Biocybernetics and Biomedical Engineering
1995 Volume 15 Number 3-4 pp. 35–51

JERZY LIEBHART*, ANNA BARTKOWIAK

A Distance-Based Regression Method as Applied in Spirometry for Estimation of Residual Volume

We applied the distance-based regression (DBR) for estimation of the residual volume (*RV*). Great advantages of DBR are that continuous and discrete predictors can be simultaneously used. We found that after including additional categorical variables an essential improvement of the goodness of fit was attained.

Till now the Residual Volume (*RV*) was estimated by a regression equation established conditionally on the type of the ventilatory defect stated in the patients.

In the paper we present one uniform algorithm for estimation of the *RV* for patients suffering from 3 types of ventilatory disorders (obturative, restrictive or mixed one) as well as for patients representing "the norm", i.e. not exhibiting symptoms of the ventilatory function loss.

The validity of the algorithm was confirmed by a high overall multiple correlation coefficient, and also by analysis of relative errors. The algorithm was additionally cross-validated for patients with obturative disorders, the most frequent pulmonary disease.

K e y w o r d s: residual volume estimation, spirometry, predictions, relative error.

1. Introduction

Residual volume (*RV*) and total lung capacity (*TLC*) are the important spirometric parameters that cannot be measured directly. The problem of estimating them from other spirometric measurements, which are more easy to obtain, was addressed formerly a.o. by Liebhart et al. [1, 2] and Krusińska et al. [3]. This was done using mainly the LSE (Least Square Error) approach with some refinements considering regression diagnostics and robust estimation of the relevant regression equation. It has appeared that while the authors managed to construct models sufficiently well estimating the *TLC* [3, 4, 1], corresponding results for *RV* were unsatisfactory [1]. This inclined us to search for another method more suitable to the problem under consideration.

The aim of this paper is the evaluation of the distance based regression method (DBR) proposed by Cuadras and Arenas [5] when applied to the problem of finding a unified and more concise formula for the estimation of

* The second author's work was partially supported by the KBN grant no. 650/2/91.

Residual Volume. The distance based regression allows for a substantially new approach that permits to include categorical variables into the set of predictors. This opens entirely new possibilities not available when using the traditional regression methods.

2. The Data

We utilize in principle the same data as those analysed by Liebhart et al. [6] or Bartkowiak and Liebhart [7]. One group of the data (Obturation) was also examined by methods of exploratory data analysis by Bartkowiak [8]. The data contains recorded values of 14 variables (Residual Volume, Total Lung Capacity and twelve other features, from which we will consider in the following only six, described below) observed in patients classified as belonging to 3 groups of pulmonary diseases and one control group representing the norm.

The group size of our data were as follows:

Group 1. Obturative type of ventilatory defect, $n=128$ patients.

Group 2. Restrictive type of ventilatory defect, $n=21$ patients.

Group 3. Mixed type of ventilatory defect, $n=28$ patients.

Group 4. Norm, i.e. without ventilatory defect, $n=28$ patients.

One can see that the groups are quite diversed in size. In particular one can notice the relatively large size of Group 1, which reflects the frequency of occurrence of the obturative type of ventilatory defect in the population. Because the distance based methods applied in the present analysis can deal with mixed-type variables we enlarged the data by four "new" variables being of categorical type.

The continuous variables are:

$Y1$: RV, Residual Volume [cm^3],

$Y2$: TLC, Total Lung Capacity [cm^3],

$X1$: Age [years],

$X2$: Height [cm],

$X3$: VC, Vital Capacity [cm^3],

$X4$: $VC\%$ evaluated as $VC/VC_{\text{predicted}} \times 100$ [%].

Spirometric parameters were measured using Spirograph PT-400, the RV and TLC values were obtained by helium method.

For the present analysis we have enlarged this set of data by inlcuding additionally four categorical variables:

$Z1$: Group membership number of the patient corresponding to the diagnosis of the ventilatory defect type — as defined below.

$Z2$: Measuring ventilatory pattern — as defined below.

$Z3$: Measuring some spirometric features — as defined below.

$Z4$: Degree of emphysema and lungs hyperinflation — as defined below.

All these categorical variables can take the values 1, 2, 3 or 4.

The "Group membership" variable $Z1$ obtained the values 1, 2, 3 or 4 depending on the type of ventilatory defect (see e.g. [9]):

1. Obturation,
2. Restriction,
3. Mixed type,
4. Norm.

The "ventilatory pattern" variable $Z2$ obtained the values 1, 2, 3 or 4 depending on the values of $X4$ and $X6$:

if $X4 \geqslant 75$ and $X6 < 75$ then $Z2:=1$,

if $X4 < 75$ and $X6 \geqslant 75$ then $Z2:=2$,

if $X4 < 75$ and $X6 < 75$ then $Z2:=3$,

if $X4 \geqslant 75$ and $X6 \geqslant 75$ then $Z2:=4$.

The "spirometric index" variable $Z3$ obtained the values 1, 2, 3 or 4 depending on the difference $\delta = X6 - X11$:

if $\delta \leqslant 0$ then $Z3:=1$,

if $0 < \delta \leqslant 10$ then $Z3:=1$,

if $10 < \delta \leqslant 20$ then $Z3:=3$,

if $20 < \delta$ then $Z3:=4$.

