DOI: 10.20472/IAC.2016.024.033

SNEZHANA GOCHEVA-ILIEVA

Plovdiv University , Bulgaria

ILIYCHO ILIEV

Technical University of Sofia, branch Plovdiv, Bulgaria

USING GENERALIZED PATHSEEKER REGULARIZED REGRESSION FOR MODELING AND PREDICTION OF OUTPUT POWER OF CUBR LASER

Abstract:

A Generalized PathSeeker Regularized Regression (GPSRR), based on data mining approach, is applied for statistical modeling and prediction of output power of copper bromide vapor lasers. The aim is on the basis of available experimental data to construct appropriate predictive models of the output power of the lasers depending on 10 operating laser characteristics in order to direct future experiments and designing new laser devices with increased output power. In particular, the influence on model performance and predictive ability of several data transformations, used to improve the normality of the distribution of the dependent variable is investigated. As a main result, numerous combined models, built by GPSRR with data mining techniques are obtained and their adequacy is established by cross-validation. It is found that the best combined models demonstrate up to 98-99% of fitting the experimental data. The combined models with the proposed preliminary transformations improve the adequacy and predictive ability of GPSRR in the region of high values of the output power by up to 10%. This was established both for learn and test random samples, showing a perfect out-of-sample performance of this type of model approach. The models are applied for predicting of laser output power for new laser devices of the same type by up to 15%.

Keywords:

Regularized regression, Generalized PathSeeker, LASSO, TreeNet (Stochastic Gradient Boosting), Copper bromide vapor laser

JEL Classification: C15

1. Introduction

It is well known that the best developed directions for finding relationships within datasets encompasses a large number of regression methods. An overview of the current state and the capabilities of prediction techniques, including regression and data mining statistical techniques, can be found in (Nisbet et al. 2009; Hastie et al. 2001). In this study one of the latest regression methods, namely the Generalized PathSeeker regularized regression (GPSRR) is used for solving a problem from laser technology. GPSRR is a new generation statistical method, developed by Jerome Friedman (2012). The GPSRR method has been realized in the spring of 2013 as a part of the Salford Predictive Modeler (SPM) software package (SPM 2013). It was shown by examples that GPSRR used in combination of preprocessing TreeNet stochastic gradient boosting and associated data mining techniques overperforms all existing predictive methods (Friedman 2012).

In this paper the GPSRR is applied to study experimental data for a family of copper bromide (CuBr) vapor lasers (Sabotinov 2006). The dependent variable is the output laser power (laser generation), considered in relation to 10 laser operating characteristics. These data have been investigated in literature by different parametric and nonparametric statistical methods. In (Gocheva-Ilieva and Iliev, 2011) multiple linear regression, factor analysis with principal component regression, nonlinear regression and multivariate adaptive regression splines (MARS) have been applied for modeling output laser characteristics – laser output power and laser efficiency of CuBr laser. In recent papers (Iliev et al. 2012, 2013) high quality models have been constructed by means of MARS and CART methods (Friedman 1991; Breiman et al. 1984; Nisbet et al. 2009). In (Gocheva-Ilieva 2014) some GPSRR models have been presented. However, for these models the predictions, especially in the most important region of the higher laser output power are not satisfactory and differ from the experiment about 10 to 20%.

The goals of this study are: (i) to find appropriate preliminary data transformation for improving the distribution of the initial data closer to the normal distribution; (ii) to construct significantly improved combined models (based on GPSRR and data mining techniques) of laser generation of CuBr vapor lasers with higher predictive ability in the region of the highest values of the output power; (iii) to apply the models for prediction of future experiment directed a development of new laser devices with increased output power. The results are obtained by means of the Salford Predictive Modeler and author's programming codes in Wolfram Mathematica software.

2. Data and methods

2.1 Description of experimental data for copper bromide vapor laser

The CuBr vapor laser is a type of metal vapor laser. It emits in the visible spectrum at two wavelengths – green (510.6 nm) and yellow (578.2 nm). The CuBr laser has better characteristics compared to lasers in the infrared spectrum (CO2 lasers) including better laser beam convergence and focus, less noise, strong beam coherence, etc. It is the most efficient source of visible light among metal vapor lasers. CuBr lasers find many practical applications in industry for the micro processing of different kinds of materials for drilling, cutting, marking, engraving, etc., in high-speed photography, military industry, nanotechnology, pulsed holography, for aerial and naval navigation, in atmospheric and ocean pollution studies, in medicine and medical research, entertainment and advertising and more (Sabotinov 2006; Foster 2005; Zureng et al. 1992; Son et al. 2014; Gocheva-Ilieva and Iliev 2011).

