ARCHIVES OF METALLURGY AND MATERIALS
Volume 53 2008 Issue 2

J. MORÁVKA[*], K. MICHALEK[**], B. CHMIEL[***]

STATISTICAL ANALYSIS OF HEATS WITH TARGETED OVERHEATING REALISED IN THE EAF AT TŘINEC STEELWORKS

ANALIZA STATYSTYCZNA WYTOPU Z CELOWYM PRZEGRZANIEM ZREALIZOWANA W PIECU ŁUKOWYM STALOWNI TŘINEC

Data sets were obtained from series of heats of tool and bearing steels produced in *electric arc furnace* (EAF) with subsequent treatment in *VOD unit* and targeted change (setting) of higher *overheating* of steel above the liquidus by chemical heating. The data can be divided to *input*, *process* and *output* quality parameters, which is typical division for so called *qualimetrics*. The paper gives a short overview with description of use of *methods of multi-dimensional statistical analysis* (exploratory analysis, analysis of time series, correlation and regression, multi-dimensional methods, PLS – Partial Least Squares) for analysis of parameters and links between quality indices for produced steel and input and set process parameters for production of steel in the EAF.

Keywords: statistical analysis, qualimetrics, targeted overheating of steel, electric arc furnace

Zestawy danych zostały pozyskane z serii wytopów stali narzędziowej i łożyskowej w elektrycznym piecu łukowym (EAF) z późniejszym nagrzewaniem w jednostce VOD i celowym przegrzaniem stali powyżej linii likwidusu przez nagrzewanie chemiczne. Dane mogą być podzielone na jakościowe parametry wejściowe, procesowe i wyjściowe, co jest typowym podziałem dla kwalimetrii. Artykuł przedstawia krótki przegląd z opisami użytych metod wielowymiarowej analizy statystycznej (analizy badawcze, analizy serii czasowych, korelację i regresję, metody wielowymiarowe, PLS – metoda częściowych najmniejszych kwadratów) do analizy parametrów i powiązań między wskaźnikami jakościowymi wytwarzanych stali, a parametrami wejściowymi i procesowymi produkcji stali w piecu łukowym.

1. Introduction

During the period 2006 and 2007 at the Třinec Steelworks (TŽ, Czech Republic), altogether 25 heats of tool steel and 51 heats of *bearing* steels were realised and data about these heats were included into the database. Steel was produced in the 12t electric arc furnace and in the VOD unit with targeted time dependent heat treatment of hot metal. The principle of the targeted time dependent heat treatment of hot metal consisted in realisation of pre-defined *overheating* of hot metal above the liquidus temperature, followed by a dwell at this temperature with subsequent cooling down to the temperature of casting.

Targeted overheating makes part of introduced technology of hot metal treatment, the substance of which is modification and degradation of so called atom arrangement at close distance, which can be achieved at certain overheating of hot metal above the liquidus temperature defined by time and temperature, or by exceeding of so called *critical temperatures*.

It is expected that application of this technology will enhance *formability of steel, reduction of contents of non-metallic inclusions* and *gas contents, increase of metal density* across the cast ingot cross section and that it will bring other positive qualitative changes.

Ultrasonic detection was used as indices of quality of produced steel. Micro-purity of samples form heats was determined in accordance with the *ASTM*. Outputs from optical emission spectrometer with application of the method *Spark-DAT* the results of which can be also considered as certain criterion of steel cleanliness were used for determination of character of non-metallic phases.

[*] TRINECKÝ INŽENÝRING, A.S., TRINEC, CZECH REPUBLIC – RESEARCHER (ING., PH.D.)
[**] VŠB-TECHNICAL UNIVERSITY OF OSTRAVA, CZECH REPUBLIC – PROFESSOR (PROF., ING., CSC.)
[***] TRINECKÉ ŽELEZÁRNY, A.S., TRINEC, CZECH REPUBLIC – RESEARCHER AND TECHNOLOGIST (ING.)