The "hyperinflation or emphysema" variable $Z4$ obtained the values 1, 2, 3 or 4 depending on the ratio $\rho = RV/TLC$:

if $\rho \leqslant 0.25$ then $Z4:=1$,

if $0.25 < \rho \leqslant 0.40$ then $Z4:=2$,

if $0.40 < \rho \leqslant 0.55$ then $Z4:=3$,

if $0.55 < \rho$ then $Z4:=4$.

The variable $Z1$ was obtained as a result of the complex examination of a patient. The diagnosis of ventilatory defect type was done on the basis of anamnesis, physical examination and laboratory findings including spirometry, blood gases analysis, chest X-ray, and — in some cases — bronchoscopy.

The variable $Z2$ was established by use of the traditional Miller's algorithm [10] that allows for differentiation between "norm" and different types of ventilatory defects using VC and FEV_1 measurements.

The variable $Z3$ is thought as a kind of an index that, in the author's opinion, should be additionally helpful in discrimination patients with the obturative type of ventilatory function loss from those with the mixed one.

The variable $Z4$ need more detailed description. It was introduced in order to evaluate the presence or absence, and the degree of emphysema and lungs hyperinflation.

The values of this variable can be assigned on the basis of medical examination; the ratio $\rho = RV/TLC$ used here is a rough approximation of this variable.

We proceeded with our analysis as follows: Firstly we performed the classical LSE analysis using the continuous variables. Secondly we carried out the distance-based (DB) regression analysis with both the continuous and categorical variables.

We considered the following combinations of variables and methods referred hereafter as variants of the evaluations:

A: $X1 \div X4$; computing the LSE regression.

B: $X1 \div X4$; computing the DB regression.

C: $X1 \div X4$ and $Z4$; computing the DB regression.

D: $X1 \div X4$ and $Z2$, $Z3$; computing the DB regression.

E: $X1 \div X4$ and $Z2 \div Z4$; computing the DB regression.

In Table 1 we show the means (AV) and standard deviations (SD) of the two predicted variables $Y1 = RV$, $Y2 = TLC$, and of the four continuous predictors $X1$, $X2$, $X3$, $X4$.

Table 1. Means (AV) and standard deviations (SD) in the original full data set and in the five Balanced Training Groups

	Original data							
	1. Obstruction $n=128$		2. Restriction $n=21$		3. Mixed type $n=28$		4. Norm $n=28$	
	AV	SD	AV	SD	AV	SD	AV	SD
RV	2277	663	1516	417	2213	646	1566	483
TLC	5530	997	4027	929	4924	937	5415	1142
$X1$	48	11	53	10	54	9	41	11
$X2$	165	8	165	9	169	6	164	8
$X3$	3261	894	2416	771	2718	687	3851	841
$X4$	77	15	56	12	59	13	93	12

	Balanced Training Groups									
	BTG 1 $n=105$		BTG 2 $n=105$		BTG 3 $n=105$		BTG 4 $n=105$		BTG 5 $n=105$	
	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD
RV	1913	616	1970	676	1877	632	1880	632	1910	652
TLC	4981	1086	5029	1205	5026	1145	5041	1172	5030	1127
$X1$	49	12	49	11	48	11	49	12	49	11
$X2$	165	8	166	8	166	8	166	8	166	8
$X3$	3052	905	3043	1020	3134	959	3147	1009	3100	967
$X4$	72	19	71	20	74	20	73	20	72	20

Some former investigations (Liebhart et al. [3, 4]) have shown that the regression equation aiming at the estimation of *TLC* and *RV* was markedly dependent on the type of ventilatory defect. This means that the appropriate one out of the four constructed equations could be chosen only after previous ventilatory function diagnosing.

Our aim was to construct a general prediction algorithm permitting to estimate *RV* in all patients, independently of their type of ventilatory defect.

We decided to use a balanced training sample as the basis for construction of the overall algorithm (i.e. suitable for all groups). Such a sample should contain patients with all types of ventilatory defects (including the norm) represented by groups of nearly equal size — otherwise a very large representation of one type of ventilatory disorders (as the large obturation group in our data) would influence the algorithm to be in the first place suitable for the largest group.

To attain our goal — of obtaining a balanced training group — we sampled at random (sampling without replacement) from the large group 1 (Obturation) $nn=28$ different patients and added them to the whole groups 2, 3 and 4 being nearly of the same size. In that way we got the first balanced training group, denoted in the following as BTG 1, containing $n=28+28+21+28=105$ patients. Next we sampled from the patients belonging to the big group 1, and which were not chosen to enter BTG 1, another $nn=28$ different patients to form an additional test group (TEST 1) to be used later for an external check cross-validation of the validity of the algorithm constructed on the basis of the data contained in BTG 1.

In a similar way we constructed another four balanced training groups named BTG 2, BTG 3, BTG 4 and BTG 5; to each of them corresponding test groups TEST 2, TEST 3, TEST 4 and TEST 5 were assigned.

The means (*AV*) and standard deviations (*SD*) for the five balanced training groups are shown in the lower part of Table 1.

All the evaluations described hereafter were carried out in parallel on the five balanced training groups and checked externally on the correspondent test groups.

3. Methods used

3.1. Computing Distances Between Individuals

The distance-based approach has a long tradition in exploratory data analysis (see, e.g. [11, 12]); nonetheless, in the context of regression, it was firstly proposed by Cuadras and Arenas [5].

The distance-based regression starts from the distance matrix **D** evaluated from the predictor variables. The distances can be defined for sets of predictors containing continuous, binary and categorical variables as well. In the paper we will use for this purpose the Gower Distance (see, e.g. [5, 11, 13, 14]) which is defined as follows.

