Prediction of strontium bromide laser efficiency using cluster and decision tree analysis

Iliycho Iliev[1,*], *Snezhana* Gocheva-Ilieva[2], and *Chavdar* Kulin[2]

[1]Technical University of Sofia – branch Plovdiv, Department of Physics, 25 Tzanko Dujsstabanov str., 4000 Plovdiv, Bulgaria
[2]University of Plovdiv Paisii Hilendarski, Faculty of Mathematics and Informatics, 24 Tsar Asen str., 4000, Plovdiv, Bulgaria

Abstract. Subject of investigation is a new high-powered strontium bromide ($SrBr_2$) vapor laser emitting in multiline region of wavelengths. The laser is an alternative to the atom strontium lasers and electron free lasers, especially at the line 6.45 μm which line is used in surgery for medical processing of biological tissues and bones with minimal damage. In this paper the experimental data from measurements of operational and output characteristics of the laser are statistically processed by means of cluster analysis and tree-based regression techniques. The aim is to extract the more important relationships and dependences from the available data which influence the increase of the overall laser efficiency. There are constructed and analyzed a set of cluster models. It is shown by using different cluster methods that the seven investigated operational characteristics (laser tube diameter, length, supplied electrical power, and others) and laser efficiency are combined in 2 clusters. By the built regression tree models using Classification and Regression Trees (CART) technique there are obtained dependences to predict the values of efficiency, and especially the maximum efficiency with over 95% accuracy.

1 Introduction

A key element in the development of laser systems is the improvement of their output characteristics – average output power, laser efficiency, etc. With the accumulation of experiment data from the operation of the lasers, there is an opportunity to gather significant new information on the relationships and dependencies between the measured operational parameters for the specific device.

The subject of investigation in this study is the new high-powered strontium bromide ($SrBr_2$) vapor laser emitting in a longitudinal pulse helium discharge [1]. The laser is characterized by a multiline radiation at several wavelengths in the range from 2.06 μm to 6.45 μm. It shows the greatest promise for practical medical applications such as the processing of biological tissues and bones with minimal damage at a wavelength of 6.45 μm [2].

In order to extract relationships from the available experiment data, in this investigation, two statistical data mining techniques are applied – exploratory cluster analysis and classification and regression trees (CART, [3, 4]). With their help, seven operational parameters are used as input independent variables - laser tube length and diameter, supplied electric power, pulse frequency, and helium buffer gas pressure, as well as the overall output laser efficiency. Through cluster analysis, on linear type classification, it is found that the parameters can be grouped in two clusters and that for efficiency the geometric parameters are the determining factors. Using the CART technique, models are obtained where the experiments are classified according to the measured laser efficiency values. Stable CART models are built, the relative importance of each operating characteristic for laser efficiency is determined. The models are validated through machine learning approach. The values for the efficiency, predicted by the models describe those measured including the maximum values with an accuracy of over 95-96%. The results of the statistical analysis can be applied to guide experimental investigations for the increase of the efficiency of the considered type of $SrBr_2$ lasers. The approach is also fully applicable for other types of laser systems.

All statistical analyses are performed by using IBM SPSS statistical software [5].

2 Description of strontium bromide laser and experimental data

Until recently, the only known source of laser generation with a wavelength of 6.45 μm were free electron lasers. But their use is difficult because of the high price and service costs, which is not acceptable for regular medical practice. A new alternative to the free electron lasers are strontium vapour lasers (atom and ion transfers) with generation at several wavelengths in the infrared spectrum: 2.06, 2.20, 2.69, 2.92, 3.01, 3.07, and 6.45 μm [6, 7]. These lasers have acceptable price, compact size, easy operation, and lower service costs than the free electron lasers. In this study we investigate the new

* Corresponding author: iliev55@abv.bg

$SrBr_2$ lasers, developed and patented at the Institute of Solid State Physics "Academician Georgi Nadjakov" of the Bulgarian Academy of Sciences [1].

