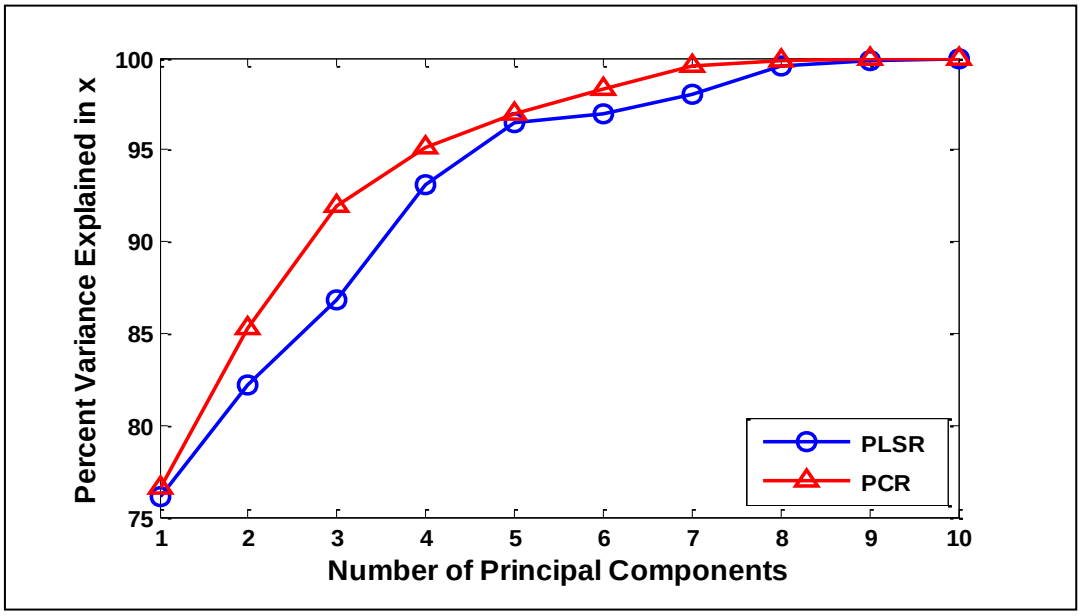
الشكل رقم (٦): البيانات المطابقة لطريقة PLSR



الشكل رقم (٧): البيانات المطابقة حسب طريقة PCR

حيث يتبين من الشكلين (٦) و (٧) ان النقاط منتشرة بشكل متقارب حسب PLSR بينما النقاط حسب PCR أكثر انتشاراً.

اما الطريقة الثانية للمقارنة بين PLSR, PCR هي برسم متغير الاستجابة ضد المكونان اللذان تبين من خلال الشكل (٣) ان المكون الثاني فسر معظم التباين الكلي هذا المكون على الرغم من انه الافضل في التفسير لمشاهدات y يتبين ان المكون الاول هو الافضل في تفسير التباين لمشاهدات المتغيرات التنبؤية في نموذج PLSR منه في PCR، كما هو واضح في الشكل رقم (٨):