Due to its wide range of practical applications the laser continues to be scientifically and experimentally developed (Sabotinov 2006). One of the important technological objective is the development of new laser devices of this type with enhanced output power. In particular, statistical modeling supports the investigation of the influence of the main laser operating characteristics (laser tube geometry, input power, neutral gas pressure, etc.) on output laser power and allows for predictions to be made. A detailed description of the laser device could be found in (Sabotinov 2006).

In this study we consider the experimental data of CuBr laser devices, developed and patented from the Laboratory of Metal Vapor Lasers at Georgi Nadjakov Institute of Solid State Physics, Bulgarian Academy of Sciences. The data are collected during the last 40 years.

The dataset comprises 10 input basic variables which determine the basic CuBr laser operation. These include: D (mm) – inner diameter of the laser tube, DR (mm) – inner diameter of the rings, L (cm) – electrode separation (length of the active area), PIN (Kw) – input electrical power, PRF (kHz) – pulse repetition frequency, PNE (Torr) – neon gas pressure, PH2 (Torr) – hydrogen gas pressure, PL (Kw/cm) – specific electrical power per unit length, C (Nf) – equivalent capacity of the condensation battery, TR (^{o}C) – temperature of CuBr reservoirs. The response variable is Y=Pout (W) – output laser power.

The sample size is n=387. We have to mention the complexity, long duration and high cost of each conducted experiment. The descriptive statistics of the data sample are represented in Table 1.

2.2 Description of GPSRR and Data Mining Techniques

The GPSRR method is a generalization of a variety of methods applying a regularization in the sense of Tikhonov (Tikhonov 1963). A regularization term is added to the error and various penalties are imposed in order to determine the regression coefficients and thus achieve better fitting to data by the obtained models. The result is a pool of regression models.

Table 1: Descriptive statistics of copper bromide laser characteristics[a]

	Pout,	D,	DR,	L,	PIN,	PL,	PH2,	PRF,	PNE,	C,	TR,
	W	mm	mm	cm	Kw	Kw/cm	Torr	kHz	Torr	Nf	^{o}C
Mean	35.9	46.96	35.30	109.35	2.14	10.84	0.38	21.73	19.94	1.25	474.4
Median	18.3	46.00	30	50.00	1.40	12.00	0.50	17.00	20.00	1.10	485.0
Std. Dev.	35.1	9.80	18	69.81	1.26	2.50	0.24	22.16	12.30	0.57	32.8
Skewness	0.71	-0.81	0.24	0.40	1.0	-0.46	-0.56	4.28	6.59	1.74	-2.38
Kurtosis	-0.94	1.51	-1.59	-1.70	-0.39	0.04	-1.28	17.26	53.36	5.25	7.11
Minimum	1.5	15	4.5	30	1	1	0.00	4	8	0	350
Maximum	120	58	58.0	200	5	16.67	0.80	126	150	3.83	590

[a] n valid for all variables is 387, Std. Error of Skewness = 0.124, Std. Error of Kurtosis = 0.247.

Unlike the conventional regression approach, the GPSRR method is implemented as a machine learning technique. Model selection is determined in accordance with the chosen type of penalty and its quality is assessed based on performance on test samples. To this end, the data are divided randomly in two parts – learn and test samples. The models are built using only the data from the learn sample and applied to predict the known values from the test sample. Various assessment and validation methods could be applied with both classic statistical indices and different techniques for cross-validation, bootstrap, etc. (Friedman 2012; SPM 2013).

Consider a dataset of n observations $\{Y_i,\mathbf{X}_i\}_{i=1}^{n}=\{Y_i,X_{i1},...,X_{im}\}_{i=1}^{n}$. As in ordinary multiple linear regression, in the GPSRR one aims to find an equation in the linear form

$$\hat{Y}=a_0+\textstyle\sum_{j=1}^{m}a_jX_j \tag{1}$$

for fitting to these data, where $\mathbf{a}=\{a_0,a_1,...,a_m\}$ is the vector of the unknown regression coefficients. To obtain more exact estimates $\hat{\mathbf{a}}$ of $\mathbf{a}$, one solves the optimization problem

$$\hat{\mathbf{a}}(\alpha)=\underset{\mathbf{a}}{\arg\min}\left[\widehat{R}(\mathbf{a})+\alpha P(\mathbf{a})\right] \tag{2}$$

where $R(\mathbf{a})$ is the empirical loss function, selected among different error criteria, e.g. the sum of squared errors (SSE) $R(\mathbf{a})=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2/n$, $P(\mathbf{a})$ is a penalty function and $\alpha>0$ is the regularization parameter.