The current stage of the development of $SrBr_2$ laser and its physical characteristics are presented in papers [8, 9], where a laser based on $SrBr_2$ with laser output power of 2.4 W is reported. In [1, 10] the laser output power reaches 4.27 W, with 90% of the laser generation at the 6.45 μm line. The obtained values of laser output power are fully comparable with these for metal strontium lasers.

The laser tube of $SrBr_2$ laser is entirely designed of quartz. Because of the high internal temperature in the active laser volume (more than 1000°C), a ceramic tube insert made out of Al_2O_3 is used to assure the thermo-chemical stability of the discharge. Principal scheme of the laser tube of $SrBr_2$ laser with its components is shown in Fig. 1.

The operation of the strontium bromide ($SrBr_2$) laser depends on the following main characteristics, or input variables: *D*1 (mm) – the internal diameter of the outer (quartz) tube; *D*2 (mm) – the internal diameter of the ceramic tube insert; *La* (cm) – distance between the electrodes (active volume); *Ceq* (pF) – equivalent capacity of the capacitor battery; *Pin* (kW) – supplied electric power in the active laser volume; *PRF* (kHz) frequency impulse repetition rate; *PHe* (Torr) – pressure of the buffer gas helium. To account for 50% loses, during the exploitation of the tube, in all numerical procedures the values of *Pin* are reduced twice, and the new variable is noted by *PIN*2.

We carry out statistical studies by using data from 167 available experiments. The dependent variable is the laser efficiency *Eff* (%). All data are collected from publications [1, 8-10].

The descriptive statistics of the processed data is given in Table 1. It could be mentioned that the distribution differs from the normal distribution and standard statistical analyses as regression are not appropriate.

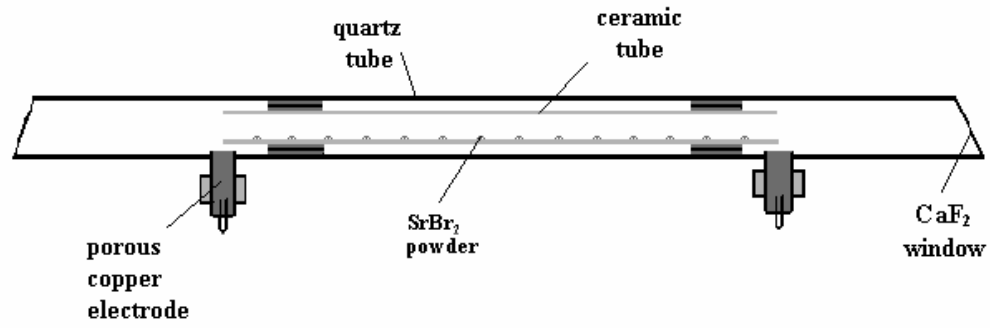

Fig. 1. Longitudinal principal scheme of strontium bromide laser.

Table 1. Descriptive statistics of experiment data[1].

	***D*1, mm**	***D*2, mm**	***La*, cm**	***Ceq*, pF**	***PIN*2, kW**	***PRF*_ KHz**	***PHE*, Torr**	***Eff*, %**
Mean	46.548	19.14	87.955	577.795	0.928	19.231	40.954	1.61
Median	46.000	19.80	98.000	632.300	0.962	19.000	41.600	1.78
Mode	46.0	20	98.0	632.3	1.05	19.0	38.00	1.16
Std. Deviation	0.724	0.869	13.281	142.422	0.116	0.817	5.894	0.400
Variance	0.525	0.756	176.391	20284.079	0.014	0.668	34.734	0.160
Skewness	0.565	-0.565	-0.565	0.181	-0.557	-1.559	-0.544	-0.412
Kurtosis	-1.702	-1.702	-1.702	-1.373	-1.233	8.556	-0.404	-1.552
Minimum	46.0	18	70.5	398.0	0.70	15.0	25.00	0.83
Maximum	47.5	20	98.0	793.1	1.05	22.0	50.00	2.08

[1] Std. Error of Skewness for all variables is 0.188 and Std. Error of Kurtosis is 0.374.