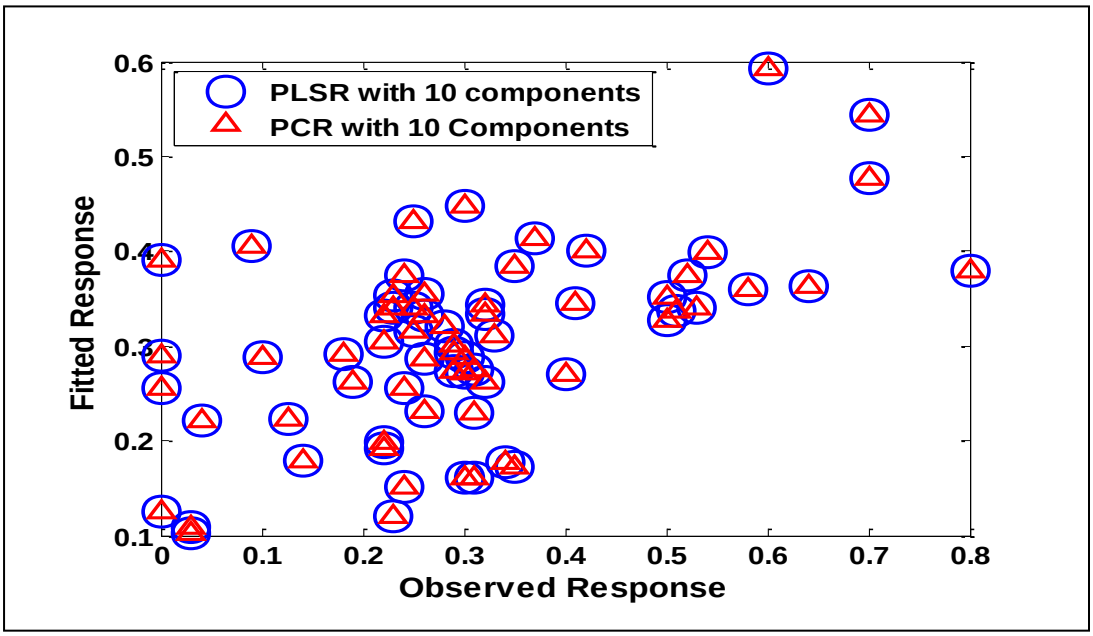
الشكل رقم (٨): عدد المكونات المفسرة لنسبة التباين في المتغيرات المستقلة لكل من PLSR و PCR

وهذه التباينات موضحة في الجدول رقم (٣):

الجدول رقم (٣): قيم التباينات المفسرة لكل من PLSR و PCR

Percentage of explained variance /PCR	Cumulative Percentage of explained variance /PCR	Percentage of explained variance / PLSR	Cumulative Percentage of explained variance/ PLSR
٧٥.٧٤٧	٧٥.٧٤٦٨	٧٦.٠٦٩	٧٦.٠٦٩
٨.٥٠٩٢	٨٤.٢٥٦٠	٦.١٣٣	٨٢.٢٠٢
٦.٦٧٣٩	٩٠.٩٢٩٩	٤.٦٥٩	٨٦.٨٦١
٣.٣٣١٢	٩٤.٢٦١١	٦.٢٠٣	٩٣.٠٦٤
١.٨١٩٥	٩٦.٠٨٠٦	٣.٣٦٥	٩٦.٤٢٩
١.٤٩٤١	٩٧.٥٧٤٧	٠.٥٦٦	٩٦.٩٩٥
١.١٨٧٧	٩٨.٧٦٢٤	١.٠٥٧	٩٨.٠٥٢
٠.٧٧٤٨	٩٩.٥٣٧٢	١.٥٦١	٩٩.٦١٣
٠.٣٠٣٩	٩٩.٨٤١١	٠.٢٢١	٩٩.٨٣٤
٠.١٣٤١	٩٩.٩٧٥٢	٠.١٦٦	١٠٠.٠٠
٠.٠٢٤٨	١٠٠.٠٠٠٠		

الشكل رقم (٥) بين ان المطابقة لمكونين اثنين كان أفضل عند استخدام PLSR منه عن PCR، والشكل التالي يبين الاختلاف في البواقي للطريقتين عند استخدام المكونات العشرة:



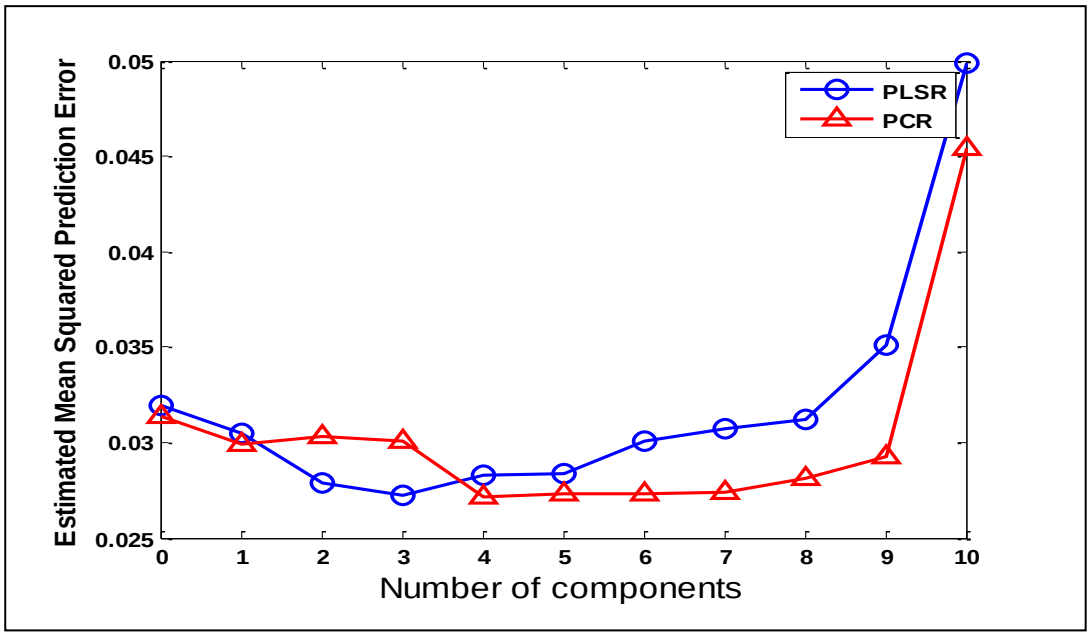
الشكل رقم (٩): البيانات المطابقة لعشرة مكونات لكل من PLSR و PCR

عند مقارنة الشكل رقم (٩) بالشكل رقم (٥) يتبين ان المطابقة لعشرة مكونات افضل من المطابقة لمكونين اثنين أي ان البواقي اقل في حالة استخدام المكونات العشرة من استخدام مكونين اثنين.

يبين الشكل رقم (٩) ان كلتا الطريقتين PLSR و PCR تقريباً متقاربتان في المطابقة ، مع هذا فان اختيار عشرة متغيرات هو اختيار عدد كبير نسبياً لذلك يمكن اللجوء الى طريقة بسيطة لتحديد اقل عدد من المكونات مع تقليل الخطأ ، وهذه الطريقة هي طريقة Cross – Validation.

اختيار عدد المكونات باستخدام Cross-Validation:

من المعتاد اختيار عدد من المكونات لتقليل الخطأ المتوقع عندما يتنبأ لمتغير الاستجابة باستخدام المتغيرات المستقلة . ان استخدام كمية كبيرة من المكونات قد يؤدي الى نتيجة مقبولة في مطابقة البيانات لكن هذا يقود إلى المطابقة المفرطة Over-Fitting لذلك فان cross-validation هي إحدى الطرق الإحصائية لاختبار عدد المكونات في PLSR و PCR. هذه الطريقة تجنبنا فرط المطابقة وذلك بحساب MSEP لكل من PLSR وPCR وحسب ما مبين في الشكل التالي:



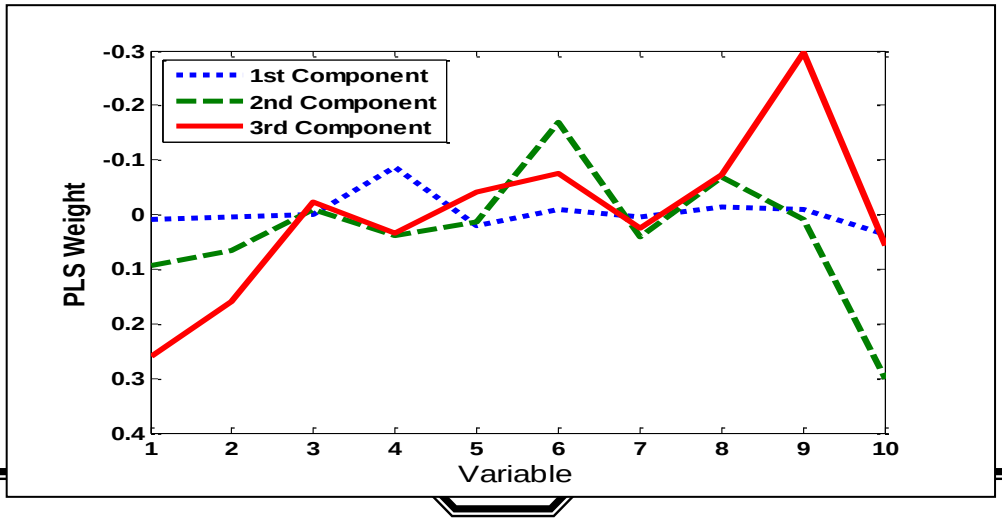
الشكل رقم (١٠): قيم MSEP لكل من PLSR و PCR

والجدول رقم (٤) يبين قيم MSEP المعروضة في الشكل رقم (١٠):

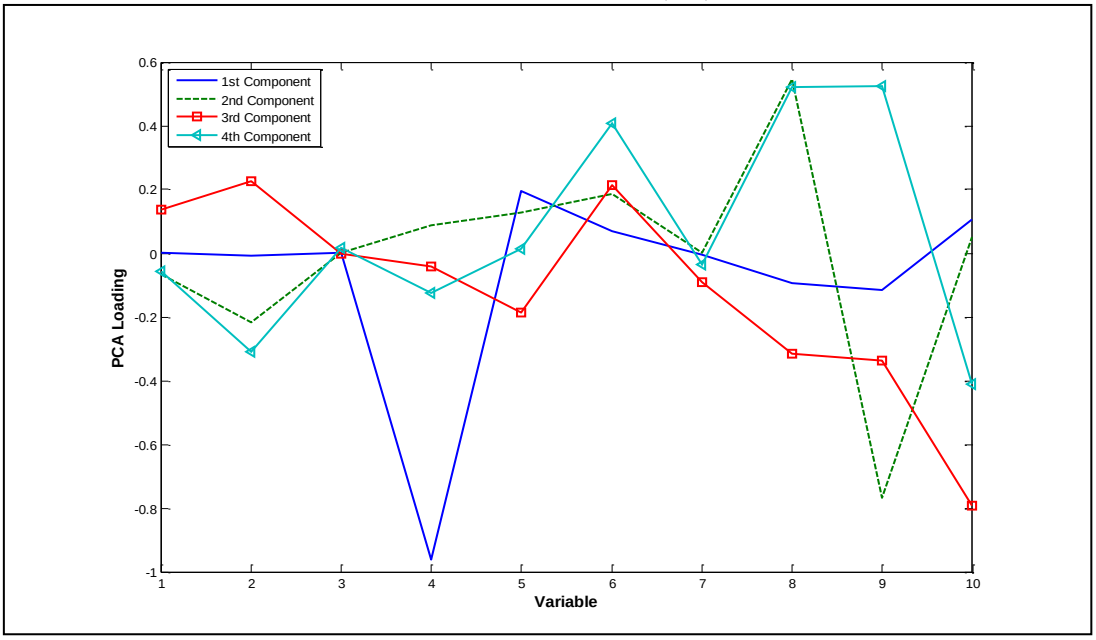
الجدول رقم (٤): قيم MSEP لكل من PLSR و PCR

MSEP - PLSR	MSEP – PCR
0.0317	٠.٠٣٢١
0.0301	٠.٠٣١٩
0.0285	٠.٠٣٢٤
0.0276	٠.٠٣١٥
0.0285	٠.٠٢٧٤
0.0293	٠.٠٢٨
0.0304	٠.٠٢٧
0.0316	٠.٠٢٦٩
0.0317	٠.٠٢٧٣
0.0376	٠.٠٢٩٧
0.0449	٠.٠٥٠٢