In the GPSRR method, the calculations are carried out through sequential path search directly in the parameter space under a given penalty $P(\mathbf{a})$ without solving the optimization problem (2) at each step. Especially, the penalty function in (2) is taken to satisfy for all $\mathbf{a}$ the condition $\frac{\partial P(\mathbf{a})}{\partial |a_j|} > 0, j = 1,...,m$.

This includes as a special case the well-known power family for penalty function $P(\mathbf{a}) = P_\gamma(\mathbf{a}) = \sum_{j=1}^{m} |a_j|^\gamma,\ 0 \le \gamma \le 2$. The case $\gamma = 2$ corresponds to Ridge regression (Hoerl and Kennard 1970) and the case $\gamma = 1$ - to the Lasso (least absolute shrinkage and selection operator) (Tibshirani 1996). Other methods of this type are represented by the generalized elastic net family with the penalty function $P_\beta(\mathbf{a})$ given by $P_\beta(\mathbf{a}) = \sum_{j=1}^{m} (\beta - 1) a_j^2 / 2 + (2 - \beta)|a_j|,\ 1 \le \beta \le 2$, where β is the coefficient of elasticity (Zou and Hastie 2005). An extension for $0 \le \beta < 1$ is obtained in (Friedman 2012).

GPSRR is realized in Salford SPM package as a very fast forward stepping algorithm with specialized variable selection procedures (Friedman 2012, SPM 2013). There is generated the collection of models by constructing a path based on selected predictors $\mathbf{X}$ as a sequence of iterations (steps) in the space of coefficients. At every step a new variable, selected to fulfill a complex of criteria is added, or the coefficient of some model variable is adjusted. The quality of models can be assured by selecting from a number of commonly used goodness-of-fit measures for learn and test samples as the coefficient of determination R^2, MSE (mean squared error), AIC, BIC, etc. and validated by ross-validation procedures.

Despite its advantages, the GPSRR method exhibits some limitations. The method does not provide automatic discovery of nonlinearities, interactions between predictors, or a missing values handling feature. To this end, TreeNet stochastic gradient boosting is applied for preprocessing the data (Friedman 2001). Also, the usage of GPSRR as a data mining engine is highly efficient in combination with model simplification features, realized in SPM, such as ISLE (Importance Sampled Learning Ensembles) and RuleLearner (rule ensembles) (Friedman and Popescu 2001, 2003). The original model, produced by TreeNet is an ensemble of hundreds or even thousands of small trees, represented new variables for the model (1). Many of them are usually equal or have very similar structure. The compression of the TreeNet model can be performed using an ISLE algorithm by removing redundant trees. The coefficients of the models are then adjusted by the GPSRR algorithm. Finally, the RuleLearner algorithm has to be applied as a post-processing technique which selects the most influential subset of nodes, thus reducing model complexity of TreeNet. Different combinations of abovementioned techniques can be also carried out to obtain the best required models.

2.3 Parametric Transformations

As noted above, the obtained initial GPSRR models in (Gocheva-Ilieva 2014) are not satisfactory, especially in the most important region of the higher laser output power, and differ from the experiment about 10 to 20%. In order to improve the quality of the models and predictions we will perform some preliminary transformations of the dependent variable.

The distribution of the original data for $Y=Pout$ is shown in Figure 1(a). To approximate the distribution more closed to the normal distribution we apply the following sequence of parametric transformations:

$$Y \to y1 \to z1 \to y2 \to z2 . \qquad (3)$$

In order to reduce the skewness of the distribution, first we apply the usual shifted Box-Cox transformation

$$y1 = Log[Y] , \qquad (4)$$

where Log means the natural logarithm.