2. Structuring of quantities and data

In compliance with the concept of the Partial Least Squares method (PLS) and so called ***qualimetrics*** [1], it is appropriate to divide the considered quantities into *four* groups, namely into *input, process, failure* and *output* quantities, which are shown in the *fig. 1*:

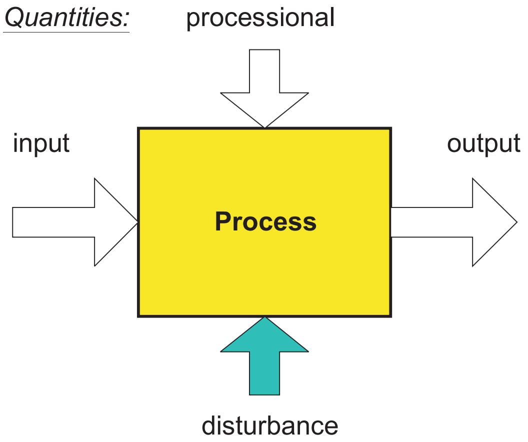

Fig. 1. Structuring of quantities at *qualimetric* analysis of data from an EAF heats

- ***Input*** quantities: physical - chemical properties of steel: chemical composition, liquidus temperature, temperature and mass of steel, etc.,
- ***Processional*** quantities: temperatures, pressures, duration of production process; for the given process these quantities (including two input quantities – steel temperature at the arrival for processing *Tprijezd* and liquidus temperature of steel *Tlik*) are shown in the *fig. 2* (so called *chronometry*, or time path of the process):

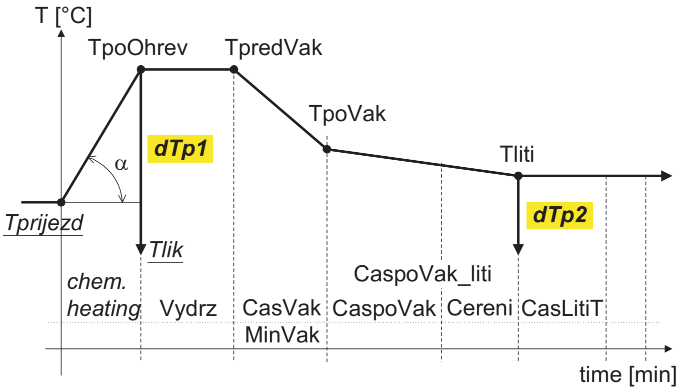

Fig. 2. Chronometry of the process – input and process quantities

These quantities can be furthermore divided into controlling quantities (in this case particularly overheating of steel above the liquidus *dTp* after chemical heating) and working quantities (the other ones),

- ***Output*** quantities: in case of qualimetrics these are the indices of product quality after its treatment. In case of steel they are *ultrasonic defects*, *outputs* from so called *micro-probes*, from the method *Spark-DAT*, parameters of cleanliness according to the standard *ASTM*, etc.,
- ***Disturbance*** quantities: they influence the product quality in an undesirable manner, in this case e.g. the fact of existence of exogenous inclusions originating from the lining and slag in the measured samples of steel (evidently mainly in the method *Spark-DAT*, where the samples of steel for an analysis were taken during casting of steel), as well as parameters of the next technological processing, namely ingot followed by their rolling into semis with an analysis from the viewpoint of *ultrasonic defects (UZ)* and cleanliness according to *ASTM* – see the *fig. 3*:

3. Exploratory analysis and analysis of time series

Basic, so called *Exploratory Data Analysis*, i.e. ***EDA*** – see e.g. literature [2]), describes random values (variables) with use of statistical indices (so called statistics) in numerical and graphical presentation in order to reveal their typical properties, as well as their peculiarities (“outliers”, etc.).
As it has already been mentioned, due to the fact that database records of heats are ordered chronologically, it is appropriate to make also a *Time Series Analysis* ***TSA*** for determination of possible dependence of parameters on time (influence of some systematic component), which could later at the regression analysis cause an auto-correlation of residues. The analysis is based on the so called Box-Jenkins methodology (classification), which defines so called *ARMA(p,q)* processes (AutoRegressive Moving Average of the order *p* and *q*) with the form of an Auto Correlation Function *(ACF)* and Partial Auto Correlation Function *(PACF)* – see e.g. [3], [4], [5]. Although the analyses *EDA* were *TSA* realised for all groups of quantities (input, process and output), for clearness we give it only for the selected representatives of these groups – see the *tab. 1.*

Heat (series, records) data were ordered according to an increasing heat number (and thus by time, i.e. chronologically), which means that they can be regarded as *time series* of quantities.

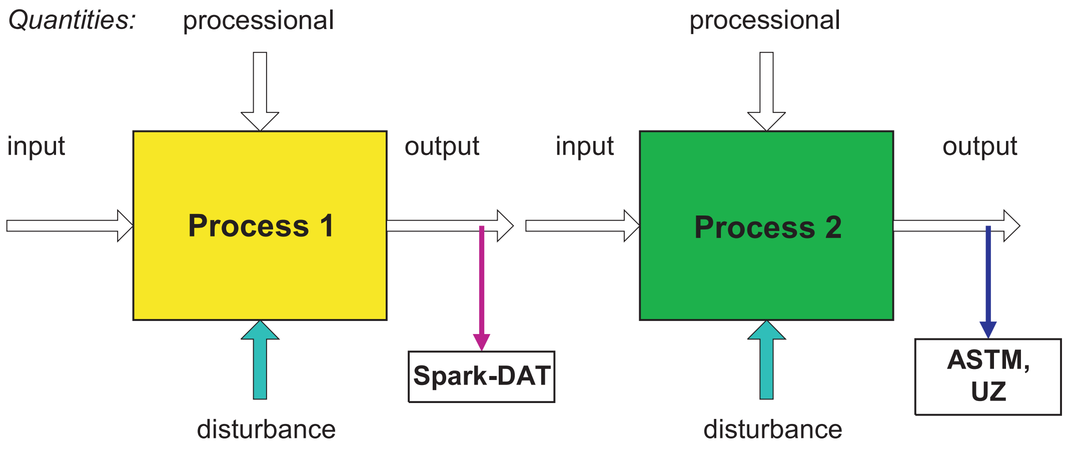

Fig. 3. Diagram of taking of samples of steel for determination of cleanliness (Quality)

TABLE 1

Results of analyses EDA and TSA for representatives of groups of quantities

Quantity	CV [%]	Distribution	Time series	Note
Alc	38.3	Gamma	WN	Analysis of total Al in steel
dTp	14.7	Normal	WN / AR(2)	Controlled overheating of steel above the liquidus
UZ2	48.2	Bi-modal	WN	Ultrasonic findings of defects

Legend:

- CV – Coefficient of Variation [%], or relative standard deviation calculated as portion of a standard deviation and arithmetic average,
- Gamma – transport (towards higher values) of skewed probability distribution,
- Bi-modal – with double modus (the most frequent values, peaks),
- WN – so called "white noise", i.e. non-correlated (internally independent) random signal,
- AR({p}) – autoregressive process only of the order p = 1, 2, . . . , k.

It is generally valid that the value of CV smaller than approx. *10÷20%* manifests small relative variability of the variable and its effect in statistical analysis can be covered by noise or other influences.

On the other hand, the values of CV higher than 100% or even 200% can indicate "outliers" (extremes), or great data incongruity caused e.g. by few positive values of the variable and the rest with zero values.