Suppose, we consider p continuous and q categorical variables. Let $\mathbf{x}_i=(x_{i1},\ldots,x_{ip})$, $x_j=(\mathbf{x}_{j1},\ldots,x_{jp})$ and $\mathbf{z}_i=(z_{j1},\ldots,z_{iq})$, $\mathbf{z}_j=(z_{j1},\ldots,z_{jq})$ denote the vectors of the continuous and categorical variables observed for the i-th and j-th individual. Let $\mathbf{G}=(G_1,\ldots,G_p)$ be the vector of ranges of the continuous variables, i.e.

$$G_k=\max_l(x_{lk})-\min_l(x_{lk}), \quad l\in[1,\ldots,n),\ k\in[1,\ldots,p].$$

The similarity $s(i,j)$ between the i-th and the j-th individual is evaluated as

$$s(i,j)=\left[\left(\sum_{k=1}^{p}(1-|x_{ik}-x_{jk}|/G_k)\right)+M_{ij}\right]/(p+q),$$

where M_{ij} denotes the number of matches (identities in the categories), i.e. number of occurrences $z_{ih}=z_{jh}$ counted over the q categorical variables $(h=1,\ldots,q)$ when considering the vectors $\mathbf{z}_i$ and $\mathbf{z}_j$.

Then the squared Gower distance $d_G^2(i,j)$ between the i-th and j-th individual is evaluated as [14, 11]

$$d_G^2(i,j)=1-s(i,j). \tag{1}$$

Alternatively, the Euclidean distance could be computed; however, this can be done using the continuous variables only. The squared Euclidean distance $d_E^2(i,j)$ between the i-th and j-th individual is computed as:

$$d_E^2(i,j)=\sum_{k=1}^{p}(x_{ik}-x_{jk})^2. \tag{2}$$

Usually it is more reasonable to compute the Euclidean distance using standardized values of the considered variables:

$$d_E^2(i,j)=\sum_{k=1}^{p}\frac{(x_{ik}-x_{jk})^2}{s_k^2},$$

where s_k^2 denotes the sample variance of the k-th variable.

It is known [5] that the results (i.e. multiple correlation coefficients and residuals when computed from a regression with p variables) obtained by the DBR method using Euclidean distances are the same as those obtained by the classical LSE regression.

3.2. Computing the Distance-Based Regression

Suppose, we have already established the distance matrix **D** between the individuals contained in the training sample. The computed distance matrix is formally of size $n \times n$, nevertheless, since the distances are symmetric, i.e. $d(i,j)=d(j,i)$, we need only one halve, it is the lower (or upper) triangle of this matrix. From this distance matrix we evaluate by methods of Multidimensional Scaling [5, 12] the corresponding principal coordinates. Generally we obtain l principal coordinates, with $l=rank(\mathbf{B})$, where **B** denotes the inner product matrix derived from the distance matrix **D**. Next we perform an ordinary regression analysis taking as explanatory variables (predictors), those principal coordinates that are mostly correlated with the predicted variable Y.

The number of principal coordinates included in the regression model depends on our decision. Statistical significancy tests for models with mixed type of explanatory variables are not known. Including a higher number of principal coordinates we obtain a better approximation of Y in the given data set; however, including too much of them we cause overfitting — and making predictions in another data set we might get a worse prediction then.

Let $\Psi_1,\ldots,\Psi_m$ ($\Psi_k=(\psi_{1k},\ldots,\psi_{nk})^T$, $k=1,\ldots,m$) denote the principal coordinates retained for further analysis. Then the considered regression model is described by the equation:

$$Y=\beta_0+\beta_1\Psi_1+\ldots+\beta_m\Psi_m+e. \tag{3}$$

The regression model is then elaborated by well known regression methods, e.g. using the LSE method [15].

3.3. Validation of the Distance-Based Regression

We will consider three statistics permitting to verify the goodness of fit of the established regression:

(a) The squared multiple correlation coefficient, called also the coefficient of determination (denoted by R^2).

(b) Residuals computed for the same (own) and foreign data sets.

(c) Relative errors computed for the same (own) and foreign data sets.

Below we describe briefly these statistics.

(a) The Squared Multiple Correlation Coefficient

The multiple R^2 gives a measure of fit of the constructed regression in the set of data from which the regression equation was evaluated. R^2 shows how large part of the total (corrected for the mean) sum of squares of the predicted

variable Y is reduced when accounting for the established regression (see e.g. Sen & Srivastava [15], p. 39).

Obviously $0 \leqslant R^2 \leqslant 1$. High values of R^2 indicate a strong linear relationship between the predicted variable Y and the predictors.

(b) Residuals

When speaking about residuals we have to distinguish between two situations:

(i) The prediction is made for an individual no. i belonging to the training sample, i.e. belonging to the *own* data set.

(ii) The prediction is made for an individual no. j belonging to the test sample, i.e. to a *foreign* data set.

In the first situation the residuals e_i $(i=1,...,n)$ are computed as the difference between the observed and predicted values of the variable Y:

$$e_i = y_i^{(\text{obs})} - y_i^{(\text{pred})}, \tag{4}$$

with the prediction done by use of equation (3). The estimates $\hat{\beta}_0,...,\hat{\beta}_m$ needed to carry out these evaluations have been obtained from the same data set to which the individual no. i is belonging.