من الشكل (١٠) والجدول (٤) يتبين ان الخطأ بالنسبة الى طريقة PLSR يقل عند المكون الثالث بينما هذا الخطأ يقل في حالة PCR عند المكون الرابع ، وبما ان الهدف دائما هو استخدام اقل عدد ممكن من المكونات وبأقل خطأ هذا يعني ان النتائج التي ت م التوصل لها ان المكونات المستخدمة حسب PLSR هي أكثر اقتصادية من PCR. ويمكن رسم كل من المكونات الثلاثة بالنسبة إلى PLSR والمكونات الأربعة بالنسبة إلى PCR والتي تبين قوة كل مكون في PLSR و PCR اعتماداً على المتغيرات الأصلية واتجاهها. وكما في الشكلين التاليين:



الشكل (١١): المكونات الثلاثة حسب PLSR



الشكل رقم (١٢): المكونات الأربعة حسب PCR

الاستنتاجات:

في هذا البحث تم تطبيق PLSR و PCR على البيانات المأخوذة من معمل الاسمنت وقد تم التوصل إلى استنتاجات:

١) بدايةً تم مقارنة الطريقتين بعد اخذ مكونين اثنين لكل من PLSR و PCR وقد تبين من خلال الرسم بعد مقارنة المشاهدات الأصلية مع البيانات المطابقة ان PLSR أفضل من PCR.

٢) تم مقارنة الطريقتين المذكورتين بمكونين اثنين من خلال قيمة R^2 حيث ان:

$R^2PLSR = 0.2477$ وان $R^2PCR = 0.0832$

مما يدل على ان PLSR يمثل المشاهدات الأصلية بطريقة أفضل مما يمثلها PCR بمكونين اثنين.

٣) بعد استنتاج ان طريقة PLSR أفضل من PCR بمكونين اثنين فقد تم مقارنة الطريقتين بعد اخذ عشرة مكونات، ومن خلال النتائج تم استنتاج ان كلا الطريقتين تمثلان البيانات الأصلية تمثيلاً متقارباً بعد مقارنتها بالبيانات المطابقة.

٤) ولتجنب مشكلة فرط المطابقة ولان الهدف هو الحصول على نموذج بأقل عدد مكن من المكونات وبأقل خطأ ممكن وبالاعتماد على طريقة Cross – Validation فقد تم استنتاج ان اختيار ثلاثة مكونات حسب PLSR أفضل من اختيار أربعة مكونات حسب PCR وذلك اعتماداً على مقياس MSEP.

المصادر:

١) أي. ام. نيفيل (1985): "خواص الخرسانة"، ترجمة المهندس حقي إسماعيل محمد الجنابي، مدرس مساعد، المعهد الفني في البصرة، حقوق الطبع والنشر محفوظة لمؤسسة المعاهد الفنية.

٢) **هيئة الموسوعة العربية السورية. دمشق.** http: // www.arab-ency.com

3) Abdi, Herve; 2003: "Partial least squares (PLS) Regression", The university of Texas at Dalla, USA.

4) Abdi, Herve; 2010: "Partial least squares Regression and projection on latent structure regressioion (PLS regression)". School of Bahaviorall and Brain Sciences, USA.

5) Carrascal, Luis M., Galvan, Ismael and Gordo, Oscar, 2009: "Partial least squares Regression as an alternative to current regression methods used in ecolohy". Journal compilation Oikos 118: 681- 690.

6) Lin, Jin-Lung, Tsay, Ruey S. (2004): "Comparisons of forecasting method with many predictors" Department of Finance, National DongHwa University and Graduate School of Business, University of Chicago.

7) MathWorks, Inc., 2008. www.mathworks.com / trademarks.

8) Refaeilzadeh, Payam, Tang, Lei, Liu, Huan, 2008: "Cross – Validation", Arizona State University.

9) Rosipal, Roman, Krarner, Nicole, 2006: "Over view and recent advances in partial least squares". Springer-Verlag Berline Heidelberg, Germany.

10) Tobias, D. Randall, 2007: "An Introduction to partial least squares regression", SAS Institute Inc., Cary, NC.