Fig. 4 shows time behaviour, histogram and ACF/PACF functions of *dTp*, i.e. the measured values of the controlled overheating of steel:

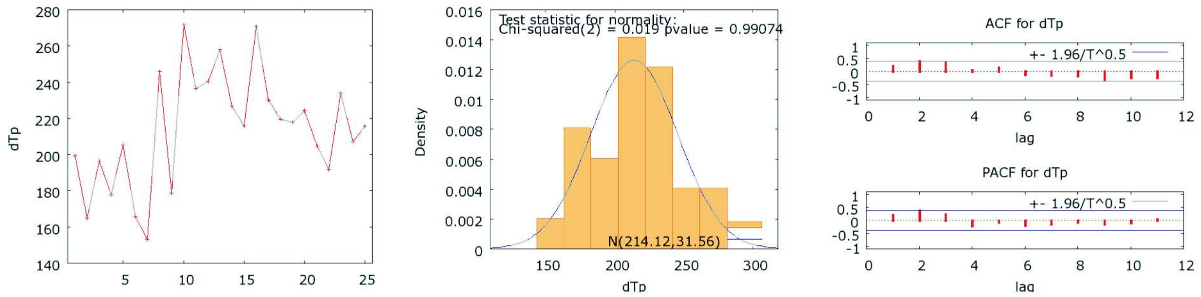

Fig. 4. Time behaviour, histogram and ACF/PACF of dTp

It is obvious from the *fig. 4* that controlled overheating in heats of tool steel had an unsteady (in the mean value) time path, however its distribution was normal (Gauss). It is possible to guess from the development of ACF and PACF that it was auto-regressive process of the 2^{nd} order AR(2), or AR(3). It means that targeted setting of overheating by a technologist for the given heat depended in some extent on the values set in two (or even three) previous heats and that it therefore was not completely random (independent).

4. Correlation analysis

Before the multi-dimensional analysis itself it is appropriate to make between the selected variables of groups yet a *correlation* analysis. This analysis serves also for defining of the next use of multi-dimensional methods – if the pairs of variables in the group of quantities do not contain a correlation coefficient $R > 0.8$, it is useless to apply the multi-dimensional method.

For example for a group of variables *Spark-DAT* the following pairs showed the highest correlations ($R > 0.80$): MGO ↔ CAMGO with R = +0.95, ALO ↔ ALCAO with R = +0.93, MGALO ↔ CAMGO with R = +0.86, MGO ↔ MGALO with R = +0.83. It is possible to plot on the basis of this overview so called *correlation diagrams* of variables – see the *fig. 5*:

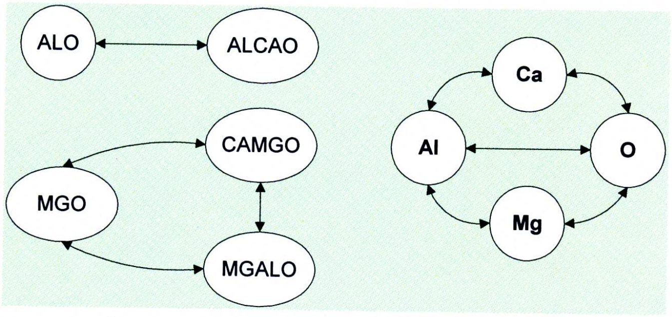

Fig. 5. Correlation diagrams of variable of the analysis Spark-DAT's elements

These correlations indicate influence of chemical „clustering“ of some chemical elements in steel, namely {Al, Ca, Mg and O}, as well as of oxides. The reason that in case of the method *Spark-DAT* the samples taken from steel are for a short moment (approx. 2 s) exposed to influence of ambient air, i.e. there is unfavourable influence of reoxidation of steel samples at taking of these samples.

5. Multi-dimensional analysis

Multidimensional Data Analysis - MDA is suitable for searching the principal and at the same time aggregated indices obtained by reduction of data dimensions with certain acceptable loss of information – see e.g. [6], [7].

MDA is based on so called *latent* variables, which are linear combination of original variables [8], [6].

This analysis contains a comparatively large set of methods – see [7]. Well-arranged structuring of methods according to types of objects, data and needs is given in [6] or in [9].