In (3) the standardizations $z1$, $z2$ are performed by selecting from two robust transformations. The first robust transformation $R1$ is given by the expression (Koekemoer and Swanepoel 2008):

$$R1: \quad z(t) = \left(t - \hat{\mu}_x\right) / \hat{\sigma}_x, \quad \hat{\mu}_x = med(t), \quad \hat{\sigma}_x = \max\left\{s_x, \left(\hat{q}_3 - \hat{q}_1\right) / \left\{\Phi^{-1}(3/4) - \Phi^{-1}(1/4)\right\}\right\} \quad (5)$$

where $med(t)$ is the sample median, s_x^2 is the unbiased sample variance, $\hat{q}_1, \hat{q}_3$ are the first and third sample quartiles, Φ is the standard normal distribution function.

The second robust transformation $R2$ is (Van der Veeken 2010):

$$R2: \quad z(t) = \left(t - \hat{\mu}_x\right) / \hat{\sigma}_x, \quad \hat{\mu}_x = med(t), \quad \hat{\sigma}_x = 1.483\, med\left(\left|t - med(t)\right|\right) . \qquad (6)$$

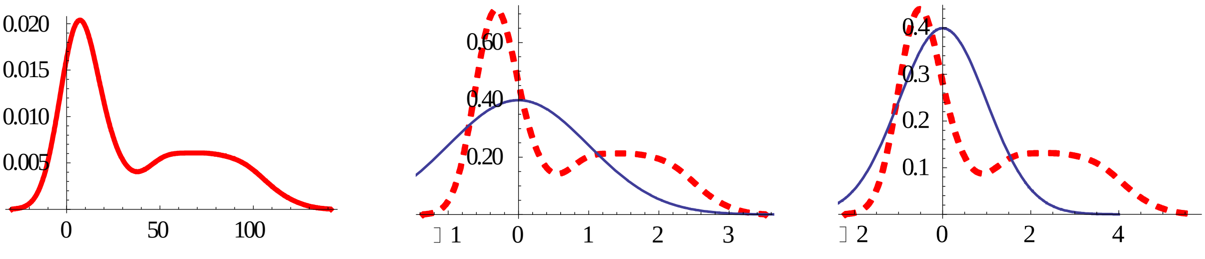
Parametric transformations (5) and (6) do not change the type of distribution. The next transformation $y2 = y2(z1)$ is the John-Drapper transformation (John and Draper 1980)

$$y_{JD}(t,\lambda) = \begin{cases} sign(t)\left\{(|t|+1)^{\lambda} - 1\right\} / \lambda, & \lambda \neq 0 \\ sign(t) Log(|t|+1), & \lambda = 0 \end{cases} . \qquad (7)$$

In (7), the optimal values of λ are estimated by the Jarque-Bera goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution (Jarque and Bera 1980; Gel and Gastwirth 2008) and by the Shapiro-Wilk test of normality (Shapiro and Wilk 1965). To this end, at a step of 0.01 for λ, the minimums of Jarque-Bera statistics were calculated by using an author's code written in Wolfram Mathematica.

The distribution of the dependent variable Y=Pout and the corresponding standardized distributions R1(Y) and R2(Y) are shown in Figures 1: (a), (b), and (c), respectively. The obtained statistics of the distribution are: Jarque-Bera statistics=47.121, p-value=0.00004, Shapiro-Wilk p-value=1.10^{-19}.

Figure 1: Smooth histograms of the initial and applied standardized robust distributions (dashed line) in comparison with the normal distribution



(a) Y=Pout

(b) R1(Y)

(c) R2(Y)

3. Results with discussions

3.1 Results from the preliminary data transformations

By applying the procedure, described in the previous section, there were obtained the following optimal values of ν for improving the distribution of $y2 = y_{JD}(z1, \lambda)$ from (3), (7):

$$\lambda_1 = 2.84, \quad \lambda_2 = 3.23 \tag{8}$$

The first value λ_1 is used in the sequence of transformations:

$$Y \to y1 = \mathrm{Log}(Y) \to z1 = R1(y1) \to y2 = y_{JD}(z1, \lambda_1). \tag{9}$$

The minimum Jarque-Bera statistics for λ_1 is 12.061, p-value=0.00974, Shapiro-Wilk p-value is 8.10^{-8}.

