Arch. Min. Sci. **62** (2017), 2, 269-277

Electronic version (in color) of this paper is available: http://mining.archives.pl

DOI 10.1515/amsc-2017-0020

KAROL FIREK*[1], JANUSZ RUSEK*

PARTIAL LEAST SQUARES METHOD IN THE ANALYSIS OF THE INTENSITY OF DAMAGE IN PREFABRICATED LARGE-BLOCK BUILDING STRUCTURES

METODA CZĄSTKOWYCH NAJMNIEJSZYCH KWADRATÓW W ANALIZIE INTENSYWNOŚCI USZKODZEŃ BUDYNKÓW WIELKOBLOKOWYCH

The paper presents the research methodology aimed at determining the building damage intensity index as a linear combination of indices describing the damage to its individual components. The research base comprised 129 building structures erected in the large-block technology. The study compared the results of a standardized approach to data mining – PCA (Principal Components Analysis) with the procedure of the PLSR method (Partial Least Squares Regression). As a result of the analysis, a generalized form of the building damage index was obtained, as a linear combination of the damage to its components.

Keywords: Principal Components Analysis, Partial Least Squares Regression, mining effects, technical condition of building

W referacie przedstawiono metodykę badań, której celem było ustalenie wskaźnika zakresu intensywności uszkodzeń budynku, jako kombinacji liniowej wskaźników opisujących uszkodzenia jego elementów składowych. Bazą do badań było 129 budynków wzniesionych w technologii wielkoblokowej. W badaniach porównano wyniki standardowego podejścia do eksploracji danych PCA (*Principal Components Analysis*) z procedurą metody PLSR (*Partial Least Squares Regression*). W wyniku analiz uzyskano uogólnioną postać wskaźnika uszkodzeń budynku jako kombinacji liniowej uszkodzeń elementów składowych.

Słowa kluczowe: analiza składowych głównych, regresja cząstkowych najmniejszych kwadratów, wpływy górnicze, stan techniczny budynków

1. Introduction

Due to the impact of mining activities, the buildings located in mining areas require, inter alia, a detailed assessment of their technical condition. This survey is carried out to evaluate their resilience to mining effects and to determine the extent of the necessary preventive safety

* AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY, AL. A. MICKIEWICZA 30, 30-059 KRAKOW, POLAND

[1] CORRESPONDING AUTHOR: kfirek@agh.edu.pl

measures, as well as to identify the extent of possible mining damage. As far as the development of mining areas is concerned, multi-family and public utility building structures, erected in the large-block technology, are of a particular importance, both in technical and social terms. During the assessment of the technical condition of these objects, evaluation of the extent and the intensity of damage is especially difficult, both in the context of identifying the causes of its occurrence, as well as the methods for its repair (e.g. Wodyński, 2007).

The research presented in this article aimed at determining the building damage intensity index as a linear combination of indices describing the damage to its individual components (both structural and secondary).

The research base created by the authors comprised 129 building structures erected in the large-block technology, located in the Legnica-Głogów Copper District (LGOM). 22 indices were pre-specified for these objects, describing damage to all their structural and secondary components (Firek, 2009). Given the relatively large number of variables (22 damage intensity indices w_{ui}) and statistically significant mutual correlations between individual indices, it was determined that there was a need for the initial evaluation and selection of the variables for the modeling of a course of a building technical wear, planned for the subsequent stages of the research. Eventually, two methods were adopted, which enabled to determine the building damage index as a dominant principal component in the input variable space. The first one was the method of principal components (PCA – Principal Components Analysis), used on regular basis to deal with such issues (e.g. Osowski, 2013). The second method was based on the results of the NIPALS iterative algorithm (Nonlinear Partial Least Squares), used in the PLSR method (PLSR – Partial Least Squares Regression) (e.g. Geladi et al., 1986; Rosipal et al., 2006; KeeSiong Ng., 2013; Varmuža et al., 2009).

The PCA method is commonly used as a tool for reducing the dimensionality of data. It involves finding the values and eigenvectors of covariance matrix built for the full set of input variables. In the basic approach, with identifying the individual principal components, only the attributes in the input variable space are analyzed (Osowski, 2013). The method of identifying principal components in the PLSR method is a greater generalization in relation to the PCA method.

Basically, the PLSR method can be broken down into two simultaneous steps (e.g. Geladi et al., 1986). The first one involves identifying the principal components in the input variable space, and the second one – the relationships binding these components with the dependent variable. As part of the presented study, the PLSR method was used at the level of identifying individual principle components and their subsequent verification. Such a course of action allows to find one of the most efficient linear combinations between input variables.