In the second situation the residuals are also computed as the difference between the observed $(y_j^{(\text{obs})})$ and predicted $(y_j^{(\text{pred})})$ values of the variable Y; however, the prediction is now done on the basis of principal coordinates obtained from a different data set. Speaking more exactly, the individual no. j has now number j in the test sample, while the prediction algorithm was evaluated on the basis of a totally different training sample.

The algorithm for predicting by the DBR method the value of Y from a new data vector $(\mathbf{x}_j, \mathbf{z}_j)$ observed in a foreign data set is more complicated; the proper algorithm is shown in the paper by Cuadras and Arenas [5].

Generally it is expected, especially when one uses a high number of principal coordinates, that in the first situation the fit of the regression would be quite good. This is not expected to be true in the second situation: the fit of the regression, when measured by the magnitude of the residuals, is usually much worse.

In both situations, before evaluating the residuals, we should fix the regression equation (3) needed to evaluate the predicted values of Y. Among others, we should fix m, the number of principal coordinates included into the regression equation.

Let us also notice, that the applied DB regression in its final pace is in fact carried out as the ordinary LSE regression, only computed on some specifically constructed variables called principal coordinates. As such, it gives the guarantee that the mean value of the residuals computed from the same data set,

from which the parameters of the regression equation were estimated, equals to zero for any number of retained principal coordinates. Moreover, the residual sum of squares of the residuals decreases (strictly: does not increase) when considering increasing numbers (m) of retained principal coordinates.

Quite different properties are found when considering residuals, evaluated in a different set of data as that which has served as the basis for constructing the prediction algorithm. Now the mean of the residuals does not need to be equal to zero. The residual sum of squares of the residuals does not need to decrease with increasing number of the retained principal coordinates. As a matter of fact, the function expressing the relation of the variance (of the residuals) from m (number of the retained principal coordinates) should first decrease (adding relevant factors to the regression should diminish the variance), and then increase (when adding superfluous factors reflecting white noise).

(c) Relative Errors

Apart from characterizing the residuals by their means and standard deviations, we consider also their relative errors defined as follows:

$$\text{relative error} = \frac{|y_i^{(\text{obs})} - y_i^{(\text{pred})}|}{y_i^{(\text{obs})}} \times 100\,. \tag{5}$$

The relative error shows — for the given individual no. i — how large part of the observed value of Y observed for this individual, is correctly predicted by the applied regression algorithm.

We do not use for validation of the constructed model any leaving-one-out method, because the second author believes strongly that this method yields the relevant information on stability of the established regression only in the case of one influential individual. When there are more of them, then the leaving-one-out method may give very unprecise results.

4. Results of Evaluations

The evaluations were conducted in five variants dealing with different combinations of the considered variables. The variants are denoted A, B, C, D, E (see Section 2).

In variant A and variant B only the variables $X1 - X4$ were considered as predictors. In variant A we computed the DB regression using Euclidean distances (hence equivalent to the classical LSE regression); while in variant B we computed analogous regression with $X1 - X4$ using the Gower distance.

The other three variants were using Gower distances with mixed variables considering $X1-X4$ and some of the variables $Z2-Z4$.

For each variant the regression predicting RV on the basis of the appropriate predictors was computed.

(a) Considering the Multiple Correlation Coefficient

The squared multiple correlation coefficients R^2 obtained when retaining $m=1$, 2, 4, 8 and 12 principal coordinates are shown in Table 2. The respective values are recorded for the five variants A, B, C, D, E of the evaluations.

For the DB regression based on Euclidean distances and using data with $p=4$ predictors only $p=4$ principal coordinates can be computed. The corresponding squared multiple correlation coefficients range 0.20–0.30, which indicates that the linear dependence between RV and the variables $X1$, $X2$, $X3$, $X4$ is poor.

Table 2. Squared multiple correlation coefficients observed in five balanced training groups when taking m principal coordinates for predicting the Residual Volume

Variants							$m=$	1	2	4	8	12	max (%)
A: LSE with $X1-X4$							BTG 1						
B: DBR with $X1-X4$							A	0.18	0.19	0.20	–	–	4(20)
C: DBR with $X1-X4$, $Z4$							B	0.17	0.23	0.35	0.46	0.57	56(95)
D: DBR with $X1-X4$, $Z2$, $Z3$							C	0.36	0.53	0.66	0.75	0.80	41(95)
E: DBR with $X1-X4$, $Z2-Z4$							D	0.10	0.18	0.30	0.44	0.57	56(95)
							E	0.23	0.36	0.52	0.72	0.82	36(95)
$m=$	1	2	4	8	12	max (%)	$m=$	1	2	4	8	12	max (%)
BTG 2							BTG 3						
A	0.23	0.24	0.24	–	–	4(24)	A	0.25	0.26	0.27	–	–	4(20)
B	0.19	0.26	0.35	0.48	0.57	54(95)	B	0.23	0.30	0.40	0.54	0.62	52(95)
C	0.25	0.48	0.71	0.78	0.83	37(95)	C	0.26	0.43	0.64	0.75	0.81	35(95)
D	0.13	0.17	0.25	0.39	0.49	56(95)	D	0.08	0.17	0.31	0.47	0.58	53(95)
E	0.22	0.35	0.57	0.74	0.79	38(95)	E	0.20	0.32	0.52	0.68	0.78	37(95)
$m=$	1	2	4	8	12	max (%)	$m=$	1	2	4	8	12	max (%)
BTG 4							BTG 5						
A	0.28	0.29	0.30	–	–	4(30)	A	0.25	0.28	0.28	–	–	4(28)
B	0.27	0.31	0.39	0.51	0.59	52(95)	B	0.20	0.25	0.33	0.46	0.57	52(95)
C	0.21	0.42	0.66	0.75	0.79	42(95)	C	0.34	0.56	0.72	0.79	0.82	38(95)
D	0.16	0.22	0.32	0.47	0.57	54(95)	D	0.10	0.19	0.31	0.48	0.58	56(95)
E	0.28	0.41	0.57	0.70	0.77	40(95)	E	0.34	0.49	0.65	0.75	0.80	42(95)