3 Statistical data mining techniques

In this study we apply two types of data mining techniques to process the experimental data for strontium bromide laser. Cluster analysis is used for grouping investigated laser variables in relatively homogeneous groups (clusters) and classification and regression tree (CART) method for determining the dependence of laser efficiency from the operational variables. This makes it possible on the basis of real measurements to find new relationships in description of laser operation.

3.1. Cluster analysis

Cluster analysis or clustering is well known statistical technique aiming to find an optimal grouping of observations or variables, for which the elements of a given cluster are similar (homogeneous) but the clusters are clearly distinguishable from each other. Cluster analysis is not well formalized unique method, but is a set of procedures for identifying the best cluster model and the appropriate number of clusters. In this paper we apply the agglomerative hierarchical cluster analysis, which is recommended for relatively small data samples. The similarity of variables is measured by the usual Euclidean distance $d(\mathbf{a},\mathbf{b})$ between two vectors (points) $\mathbf{a}=(a_1,a_2,\ldots,a_n)$ and $\mathbf{b}=(b_1,b_2,\ldots,b_n)$, defined by $d(\mathbf{a},\mathbf{b})=\sqrt{\sum_{k=1}^{n}(a_k-b_k)^2}$ or squared Euclidean distance. Two clusters are considered to be similar and are grouping when the selected distance between them is small. To avoid different physical measures, all variables are preliminary transformed into dimensionless.

In the beginning of CA all objects (i.e. variables) are considered as initial clusters. The hierarchical procedure is a step by step process for consecutive formation of a new cluster at each step, by merging the two clusters with the smallest distance between them. Different cluster methods with specific distance between clusters could be applied. We will use the following four cluster methods, provided in SPSS: Average linkage (Within-groups), Average linkage (Between-groups), Nearest neighbor (Single linkage) and Ward's method. [5, 11, 12]. As an example, in Average linkage (Between groups), the distance $D(A,B)$ between two clusters A and B is defined using all n_A points of A and all n_B points of B by the expression

$$D(A,B)=\frac{\sum_{i,j}d(x_i,x_j)}{(n_A+n_B)(n_A+n_B-1)} \qquad (1)$$

The optimal number of clusters is achieved when the distance between the formed clusters is relatively large. Usually it can be equal or bigger than 20% of the maximum of all initial distances d. For convenience, the distances are graphically presented in rescaled units, where the maximum is 25 units.

3.2. CART method

Classification and regression tree (CART) method, known also as "decision trees" is a powerful data mining technique, proposed by Breiman et al. in 1984 [3]. This is a supervised learning algorithm used both for classification and regression purposes. If the response variable Y is binary-valued the CART technique is used for classification problems, whereas for regression problems Y could be a continuous variable. The algorithm of the method is capable to process very large or small datasets of arbitrary type of variables $X_1,X_2,\ldots,X_p$. CART tree can handle linear, nonlinear effects and interaction terms in a rule sequence and build the easily interpreted models. CART is characterized as a multivariable nonparametric statistical method [12].

The CART methodology is based on a recursive type partitioning of the initial data of $\mathbf{R}^p$ into non-overlapping multidimensional sub-regions (classes). The structure of a CART tree consists of a root node, internal nodes and terminal nodes, each of them representing a subset of cases falling within a sub-region. At any step the cases in a current node τ could be or not be splitted into two daughter nodes by asking the unique question of the type $x_{i,j}\le\theta_j$, where $x_{i,j}$ is the i-th value of the variable X_j in the current node, and θ_j is a threshold value. The CART algorithm selects the values of i and j from all possible variables $X_j, j=1,2,\ldots,p$ and θ_j from cases in the current node which gives the minimum error in predicting the dependent variable Y. This way, beginning from the root node, each node identifies a specific decision sequence of rules. In regression problems the predicted value $\hat{Y}_\tau$ for cases in a node τ is simply equal to the mean value of their corresponding responses. The usual way to calculate the overall error is the least square error $\sum_{\tau=1}^{M}\left(Y_\tau-\hat{Y}_\tau\right)^2$. The terminal nodes give the final decision about the distribution of dataset and model prediction of existing or new observations.