A measure for the verification of individual principal components is the coefficient of determination R^2 for univariate regression models which were created based on the individual principal components identified using the PCA and PLSR methods. This measure gives an idea of a degree of contribution of a given principal component in explaining the variance contained in the output variables.

The advantages of the PLSR method, as compared with the PCA method, include among others (e.g. Wise B. M.):

– a possibility to identify any number of principal components, dependent exclusively on the number of variable attributes, not on the number of cases collected in the database,
– a possibility to iteratively identify principal components in the input variable space, while maintaining a relationship with the dependent variable,

- an identified set of principal components, due to the iterative procedure used, is insensitive to the deviation of the variables from the requirements set for the standard linear models,
- it is possible to analyze both continuous and categorical variables.

2. Methodology of the research

The **PCA** method involves determination of the values and eigenvectors of the covariance matrix. Therefore new, mutually uncorrelated linear combinations of input variables are obtained. It is aimed for each of the new variables to be characterized by the largest possible variance (e.g. Osowski, 2013, Wise B. M.).

The **PLSR** method assumes that both the variables from the input space **X** and output space **Y** can be written as the sum of the linear combinations of the vectors **t** and **p** as well as **q** and **u**, i.e. the so-called scores and loadings. These relationships have the following form (Rosipal et al., 2006):

$$\mathbf{X} = \sum_{i=1}^{a} \mathbf{t}_i \mathbf{p}_i = \mathbf{T}\mathbf{P}^T \tag{1}$$

$$\mathbf{Y} = \sum_{i=1}^{a} \mathbf{q}_i \mathbf{u}_i = \mathbf{Q}\mathbf{U}^T \tag{2}$$

where:

$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i, ..., \mathbf{x}_m\}$ — the input variable space,
$\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_i, ..., \mathbf{y}_m\}$ — the output variable space,
a — arbitrarily determined number of principal components,
$\mathbf{t}, \mathbf{p}$ — parameters of linear combinations: scores,
$\mathbf{q}, \mathbf{u}$ — parameters of linear combinations: loadings,
$\mathbf{T}$,Q — block scoring matrix (n X a)
$\mathbf{P}^T, \mathbf{U}^T$ — block loading matrix (a X m).

The PLSR method in the classical form based on the NIPALS iterative algorithm lead to finding vectors **w** and **c** (normalized vectors **t** and **u**) in order to consequently maximize the covariance between the projected input and output variable space to the normalized directions **t** and **u**, respectively (**w** and **c**) (KeeSiong Ng., 2013):

$$\left[\operatorname{cov}(\mathbf{t}, \mathbf{u})\right]^2 = \left[\operatorname{cov}(\mathbf{Xw}, \mathbf{Yc})\right]^2 = \max_{|\mathbf{r}|=|\mathbf{s}|=1} \left[\operatorname{cov}(\mathbf{Xr}, \mathbf{Ys})\right]^2 \tag{3}$$

As a result, using the NIPALS iterative procedure (Geladi et al., 1986), the decomposition of the vectors of the state of the input variables and the dependent variable (1) and (2) was obtained, as assumed at the beginning. The individual component vectors of the matrix **P** are the tangent coefficients of the subsequent principal components identified during the construction of the PLSR model.

3. Database of prefabricated large-block buildings

3.1. Technical characteristics of the study group of buildings

The research studies used the information collected during the building surveys carried out with the participation of the authors in the years 2002-2010 (surveying documentation, 2002-2010). A database of 129 residential and public utility buildings, not older than 35 years, located in the Legnica-Głogów Copper District (LGOM) was created. All the structures subject to the analysis were erected in the large-block technology, in the systems of large blocks (WBL) and large blocks for school buildings (SzWBL). Most of the structures were built as multi-segmented, split by expansion joints. In 93.8% of the cases, the surveyed buildings had no more than five storeys. The foundation of most buildings was on a constant level, mostly with a basement (85.3%). The dominant construction system (68.0%) was a system of transverse load-bearing walls. The analyzed structures were characterized by high rigidity, resulting from the reinforced concrete prefabricated structure made of the so-called Żerań brick. This system consisted of wall diaphragms, connected by vertical joints and stiffened horizontally by floor slabs, connected with a tie beam to the walls. The foundation and basement walls of all the studied buildings were made as reinforced concrete or concrete monolithic. The ceilings above basements, in most cases (89.1%), and all the ceilings in higher floors, were made as prefabricated reinforced concrete. According to the database, the buildings had a ventilated bipartite flat roof, primarily made of prefabricated roofing beams, based on honeycomb walls. Most of the surveyed buildings (96.9%) had preventive protection against mining impacts. These included reinforced concrete footings founded on a constant level with additional longitudinal reinforcement and reinforced concrete tie beams at the ceiling levels. The ceilings above the basements and last floors were monolithized with reinforced concrete topping and reinforced concrete shear studs in the corners of the basement walls.