11) Yan, Xin, Gang Su, Xiao, 2009: "Linear Regression Analysis, Theory and Computing". puplished by world scientific puplishing. Co. Pte. Ltd, London.

الملحق A: بيانات الاسمنت من كانون الاول / ٢٠٠٣ ولغاية نهاية حزيران / ٢٠٠٤

t	y Autoclave	Frl	Mgo	cao	fe2o3	ai2o3	sio2	L.S.F.	In.R	L.O.I.	SO3
1	0.26	1.45	2.83	62.66	2.49	5.4	21.7	87.18	0.21	1.19	2.64
2	0.24	1.68	3.03	61.85	2.5	5.43	21.59	88.41	0.21	1.25	2.54
3	0.26	1	2.7	62.4	2.44	5.47	21.72	88.76	0.2	1.21	2.45
4	0.26	1.87	2.51	62.92	2.46	5.89	21.98	86.5	0.2	1.02	2.54
5	0.53	1.75	2.94	62.5	2.42	5.65	21.86	86.81	0.21	1.08	2.65
6	0.64	1.75	2.7	62.23	2.44	5.83	21.79	86.98	0.2	1.08	2.64
7	0.34	1.15	2.9	62.37	2.4	5.68	21.8	87.73	0.2	0.72	2.52
8	0.33	1.83	2.82	62.69	2.42	5.37	21.75	88.34	0.2	0.79	2.53
9	0.22	1.82	2.71	62.69	2.45	5.04	21.62	89.73	0.21	0.79	2.52
10	0.41	1.7	2.71	63.01	2.48	5.38	21.72	88.16	0.23	0.79	2.87
11	0.58	1.95	2.96	62.69	2.57	5.46	21.99	87.94	0.2	0.64	2.5
12	0.35	1.38	3.2	63.12	2.4	5.86	21.43	88.51	0.21	1.22	2.56
13	0.54	1.9	3.3	62.28	2.56	5.47	21.47	88.77	0.2	1.14	2.67
14	0.29	1.88	3	62.47	2.4	5.86	21.34	88.44	0.2	1.2	2.58
15	0.51	1.45	2.83	62.25	2.44	5.4	21.7	87.91	0.2	1.19	2.64
16	0.04	1.97	3	62.35	2.4	5.94	21.9	87.53	0.2	0.7	2.15
17	0.4	2	2.79	62.84	2.44	5.79	21.92	87.69	0.2	1.41	1.99
18	0.5	2.3	3.24	62.69	2.5	5.38	22.16	87.16	0.21	0.87	2.12
19	0.3	1.78	3.01	62.84	2.43	5.48	22.02	87.96	0.21	0.75	1.99
20	0.7	2.68	3.01	62.53	2.63	5.2	22.13	86.69	0.2	0.92	2.28
21	0.6	2.63	2.4	62.84	2.61	5.36	22.04	87.36	0.21	0.95	2.73
22	0.7	2.5	2.92	62.24	2.46	5.37	21.77	87.53	0.2	0.95	2.8
23	0.3	2.33	2.7	63.01	2.47	5.64	22.25	86.41	0.21	0.85	2.59
24	0.23	1.69	2.65	62.31	2.68	5.67	22.01	86.36	0.2	0.7	2.6
25	0.35	2	2.46	62.61	2.48	5.72	22.19	86.19	0.21	1.1	2.4
26	0.09	2	2.99	62.58	2.6	5.98	22.16	85.67	0.21	0.91	2.65
27	0.3	1.72	2.96	63.11	2.55	5.68	21.81	88.23	0.23	0.8	2.56
28	0.03	1.43	2.9	63.26	2.48	5.77	21.36	90.29	0.2	0.79	2.26
29	0.03	1.48	2.7	63.69	2.42	5.9	21.22	91	0.22	0.75	2.46
30	0.22	1.2	2.79	62.69	2.46	5.79	21.38	89.1	0.23	1.41	2.5
31	0.14	1.01	2.7	62.26	2.43	5.77	21.94	86.32	0.2	0.75	2.59
32	0.24	1.5	3	62.01	2.56	5.94	21.26	88.16	0.23	1.1	2.67
33	0.24	1.16	2.7	62.9	2.4	5.81	21.18	89.83	0.2	0.7	2.76
34	0.32	1.8	3.1	62.85	2.45	5.66	21.36	89.21	0.21	1.1	2.8
35	0.1	1.8	3	62.81	2.67	5.55	21.42	89.3	0.23	1	2.5
36	0.22	1.44	2.92	61.97	2.45	5.71	21.87	86.42	0.2	1.27	2.5