For the DB regression based on Gower distances it is theoretically possible to compute $n-1$ principal coordinates, with n being the size of the training

sample. In our evaluations we started from one principal coordinate and continued to add more and more of them, till *either* a 95% reduction of the total variance was achieved, *or* the number of extracted principal coordinates equaled to $n-1$. Then we stopped the process. In Table 2 under the heading MAX (%) *either* the maximal number of principal coordinates possible to obtain — *or* that number of principal coordinates which gave a 95% reduction of the total variance are shown. Using the sets B we got extracted 52 – 57 principal coordinates satisfying the above specified conditions (let us remember that the sample sizes were 105).

Comparing the results from B with A one can state that for every BTG the values of R^2 obtained when using Gower distance are higher as those obtained from an analogous Euclidean distance. The differences in R^2 range 0.05 – 0.15, which means that the DB regression with Gower distances can decidedly improve the quality of description of the relations between the predicted RV and the predictors appearing in the considered data set.

Adding the categorical variables $Z4$ or $Z2$, $Z3$, $Z4$ the quality of prediction improves even more: we obtain then multiple correlation coefficients ranging 0.52 – 0.72 for $m=4$ principal coordinates. This is really a great improvement.

The above conclusion is valid in the context when the considered goodness of fit measures is evaluated for the same data on which the regression equation was constructed.

(b) Considering Residuals from Own and Foreign Data Sets

Obviously the own residuals will decrease with increasing number of principal coordinates included into the regression equation (3). Therefore the investigation of these residuals is not so much interesting for us.

A much more interesting task yielding non obvious results is to investigate some foreign residuals. In our analysis we had 5 test samples corresponding to 5 BTG group. We decided to proceed in such a way, that after evaluating the predicting regression, say, in the first BTG group we used the established regression for evaluating residuals in the corresponding first test group. In turn we did it for the pairs of groups (BTG 2 – Test 2), (BTG 3 – Test 3), (BTG 4 – Test 4), (BTG 5 – Test 5).

The distributions of the residuals evaluated in the first two test samples are shown in Figs. 1 and 2 in the form of box-and-whisker plots.

The plots for the remaining test groups look very similar.

From these plots one can see that the most favourable variant for estimation of RV are variants C and E. These variants yield residuals with the smallest spread.

However, let us notice, that all the means are above 0, hence the predicted values are somehow underestimated.

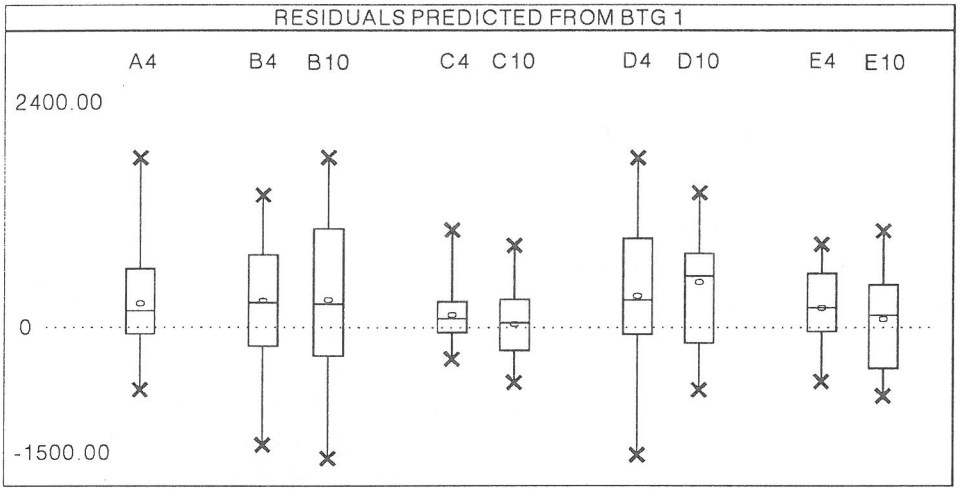

Fig. 1. Box-and-whisker plots of residuals evaluated in Test group 1 for various subsets of predictors. Letter symbols denote subsets of variables (see Table 2 or Section 2), digits denote numbers of principal coordinates (obtained from BTG 1) used for prediction