3.2. Technical condition and the extent of damage to the surveyed buildings

The measure of the technical condition of buildings is the degree of wear. In this research, the degree of technical wear was determined for individual buildings using the method of weighted average, taking into account individual construction and technological solutions (e.g. Wodyński, 2007). Most of the studied buildings had the degree of wear within the range of 20-50%. This means that due to the aging processes, the technical condition of the buildings ranged between good and average (Wodyński, 2007).

In order to examine the importance of damage in the technical wear of each building, qualitative damage intensity index w_{ui} was determined for the individual structural and non-structural components (Table 1). This index was defined in (Firek, 2009) in a 6-point scale, in which $w_{ui} = 0$ means that the damage does not occur, $w_{ui} = 1$ – slight damage, $w_{ui} = 2$ – moderate damage, $w_{ui} = 3$ – intensive damage, $w_{ui} = 4$ (and 5) – very intensive damage.

Preliminary analysis of the value of the damage intensity index w_{ui} in the study group of buildings proved that most of the objects were damaged slightly or moderately.

In the analyzed group of buildings, the elements of the load-bearing structure did not exhibit any damage that could threaten the safety of the structure. The elements which were damaged to the greatest extent were elevation layers and load-bearing overground walls. For most objects (80.6%), the index of damage to the elevation layers w_{u17} was qualified as grade 2 or 3 (moderate or intensive damage).

4. Research results

Prior to the study, the input variables were analyzed for their variance The initial database comprised damage indices, sequentially w_{u1}, w_{u2}, to w_{u22}. Preliminary analysis of the variance of individual attributes allowed to leave the following damage indices for further analysis: w_{u2}, w_{u3}, w_{u7}, w_{u11}, w_{u12}, w_{u13}, w_{u17}. Other indices were rejected after the analysis of their diversity of values for all the buildings collected in the database.

Finally, a reduced set of damage indices was subjected to the PCA and PLSR analyses, performing calculations in Matlab (Matlab ...). As a result of these analyses, seven principal components each were obtained, respectively. Then, using the individual principal components, univariate simple linear regression models were determined for prediction of the degree of technical wear of buildings for which the values of coefficients of determination R^2 were determined.

In addition, for all the principal components obtained by the two methods, relative reduction in the raw variance contained in the input variables, due to the identification of respective principal components, was evaluated.

Tables 2 and 3 illustrate the linear formulas of the identified principal components. The results are presented for the standardized variables. Standardization of the variables allows for the assessment of relative contributions of the individual indices, introduced by specific damage to the appropriate principal component.

TABLE 1

Intensity indices of damage to components of a building structure
Source: own study

Designation	Index description
1	**2**
Components of load-bearing structure	
w_{u1}	intensity of damage to foundations
w_{u2}	intensity of damage to load-bearing walls of the basement or to foundation walls
w_{u3}	intensity of damage to interior and exterior overground load-bearing walls (including lintels and walls under windows)
w_{u4}	intensity of damage to firewalls
w_{u5}	intensity of damage to components of framed load-bearing structure (columns, transoms)
w_{u6}	intensity of damage to ceilings over basements
w_{u7}	intensity of damage to ceilings over higher levels, flat roof (roof covering)
w_{u8}	intensity of damage to stairs
w_{u9}	intensity of damage to balconies and loggias (as well as eaves and cornices)
w_{u10}	intensity of damage to roof structure
Secondary components (finishing elements)	
w_{u11}	intensity of damage to partition walls
w_{u12}	intensity of damage to internal plasters and wall cladding

1	2
w_{u13}	intensity of damage to floors (floor layers)
w_{u14}	intensity of damage to chimney walls
w_{u15}	intensity of damage to bracings (elements ensuring spatial rigidity)
w_{u16}	intensity of damage to infill (or curtain) walls
w_{u17}	intensity of damage to facade (facade layers)
w_{u18}	intensity of damage to damp-proof insulation
w_{u19}	intensity of damage to roofing
w_{u20}	intensity of damage to flashings, gutters and downpipes
w_{u21}	intensity of damage to joinery
w_{u22}	intensity of damage to exterior elements (entrances to buildings, platforms, terraces, wall trims, etc.)