For analysis of qualimetric data the following suitable methods were used: *Principal Component Analysis - PCA* and *Cluster Analysis - CLU*.

5.1. Principal components method (PCA)

As an example of use of the *PCA* we give processing of the set of *temperatures* from the group of process quantities. The *fig. 6* shows *"scree" plot of eigenvalues* and *diagram of component weights* for six temperatures and for the temperature of overheating above the liquidus at the end of chemical heating:

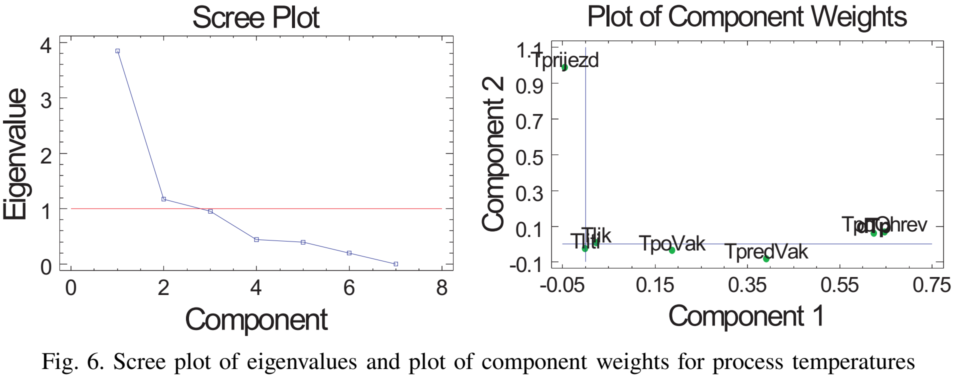

Fig. 6. Scree plot of eigenvalues and plot of component weights for process temperatures

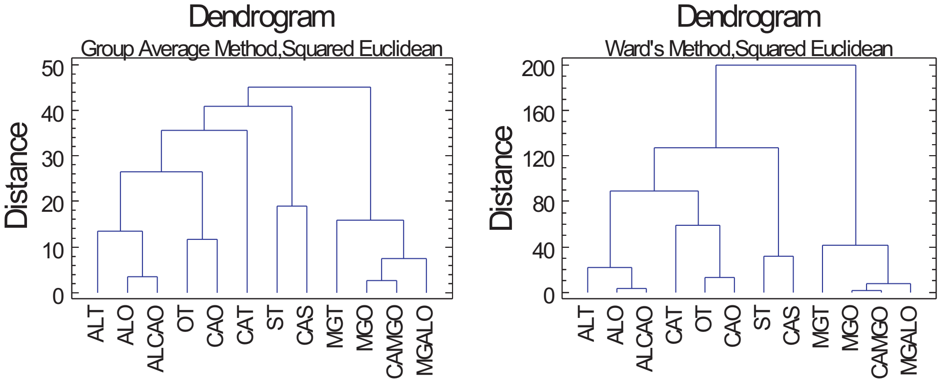

Fig. 7. Dendrogram of Spark-DAT variables – average method and Ward's method

It is obvious from the *fig. 6* that process temperatures can be approximately (in a simplified manner, but sufficiently and "usefully") described by 2 principal components, which describe approx. 70 % of variability (scatter) of the set of temperatures.

The first principal value is best characterised either by *overheating* (dTp) or by temperature *after chemical heating* (TpoOhrev), between which, however, exists a perfect multi-collinearity.

The second principal value contains (in case of original variables) a dominant variable *"arrival" temperature*, which is independent on other temperatures.

5.2. Cluster analysis (CLU)

Cluster analysis was used e.g. for variables of the method *Spark-DAT*, where it provided for the variables (attributes), i.e. for individual variables of chemical elements of impurities and for raw data the results visible in so called *dendrograms* – see the *fig. 7*, representing the clusters only for 12 selected variable elements in standardised values:

Both dendrograms indicate for all non-zero variables of elements the logical fact already mentioned in the previous exploratory analysis, that it that there are *four* or *three* dominant cluster. These clusters correspond to the results of the previous correlation analysis.