In machine learning the model performance is estimated not only for its overall accuracy, but rather than from the prediction accuracy for a random test subsample, which has not been used in the modelling procedure. This procedure is called cross-validation [3].

4 Results

We applied the two described techniques, which have to be considered as independent each to other. Cluster analysis is used to classify the variables by similarity principle and to determine the place of the efficiency Eff between the operational variables. CART method is used to establish the relationships in order to construct appropriate regression model, capable to predict the efficiency values and this way to guide the experiment.

4.1 Results from cluster analysis

We used four different cluster methods with Euclidean and squared Euclidean distance. In all analyses the optimal cluster solutions are very similar. They consist of two clusters. As a presentative cluster model, we selected the model obtained by the average linkage (between groups) cluster method with squared Euclidean distance. To note that all variables were preliminary transformed into dimensionless form using the zscores transformation of an variable x, given by the expression $zx_i = (x_i - \bar{x})/s,\ \ i = 1,...,n$, where $\bar{x}$ is the mean value and s is standard deviation of the variable, n is the volume of the sample. The dendrogram of all clusters is shown in Fig. 2.

From Fig. 2 it can be seen that the solution with two clusters: $C1 = \{La, D2, PIN2, PHe, Ceq, Eff\}$ and $C2 = \{D1, Prf\}$ has the largest rescaled distance between all clusters.

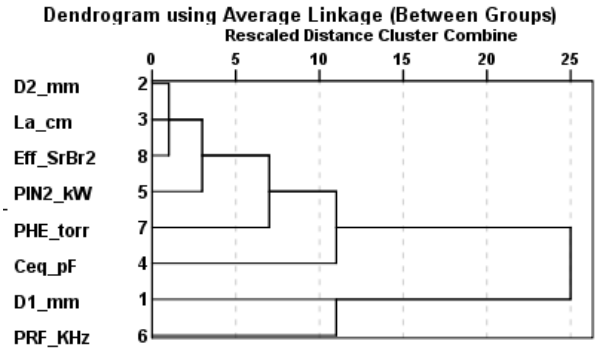

Fig. 2. Dendrogram of all seven input variables and laser efficiency *Eff*, obtained by the Average linkage (Between groups) cluster method with squared Euclidean distance.

The gap from the two clusters *C*1 and *C*2 shows that they are divided by 14 rescaled units one from another. This corresponds to the bigger distance from 224.138 to 512.718 obtained at stage 7 for the two-cluster solution. From Fig. 2 it is also observed that laser efficiency *Eff* (number 8) is grouped at stage 2 with the laser tube length *La* (number 3) and internal diameter of the ceramic tube *D*2 (2). This result could be interpreted, that these two variables are closely related to the efficiency. Then to this cluster is added *PIN*2 (5), etc. to obtain cluster *C*1.

Very similar results are obtained by other cluster methods. For the same type of distance the different cluster methods give the biggest rescaled distance for 2 clusters as follows: Average linkage (Within groups) – 7 rescaled units, Nearest neighbour (Single linkage) – 10 rescaled units, Ward's method – 18 rescaled units. In all models the stages of grouping clusters are the same. We can conclude that the two-cluster solution, consisting of

$$C1 = \{La, D2, PIN2, PHe, Ceq, Eff\}$$
$$C2 = \{D1, Prf\} \quad (2)$$

is very stable and laser efficiency *Eff* for the examined dataset of experiments is closely related to *La*, *D*2, and *PIN*2.