TABLE 2

Summary of the forms of the subsequent principal componentsobtained by the PCA method
for standardized variables
Source:

Coefficients of the linear combination	**Forms of the subsequent principal components** $PC_i = a_2w_{u2} + a_3w_{u3} + a_7w_{u7} + a_{11}w_{u11} + a_{12}w_{u12} + a_{13}w_{u13} + a_{17}w_{u17}$						
	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
a_2	0,21	0,52	0,77	–0,09	–0,26	0,11	0,04
a_3	0,44	0,22	0,05	0,01	0,80	–0,18	–0,29
a_7	0,37	0,23	–0,26	0,79	–0,32	–0,11	–0,08
a_{11}	0,46	–0,28	–0,09	–0,16	–0,14	0,73	–0,37
a_{12}	0,44	–0,37	0,14	0,10	0,15	0,05	0,79
a_{13}	0,44	–0,25	–0,02	–0,40	–0,38	–0,63	–0,20
a_{17}	0,18	0,59	–0,56	–0,42	–0,07	0,12	0,34

Table 4, on the other hand, illustrates the results of a simulation and evaluation of univariate regression models for the individual principal components identified by each method.

To complement the analysis, Table 5 summarizes the degree of variance explained in the input variable space by the subsequent principal components by the PCA and PLSR methods.

Tables 2 and 3 highlight these principal components, which defined as one generalized variable in the simple linear regression model give the best degree of fitting, and thus the contribution of these components in explaining the variance of the output variables is the highest (cf. Tab. 4). It is important that the dominant principal component determined by the PLSR method is more effective in this regard. The efficiency, in this case, lies in the fact that one dominant component of the PLSR methods used in the simple regression model allows for an explanation of the variance of more than two principal components identified by the PCA method.

On the other hand, Table 5 summarizes the levels of explaining the variance contained in the input variable space. As it can be seen at this stage, the PLSR method has a higher degree of explaining the variance contained in the variables already at the level of the first principal component. Comparing both methods, it can be stated that in order to achieve the same level of explaining the variance of the input variables, the PCA method (to catch up with the result of the PLSR method) needs at least two principal components (cf. Tab. 5).

TABLE 3

Summary of the forms of the subsequent principal componentsobtained by the PLSR method
for standardized variables
Source:

Coefficients of the linear combination	**Forms of the subsequent principal components** $PC_i = a_2w_{u2} + a_3w_{u3} + a_7w_{u7} + a_{11}w_{u11} + a_{12}w_{u12} + a_{13}w_{u13} + a_{17}w_{u17}$						
	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
a_2	0,44	–0,04	–0,57	0,64	–0,63	0,09	0,05
a_3	0,46	–0,32	–0,10	–0,40	0,17	0,55	–0,56
a_7	0,40	–0,28	–0,26	–0,11	0,60	–0,66	0,43
a_{11}	0,20	–0,49	0,47	0,03	–0,16	–0,41	–0,44
a_{12}	0,10	–0,56	0,28	0,51	0,15	0,28	0,06
a_{13}	0,21	–0,46	0,46	–0,40	–0,40	0,10	0,55
a_{17}	0,58	0,22	0,31	–0,05	0,10	0,00	0,06

TABLE 4

Summary of the influence of individual principal componentsobtained by the PLSR and PCA methods
on the variance of the explained variable s_z
Source:

Number of the principal component	**PCA**	**PLSR**
	Values of the coefficient of determination R^2	
PC_1	2,48	8,16
PC_2	6,25	0,62
PC_3	2,03	0,01
PC_4	0,00	0,44
PC_5	0,04	0,75
PC_6	0,03	0,49
PC_7	1,11	0,01

TABLE 5

Summary of the influence of the principal components obtained by the PCA and PLSR methods
on the variance contained in the input space of the standardized variables
Source:

Number of the principal component	**PCA**	**PLSR**
	The degree of explaining variance contained in the input variables (%)	
PC_1	50,36	80,28
PC_2	23,24	14,55
PC_3	8,89	0,88
PC_4	8,07	1,42
PC_5	4,96	0,64
PC_6	2,54	1,02
PC_7	1,94	1,19