6. Regression analysis

Multiple linear regression analysis belongs in some sense also to the methods of multi-dimensional analysis. Its advantage consists in a long-term and extensive elaboration of various diagnostic methods for definition and classification of regression quality.

When multi-dimensional linear regression was applied on the analysed data the explained variable (regressand) was the representative of the *index of quality,* and the explaining variables (regressors) were both the representatives of *input* quantities, and the representatives of *process* quantities.

The results of multiple linear regressions for the regressand, i.e. for the explained variable *UZ2*, are summarised in the *tab. 2*:

TABLE 2

Results of regression with selected regressors for the quality index *UZ2*

Regression coefficients b_j						F-test	R^2 [%]	DW	Note
b_0	Alc	dTp	T poOhrev	T prijezd	x				
(1222)	(-874)	(+1.13)	–	(-0.80)	–	(1.66)	23.8	1.960	
(-513)	(-757)	- -	(+1.18)	(-0.79)	- -	(1.85)	25.8	1.979	
-62 617	–	- -	–	–	43.08	29.0	61.7	2.185	x = Tlik
- -	- -	- -	- -	- -	5.49	160.3	89.4	2.088	x = CaspoVak

Legend:

- b_j – regression coefficient of variable, b_0 is an absolute term,
- F-test – value of F-statistics of the model significance as a whole,
- DW – value of Durbin-Watson' statistics of auto-correlation of the 1st order of residues.

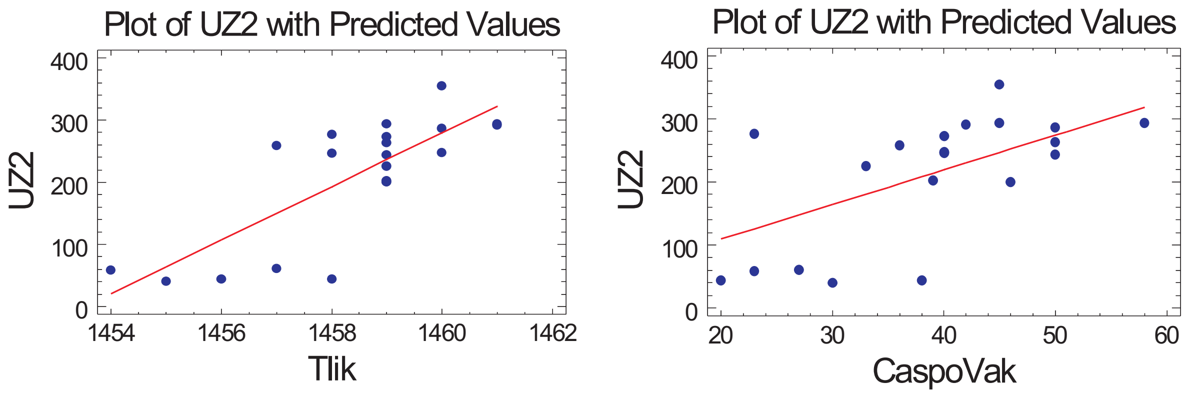

Fig. 8. Prediction diagrams of the model *UZ2* with regressors *Tlik* and *CaspoVak*

The *fig. 8* shows prediction diagrams for the last two models in the *tab. 2*. It is obvious from this picture, that although the determination coefficients are comparatively high, i.e. approx. 62 % and 89 %, due to small volume of data the prediction diagrams (so far – due to lack of data from heats) are no too "tight" and convincing.

Apart from so-called „classical" *static* regression we considered also *dynamic* regression [5], which contained so-called delayed variables of regressors and regressands. This regression makes it possible to intercept auto-correlation of the process variables caused by dynamics and inertia of the process. Dynamic regression analysis was used at static regression models, in which auto-correlation of residues occurred. Dynamic regression confirmed in principle the results of static regression, while in some cases it gave the determination coefficient (regression rebate) higher even by 20%, and residues of dynamic regression models fulfilled the require pre-requisites of independence, homoskedasticity and normality.