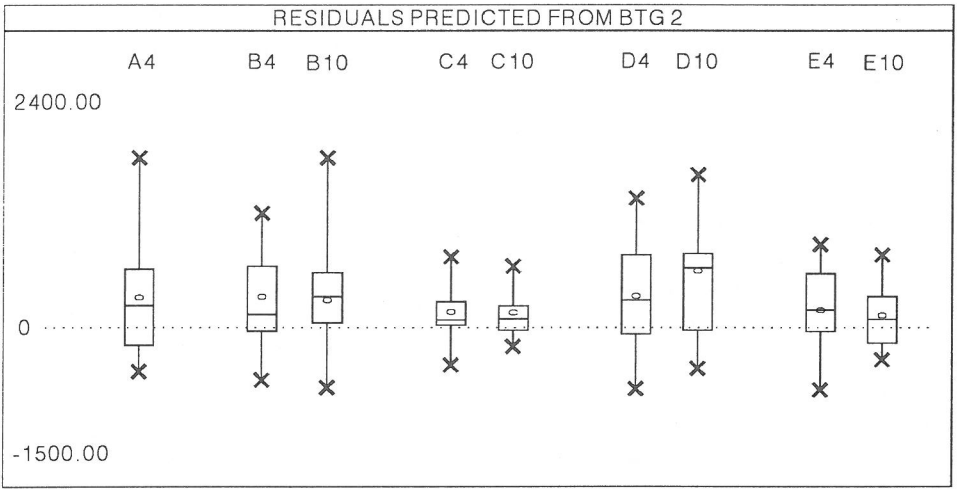

Fig. 2. Box-and-whisker plots of residuals evaluated in Test group 1 for various subsets of predictors. Letter symbols denote subsets of variables (see Table 2 or Section 2), digits denote numbers of principal coordinates (obtained from BTG 2) used for prediction

(c) Considering Relative Errors in Own and Foreign Data Sets

We have investigated the relative errors in all the 5 test groups. In Table 3 we show the means of these errors — in dependence of m, the number of principal coordinates introduced into the regression equation.

Table 3. Means (AV) and standard deviations (SD) of relative errors computed in own and foreign data sets when retaining $m=2, 4, 6, 8, 10$ and 12 principal coordinates. Key to variants A, B, C, D, E — see Table 2

↓ Groups		$m=2$		$m=4$		$m=6$		$m=8$		$m=10$		$m=12$	
Variants ↓		AV	SD	AV	SD	AV	SD	AV	SD	AV	SD	AV	SD
Own data set													
BTG 1	A	27	(27)	27	(29)	–		–		–		–	
	B	26	(28)	25	(26)	24	(24)	22	(23)	21	(22)	20	(22)
	C	21	(22)	18	(20)	16	(15)	15	(13)	14	(13)	13	(12)
	D	27	(25)	24	(25)	23	(22)	22	(20)	21	(18)	20	(18)
	E	24	(24)	21	(21)	17	(20)	16	(14)	14	(11)	13	(11)
BTG 2	A	28	(31)	27	(32)	–		–		–		–	
	B	27	(27)	25	(25)	23	(22)	22	(22)	21	(20)	19	(18)
	C	23	(21)	17	(18)	16	(15)	14	(14)	13	(11)	13	(10)
	D	29	(30)	27	(28)	25	(26)	24	(27)	23	(28)	22	(26)
	E	26	(26)	20	(23)	18	(18)	16	(16)	15	(14)	14	(13)
Foreign data set													
TEST 1	A	21	(16)	21	(15)	–		–		–		–	
	B	23	(24)	28	(23)	29	(23)	30	(20)	29	(21)	28	(23)
	C	12	(9)	11	(8)	12	(8)	12	(8)	14	(10)	13	(9)
	D	26	(21)	27	(25)	26	(20)	28	(17)	27	(15)	26	(14)
	E	16	(09)	15	(10)	14	(11)	14	(10)	18	(13)	17	(12)
TEST 2	A	17	(13)	18	(13)	–		–		–		–	
	B	18	(15)	26	(22)	21	(12)	22	(15)	24	(15)	26	(15)
	C	23	(10)	10	(8)	10	(10)	10	(8)	9	(8)	12	(9)
	D	20	(9)	18	(12)	19	(10)	23	(10)	23	(11)	20	(13)
	E	19	(12)	16	(10)	10	(8)	9	(5)	9	(6)	11	(7)

In Table 3 we show only the results for the first two test groups. Again one can note the steady diminishing of the relative errors depending on the number of principal coordinates used for construction of the regression function. Surprisingly, we got in the test groups smaller relative errors as in the original BTG groups.

Let us emphasize once more that the above conclusion is valid in the context when the considered goodness of fit measure is evaluated on the same data on which the regression equation was constructed. The behaviour of the established regression applied to another set of data is more important. To evaluate the true practical importance of the established regression we should look at the residuals and the relative errors evaluated in test groups.

In Figures 3 and 4 we show graphs exhibiting the relative errors as function of m, the number of PC's used for construction the respective regression. Again we show only the graphs obtained from the first two test groups, since the other graphs exhibit the same pattern.