37	0.23	0.94	2.53	62.52	2.42	5.93	21.37	88.62	0.22	1.05	2.56
38	0.3	1.26	2.43	62.11	2.4	5.75	21.6	87.55	0.22	1.65	2.5
39	0.18	1.45	2.53	62.38	2.43	5.59	21.77	87.46	0.23	1.14	2.59
40	0.26	1.64	2.87	62.5	2.38	5.49	21.96	87.15	0.21	1.2	2.6
41	0.25	1.5	2.8	62.08	2.4	5.55	21.63	87.66	0.2	1.2	2.6
42	0	1.74	2.7	62.43	2.49	5.53	22	86.76	0.23	1.26	2.6
t	y Autoclave	Frl	Mgo	cao	fe2o3	ai2o3	sio2	L.S.F.	In.R	L.O.I.	SO3
43	0.8	1.9	3	62.35	2.54	5.81	21.79	87.08	0.2	1.4	2.45
44	0.37	2.01	3.02	62.24	2.6	5.91	21.85	86.46	0.21	1.4	2.5
45	0.25	1.78	2.9	62.71	2.52	5.68	21.8	87.73	0.21	1.23	2.49
46	0.23	1.3	3.58	62.1	2.3	5.51	21.85	86.77	0.21	1.39	2.83
47	0.31	1.3	3.91	62.04	2.4	5.45	21.46	88.06	0.23	1.2	2.85
48	0.42	1.65	3.4	62.15	2.59	5.21	21.2	89.44	0.21	1.4	2.8
49	0.29	1.78	3.9	62.16	2.42	5.28	21.32	89.41	0.2	1.3	2.45
50	0.5	1.7	3.1	62.64	2.7	5.13	21.23	90.4	0.2	1.31	2.5
51	0.28	1.6	3	62.5	2.68	5.2	21.45	89.3	0.23	1.28	2.48
52	0.31	1.1	3.9	62.75	2.62	5.32	21.2	90.43	0.2	1.1	2.51
53	0.31	1.2	3.3	63.05	2.41	5.41	21.32	90.21	0.21	1.12	2.76
54	0.32	1.27	2.85	62.24	2.24	5.36	21.77	87.68	0.23	1.13	2.68
55	0.19	1.4	3	62.24	2.4	5.07	21.28	89.95	0.23	1.3	2.53
56	0.25	1.5	3	62.42	2.32	5.2	21.35	89.79	0.02	1.41	2.55
57	0.32	1.72	4.1	62.03	2.5	5.48	21.73	87.22	0.2	1.2	2.58
58	0.52	2.06	3.7	62.09	2.38	5.6	21.9	86.69	0.21	1.19	2.52
59	0.29	1.6	4.1	62.53	2.6	5.35	21.31	89.52	0.22	1.09	2.65
60	0.126	1.53	4	62.69	2.44	5.31	21.1	90.32	0.2	0.98	2.7
61	0.22	1.48	3.86	62.62	2.55	5.33	21.49	89.28	0.21	0.86	2.44
62	0	2.29	3.85	62.62	2.52	5.08	21.52	88.95	0.23	0.85	2.35
63	0	1.8	3.35	62.52	2.49	5.66	21.64	88.22	0.22	1.12	2.35
64	0	1.34	3.58	62.66	2.5	5.58	21.27	90.15	0.2	1.41	2.1