From the analysis of the contained results it can be concluded that:
– the PCA method can be used as a tool for reducing dimensionality, as evidenced by a high level of variance explained by subsequent principal components (cf. Tab. 5). However, its use in a situation when one generalized variable is sought for future model analyses is limited for the following reasons:
  • principle components obtained by this method are characterized by a certain dispersion. Comparing the results contained in Table 4, it can be noticed that the PCA method requires at least two principal components to be generated to equal the level of the influence on the explanation of the variance of the dependent variable with one, first component of the PLSR method; a similar situation occurs in the case of a variance analysis contained in the input variable space (cf. Tab. 5),
  • linear combination coefficients obtained by the PCA method, for the dominant principal component, take negative values (cf. Tab. 2), which is in contradiction to the observations and physics of the phenomenon (Wodyński, 2007),
– on the other hand, in the PLSR method, the principal components are generated so as to ensure a high level of explanation of the variance in both the space of the input variable and the dependent variable s_z (Tab. 4 and 5). The first identified principal component allows to explain over 80% of the variance contained in the input space, as well as, more importantly, helps to explain more than 8% of the variance contained in the raw set of the output variables s_z. In addition, the linear combination coefficients for the first, dominant principal component emphasize that increased intensity of damage is accompanied by an increase in technical wear.

5. Summary and conclusions

The subject of the study comprised data on damage to individual structural and secondary components of a group of large-block buildings located in the Legnica-Głogów Copper District (LGOM). The main objective of the analysis was to identify a methodology allowing for a preliminary data analysis and for identifying damage index of the whole building, which could be used to analyze the effect of mining impacts on the technical condition of this kind of a structure.

The methods in the field of Data Mining, i.e. the PCA and PLSR analyses were adopted for the research study.

The following conclusions were drawn therefrom.

It was found that the PLSR method was more efficient at the level of a preliminary data analysis than the PCA method. This is an effect of the iterative identification of principal components in the input variable space in the PLSR method, which occurs with a permanent relationship with the dependent variable.

It was proved that the first principal component may be the initial approximation of a generalized degree of damage to a prefabricated large-block building; it is generated by the PLSR method and takes the following form:

$$\begin{aligned} w_u = {} & 0,057 w_{u2} + 0,058 w_{u3} + 0,064 w_{u7} + 0,039 w_{u11} + \\ & + 0,015 w_{u12} + 0,037 w_{u13} + 0,044 w_{u17} + 7,03 \qquad (\%) \end{aligned}$$

This formula can be used to assess the technical condition of buildings erected in the large-block technology.

The article was prepared within the scope of the AGH Statutory Research no. 11.11.150.005.

References

Arkusze inwentaryzacyjne, 2002-2010. *Arkusze inwentaryzacyjne zabudowy miasta Polkowice i Lubin wraz z oceną odporności na wpływy górnicze, wykonane w Katedrze Geodezji Inżynieryjnej i Budownictwa AGH w latach 2002-2010*. Prace niepublikowane.

Firek K., 2009. *Proposal for classification of prefabricated panel building damage intensity rate in mining areas*. Archives of Mining Sciences, Vol. 54, Iss. 3.

Geladi P., Kowalski R.B., 1986. *Partial least-squares regression: a tutorial*. Journal: Analytica Chimica Acta – ANAL CHIM ACTA, vol. 185, no. 1, p. 1-17.

KeeSiong Ng., 2013. *A Simple Explanation of Partial Least Squares*. keesiong.ng@gopivotal.com. Draft, April 27, 2013.

Matlab. © 1994-2015 The MathWorks, Inc.

Osowski S., 2013. *Metody i narzędzia eksploracji danych*. Wydawnictwo BTC, Warszawa.

Rosipal R., Kramer N., 2006. *Overview and Recent Advances in Partial Least Squares*. *In Subspace, Latent Structure and Feature Selection Techniques*. Saunders C., Grobelnik M., Gunn S., Shawe-Taylor J. (eds.), Springer, p. 34-51.

Wise B.M., *Properties of Partial Least Squares (PLS) Regression, and differences between Algorithms*. Eigenvector Research, Inc. 3905 West Eaglerock Drive, Wenatchee, WA 98801.

Wodyński A., 2007. *Zużycie techniczne budynków na terenach górniczych*. Uczelniane Wydawnictwa Naukowo Dydaktyczne AGH, Kraków.

Varmuza K., Filzmoser P., 2009. *Introduction to Multivariate Statistical Analysis in Chemometrics, CRC, Boca Raton*.