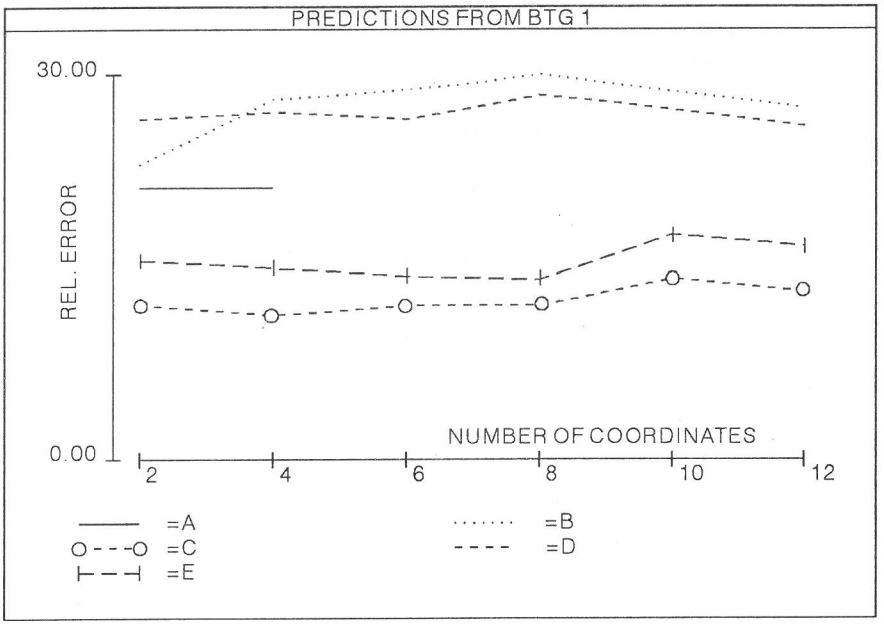

Fig. 3. Test group 1. Relative errors versus m, the number of principal coordinates included into the predicting regression equation constructed from BTG 1

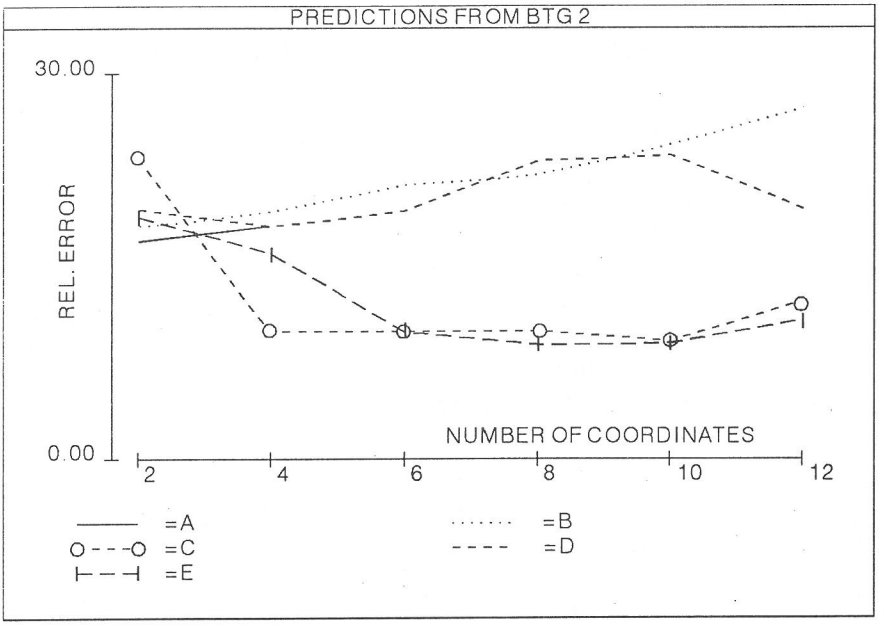

Fig. 4. Test group 2. Relative errors versus m, the number of principal coordinates included into the predicting regression equation constructed from BTG 2

One can see in these graphs two systematic features:

— The magnitude (level) of the error does not depend much on m, the number of principal coordinates.

— The variants C and E yield markedly lower errors as the variants A, B and D.

Summarizing the results of this section, one can say that calculating DB regression with Gower distance we generally improve the goodness of fit of the established regression. Adding to the continuous variables $X1$, $X2$, $X3$, $X4$ the categorical variables $Z2$, $Z3$, $Z4$ and $Z4$ alone we can improve substantially the goodness of fit.

5. Discussion

This study was designed to assess whether the DB regression method can deal with the problem of nonhomogeneous data and nonlinear relations, with the future goal to propose an algorithm of RV estimation ready to be used in daily medical practice. The latter can be achieved only after collecting the data anew as to include the information that appeared to be necessary in the light of present analysis.

DB regression, by use of Gower distances, performed without introducing any categorical variables, did not bring about any substantial improvement in RV prediction expressed by relative errors values when compared to ordinary LSE regression (Tables 2, 3).

We first introduced to the data set analysed by DB regression two categorical variables $Z2$ and $Z3$. The former was constructed on the basis of the Miller's algorithm for ventilatory defect types differentiation, the latter, in our opinion, should be helpful in discrimination between obturative and mixed ventilatory defect types. From Tables 2, 3 and Figures 1 – 5 is evident that the results appear unsatisfactory. This is in agreement with earlier suggestion of Liebhart et al. [2] that Miller's algorithm is too simplified to give sufficiently precise diagnosis of ventilatory function loss.

The detailed analysis of individual patients data showed that the error of estimation was dependent on the presence and the degree of lungs hyperinflation or emphysema.

To cope with this problem we decided to introduce additionally a categorical variable describing this type of disorder. This variable is denoted in our analysis as $Z4$.

The results obtained for the present data with variables $Z2$, $Z3$ and $Z4$ and particularly with variable $Z4$ alone, showed amazing improvement in correlation coefficients between the measured and estimated RV values, in the mean relative error as well as relative error variance. This confirmed our

presumption about the significance of lungs hyperinflation and emphysema in the regression of *RV* from other spirometric parameters. The decrease in relative error of estimation from about 25% to 10–15% can be considered satisfactory, especially when remembering that the helium method of *RV* measurements is burdened with an error ranged about 15% (see, e.g. Cotes [9]). From Figs. 1 and 2 is difficult to assess which is the optimal number of principal coordinates. It seems to lie somewhere between 4 and 10 coordinates.