7. Partial least squares

Lately the method *Partial Least Squares – PLS* is getting used more and more often (and introduced into static SW) – see [1]. This method is based on a synthesis of the principle of the *principal components method* and *multiple linear regression*.

A peculiarity (and certain *advantage*) of the *PLS* method in comparison with other statistic multi-dimensional methods is the fact that number of data samples can be even smaller than number of regressors, i.e. *n*¡*p*. Especially in case of regression analysis an opposite relation is strictly required, that is *n* >> *p*, i.e. that number of data must be substantially larger (at least double by preferably by an order) than number of regressors.

PLS was used for a set of regressands of the *Spark-DAT* method and a set of regressors, containing both *input* and *process* quantities. The optimum number of 7 components was established with use of a diagram for comparison of models by the method of so called cross validation – see the *fig. 9* on the left.

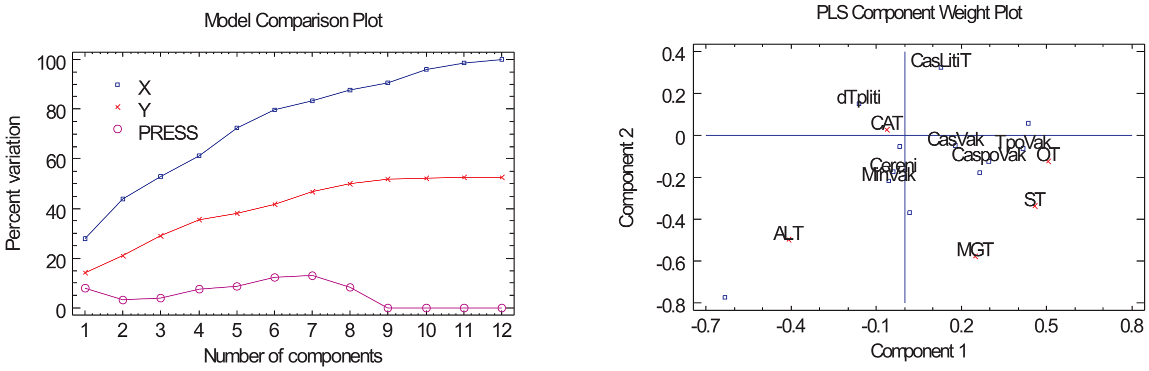

Fig. 9. Diagram of comparison of models and plot of component weights of the PLS method

Diagram *PLS* of component weights (*fig. 9* on the right) shows that total oxygen *OT* correlates with process variables temperature and time for vacuum treatment (*TpoVak*, *CaspoVak*), *CAT* is close to the beginning and therefore does not have big variability or weight, variables of the elements *ALT*, *MGT* and *ST* are comparatively distant from the process variables.

In conclusion it can be said that the *PLS* method provided for the regressand *ALT* and *MGT*, and their statistically significant models, practically of the same value of regression coefficients as *multiple regression analysis*, which, however, indicated insignificant models.

The *PLS* method has, nevertheless, several „degrees of freedom“, it is more complex and it lacks rich regression diagnostics – which are its *disadvantages* in comparison with multiple regression analysis.

It seems that the main advantages in comparison to regression remain, however, a possibility of determination of dependencies simultaneously fro the whole group of regressands and possibility of calculation even for number of data smaller tan the number of regressors.