4.2 Results from CART

There were carried out different CART analyses, by experimenting with the so called control parameters of the method and removing the correlated variables. As the number of experiments N=167, it is taken a minimum from 10 to 20 cases in the parent node of the tree and a minimum of 5 to 10 cases in the child node, respectively. We obtained similar solutions by changing the minimum cases in parent and child nodes and removing some of the variables. All models are tested by the usual 10-fold cross-validation procedure against overfitting [3, 12].

We present in this paper one of the best from the obtained CART models. It was established that variable Seq has no significant influence on the efficiency, and *D*1 correlate strongly with *D*2, so that these variables were removed from the CART procedure. The plot of the regression tree until depth three is shown in Fig. 3. The maximum predicted value of *Eff* is calculated in the node

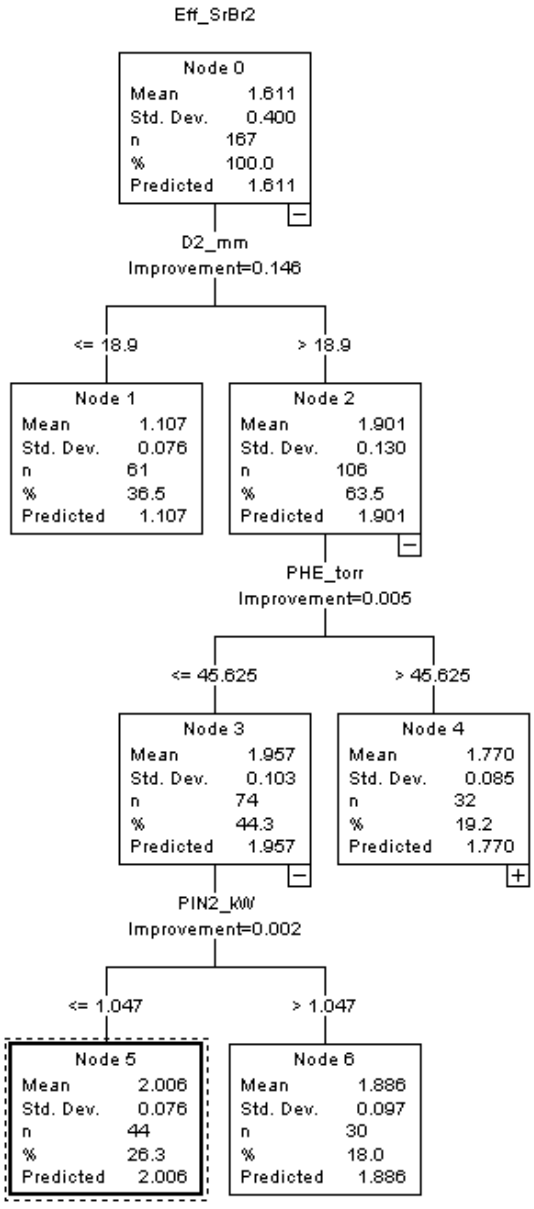

Fig. 3. CART tree for laser efficiency *Eff*.

5 and is equal to 2.006%. This value gives 96.4% accuracy in comparison with the maximum experiment value of *Eff*, which is 2.08% (see also Table 1). This node contains 44 experiments, or 26.3% of all data. To note that the branch from node 4 is truncated and not included in this figure.

CART model is very intuitive and could be easily used for implementation. For example, by following the tree rules in Fig. 3 it is seen that every experiment with *D*2>18.9 mm falls into node 2. Then if *PHe* $\leq$46.625 Torr it goes to node 3. In the case of supplied power *PIN*2 $\leq$ 1.047 kW, it will belong to node 5 and one can expect a maximum laser efficiency *Eff*.

The relative variable importance in the model is as follows: *La* and *D*2 - 100%, *PIN*2 – 86.6%, *PRF* – 76.3% and *PHe* – 70.9%. Thus, from the observed experiment data *D*2 and *La* have the highest influence on efficiency *Eff*. This corresponds to the results from cluster analyses. Some difference with respect to cluster model could be explained by the nonlinear regression dependence between *Eff* and all variables in the CART model.