It is a nice feature of the residuals obtained in our analysis that they are somehow stable within the considered range. Nonetheless, some more rigid rule for choosing the relevant principal coordinates are needed. Some work in this respect is in progress (Bartkowiak [16]).

The categorical variable *Z1*, defined as the disease group category (see Section 2) was not used so far in the constructed regression. It is possible that *Z1* contains some useful information on the shape of the established regression and, when included into the evaluations, would yield a further substantial improvement of the fit. This topic was investigated in another study by Bartkowiak and Liebhart [17]. They stated that Z1 plays a considerable role in predicition of RV – however only then if it is combined with the variable *Z4*.

References

1. Liebhart J., Krusińska E., Karkowski Z., Małolepszy J., Liebhart E.: On discrepancies between indirect estimation of residual volume (RV) and total lung capacity (TLC). Modelling Simulation & Control, 1988, 13, 35–42.
2. Liebhart J., Krusińska E.: A new method for differentiation of different types of ventilatory defects (in Polish). Pneumonol. Pol. 1985, 53, 527–535.
3. Krusińska E., Liebhart J., Bartkowiak A., Karkowski Z., Małolepszy J., Liebhart E., Marciniak W.: Various methods of variable selection in linear regression. Presentation of the results of Total Lung Capacity estimation. (In Polish), Człowiek Populacja Środowisko 1988, V, 28, 69–102.
4. Liebhart J., Karkowski Z., Krusińska E., Liebhart E., Małolepszy J.: Determination of Total Lung Capacity using data from forced expiratory flow volume curves and simple measurements of the chest. (In Polish). Pneumonol. Pol. 1985, LIII, 11–12, 520–526.
5. Caudras C.M., Arenas C.: A distance based regressions model for predictions with mixed data. Commun. Statist. Theory Meth. 1990, 19 (6), 2261–2279.
6. Liebhart J., Bartkowiak A., Liebhart E.: The impact of outliers in the regression estimating TLC from age and some spirometric observations. Modelling Simulation & Control, 1989, C, AMSE Press, 15 (4), 1–18.
7. Bartkowiak A., Liebhart J.: Regression diagnostic for the equations estimating *TLC* from age and some spirometric observations. Modelling Simulation & Control 1990. C, AMSE Press, 21, 2, 1–14.
8. Bartkowiak A.: Exploratory data analysis, its historical development, what it is today. Biocybernetics & Biomedical Engineering, 1995, 15 N61–2, 85–120.
9. Cotes J.E.: Lung Funtion. Assessment and Application in Medicine. Blackwell Scientific Publications. Oxford, 1965. Polish edition: PZWL, Warszawa 1969.

10. Miller W., Wu N., Johnson R.: Convenient method of evaluating pulmonary function with a single breath. Anaesthesiology 1956, 17, 480.
11. Cuadras C.M.: Distancias Estadisticas. Estadistica Espanola 1988, 30, 119, 295–378.
12. Mardia K.V., Kent J.T., Bibby J.M.: Multivariate Analysis, Academic Press, 1979.
13. Cuadras C.M., Fortiana J.: Aplicación de las distancias en Estadistica. (In Spanish with broad English Summary). Qüestiió 1993, 17 (1), 39–74.
14. Gower J.C.: A general coefficient of similarity and some of its properties. Biometrics, 27, 1971, 857–874.
15. Sen A., Srivastava M.: Regression Analysis, Theory, Methods and Applications. Springer, New York, Berlin 1990.
16. Bartkowiak A.: Practice of computing distance-based regression — How many PC's are relevant? Joint Statistical Meetings, Toronto August 13–18, 1994. Proceedings of the Statistical Computing Section. Submitted.
17. Bartkowiak A., Liebhart J.: Estimation of the Spirometric Residual Volume (RN) by a Regression Built from Gower Distances. Biom. Journal 1995, 32, no 2, 131–149.

*Medical Academy
Department of Internal Diseases and Allergology
Wrocław
**University of Wrocław
Institute of Computer Science

Regresja metodą „distance" zastosowana w spirometrii do wyznaczania objętości zalegającej (*RV*)

Zastosowano regresję metodą „distance" (DBR) do estymowania trudno mierzalnego wskaźnika spirometrycznego — objętości zalegającej (*RV*). Wielką zaletą metody DBR jest to, że umożliwia ona jednoczesne wykorzystanie zarówno ilościowych ciągłych, jak i dyskretnych zmiennych objaśniających. Po włączeniu do zbioru regresji zmiennych skategoryzowanych stwierdziliśmy znaczącą poprawę jakości estymacji.

Dotychczas *RV* próbowano estymować przez zastosowanie równań regresji swoistych dla każdego z 3 typów upośledzenia wentylacji płuc oraz dla normy (równanie ogólne okazało się bardzo niedokładne). W pracy przedstawiamy uniwersalny algorytm wykorzystujący regresję metodą „distance", pozwalający na estymowanie *RV* u chorych z obturacyjnym, restrykcyjnym i mieszanym typem upośledzenia wentylacji płuc oraz u osób bez niewydolnosci wentylacyjnej. Jakość estymacji potwierdzają wartości współczynników korelacji wielokrotnej (Table 2) oraz wyniki analizy błędów względnych (Table 3). Dodatkowo dla chorych z najczęstszym, tj. obturacyjnym typem upośledzenia wentylacji płuc omawiany algorytm poddano ocenie przez „cross validation".