8. Conclusion

On the basis of results of multi-dimensional statistical analysis of heat data sets ordered by time for tool steel (25 heats) and bearing steel (51 heats) with targeted temperature-time treatment of hot metal, i.e. with targeted change (setting) of higher *overheating of steel* above the liquidus temperature (after termination of chemical heating), it is possible to draw the following conclusions:

- solution of the issue of finding dependence of quality on the values of input and process values belongs to the discipline of so called ***qualimetrics***, which is certain specific area of statistical analysis, using the whole set of uni-dimensional or multi-dimensional statistical methods,
- before the statistical analysis, during its execution and also at its completion it is necessary to ***consult*** with a specialist-technologist all technological connections and also correct technological interpretation of obtained results,
- the principal ***imperfections of data*** were the following ones: *small scope*, *insufficient variability of quantities*, *non-planned* (non-orthogonal) *experiments*, *missing quantities* (input and process),
- ***multiple linear regression analysis*** provided us (and it generally provides) the biggest volume of "useful" basic and diagnostic bits of information – more than any other used method,
- for future research aimed at determination of the influence of input and process quantities on the output quantities (quality, micro-purity) of steel or products made of steel it would be appropriate to elaborate so called *plan of targeted experiment* in conformity with the ***DOE*** method (Design Of Experiments) - see [10], which will ensure an orthogonality of partial experiments with impact on reduction of scatters (and therefore increase of statistical significance) of estimates of regression coefficients and increase of the model determination coefficient,
- it is important to compare the results of the used diagnostic methods (Spark-DAT, ASTM, US findings) and complete them by results of ***metallographic analysis*** and ***plastometric*** tests of samples [11]. It would also be appropriate to consider also use of other analytical methods for evaluation of the influence of *targeted overheating*, which provide more complex information about overall degree of contamination of steel, such as e.g. *a real portion of inclusions and their number*, their *distribution into distribution groups*, *micro-analytic* classification into *ternary diagrams,* etc.

REFERENCES

[1] K. Kupka, PLS regresní modely v řízení kvality [PLS regression models at quality control]. In Proceedings from the national seminar Data analysis 2005/II for technology, research and other applications, pp. 125-127, Lázně Bohdaneč (2005).

[2] M. Meloun, J. Militký, Statistické zpracování experimentálních dat [Statistic processing of experimental data]. PLUS, 839 Praha (1994).

[3] J. Anděl, Statistická analýza časových řad [Statistic analysis of time series]. SNTL, 282 Praha (1975).

[4] T. Cipra, Analýza časových řad [Analysis of time series]. SNTL/ALFA, 248 Praha (1986).

[5] J. Arlt, Moderní metody modelování ekonomických časových řad.[Advanced methods of modelling of time series] Grada Publishing, 312 Praha (1999).

[6] M. Meloun, J. Militký, Kompendium statistického zpracování dat. Metody a řešené úlohy včetně CD [Compendium of statistic data processing. Methods and solved tasks – with CD]. Academia, 764 Praha (2002).

[7] M. Meloun, J. Militký, M. Hill, Počítačová analýza vícerozměrných dat v příkladech [Computer analysis of multi-dimensional data with examples]. Academia, 456 Praha (2005).

[8] M. Meloun, Výhody a přednosti vícerozměrné analýzy dat [Advantages of multi-dimensional data analysis]. In Proceedings from the national seminar *Data analysis 2002/II* for technical and research practice, 7-28 Lázně Bohdaneč (2002).

[9] P. Hebák, Vícerozměrné statistické metody [Mutli-dimensional statistic methods]. In Proceedings

from the national seminar *Data analysis 2005/II* for technology, research and other applications, 190-198 Lázně Bohdaneč (2005).

[10] J. Tošenovský, D. Noskievičová, Statistické metody pro zlepšování jakosti [Statistic methods for quality enhancement]. Montanex, 362 Ostrava (2000).

[11] J. Cibulka, J. Morávka, K. Michalek, B. Chmiel, Hodnocení teplotně-časových parametrů výroby oceli pro zvyšování výsledné jakosti pomocí vícerozměrných statistických metod [Evaluation of temperature-time parameters of steelmaking for enhancement of resulting quality with use of multi-dimensional statistic methods]. Technical report. Ostrava : VŠB-TU Ostrava, 51 May 2007.

Received: 3 April 2008.