The overall fitting of the selected CART model to the entire (learn) dataset of experiments is illustrated in Fig. 4. The simple comparison between the experiment and predicted values of efficiency *Eff* shows that the coefficient of determination $R^2 = 96.0\%$, i.e. the model explains 96% of the dataset.

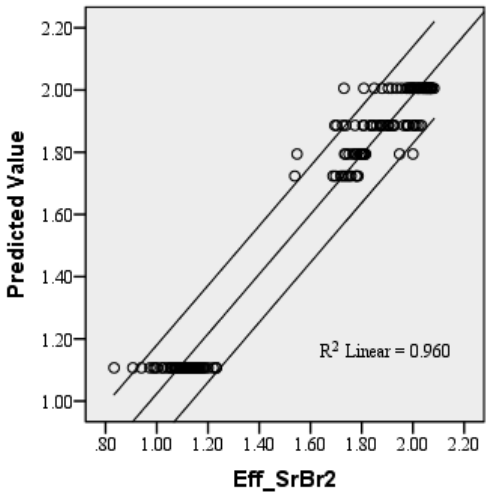

Fig. 4. Experimental versus predicted values of laser efficiency *Eff*, using the selected CART model with 5% confidential interval.

5 Conclusions

By applying two data mining statistical techniques the relations between laser efficiency and seven operational characteristics of $SrBr_2$ laser are obtained based on available experiment data. It was obtained a cluster model for linear type classification of variables in two cluster solution. The more precise results are obtained by CART modelling. The obtained CART model determined the nonlinear type of influence of the operational characteristics on laser efficiency. The predicted values reached over 95% from the maximum measured laser efficiency. There are extracted the main local variation of the investigated laser operational characteristics in order to achieve higher efficiency. The proposed approach can be used for guiding future experiments.

This study was supported by the Scientific and Research Sector of Technical University of Sofia, and the Scientific and Research Department (NPD) of University of Plovdiv Paisii Hilendarski, grant MU17-FMI-003.

References

1. A. Temelkov, N. Vuchkov, I. Freijo-Martin, A. Lema, L. Lyutov, N. Sabotinov, J. Phys. D: Appl. Phys. **42**(11), 115105 (6 pp), (2009)
2. G. Edwards, R. Logan, M. Copeland et al., Nature, **371**.6496, 416-419 (1994)
3. L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and regression Trees* (Wadsworth Advanced Books and Software, Belmont, Canada, 1984)
4. X. Wu, V. Kumar (eds.), *The Top ten algorithms in data mining* (Chapman & Hall/CRC, Boca Raton, 2009)
5. IBM SPSS Statistics, http://www-01.ibm.com/software/analytics/spss/products/statistics/
6. G.D. Chebotarev, E.L. Latush, O.O. Prutsakov, A.A. Fesenko, Quantum Electron. 38(4), 299-308 (2008)
7. G.D. Chebotarev, E.L. Latush, A.A. Fesenko, Quantum Electron. 38(4), 309-318 (2008)
8. K.A. Temelkov, N.K. Vuchkov, B.L. Pan, N.V. Sabotinov, B.Ivanov, L.Lyutov, J. Phys. D: Appl. Phys. **39**, 3769-3772 (2006)
9. K.A. Temelkov, N.K. Vuchkov, B.L. Pan, N.V. Sabotinov, B.Ivanov, L.Lyutov, Proc. of SPIE, Bellingham, WA, **6604**, 660410, 1-5 (2007)
10. K.A. Temelkov, N.K. Vuchkov, B. Mao, E.P. Atanasov, L. Lyutov, N.V. Sabotinov. IEEE J. Quant. Electron. **45**(3), 278-281 (2009)
11. S. Gocheva-Ilieva, I. Iliev, *Statistical models of characteristics of metal vapor lasers* (Nova Science, New York, 2011)
12. A.J. Izenman, *Modern multivariate statistical techniques: regression, classification, and manifold learning* (Springer, New York, 2008