The second optimal value λ_2 is found for the sequence of transformations:

$$Y \to y1 = \mathrm{Log}(Y) \to z1 = R2(y1) \to y2 = y_{JD}(z1, \lambda_2). \tag{10}$$

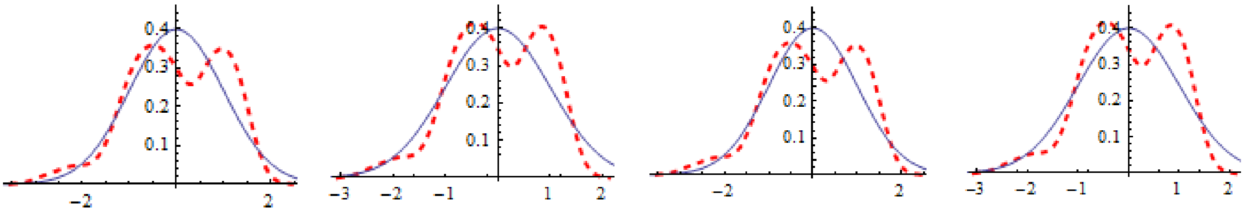
The corresponding statistics are: Jarque-Bera=12.134, p-value=0.00942, Shapiro-Wilk p-value = 6.10^{-8}.

The final transformations in (9) or (10) are performed by $z2 = R1(y2)$ or $z2 = R2(y2)$. The notations of a given sequence of transformations of Y=Pout in (3) are shortly denoted by

$$Rs[\lambda_j]Rk, \; s = 1,2; \; j = 1,2; \; k = 1,2. \tag{11}$$

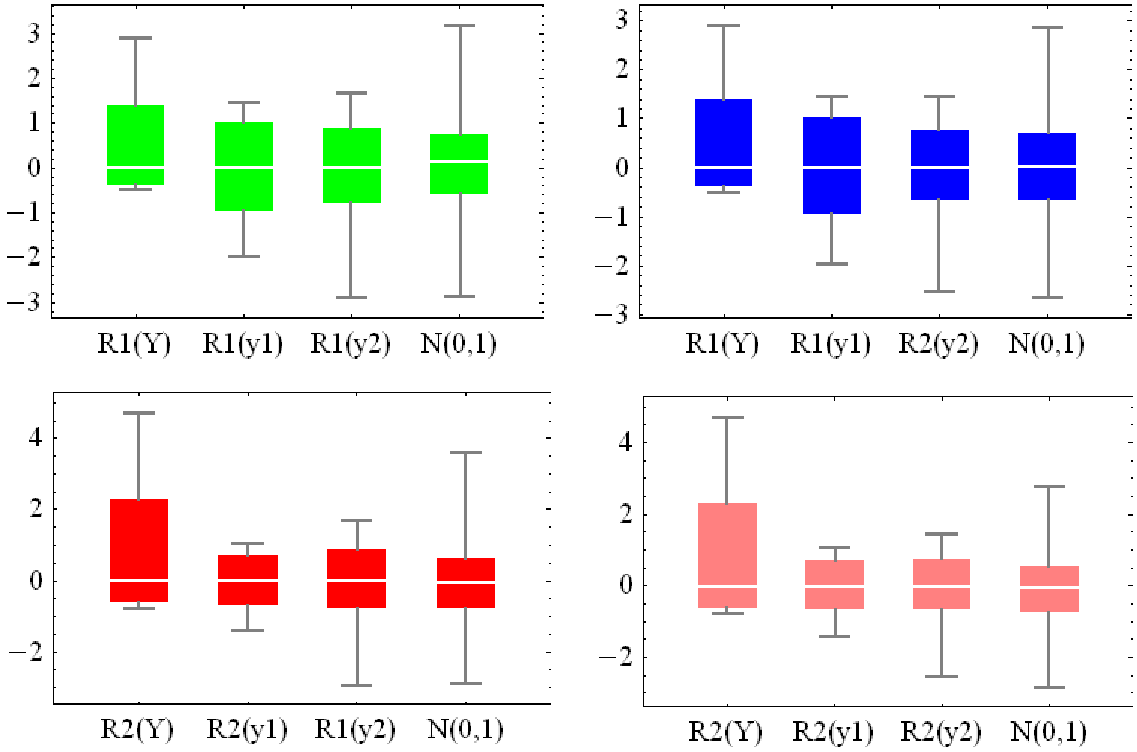
The histograms and boxplots of the obtained improved distributions by (11) are shown in Figure 2 and Figure 3, respectively. It is observed that the transformations R1 [λ_1] R2 and R2 [λ_2] R2 give the distributions more closed to the normal. Although these distributions are not very close to normal, they help to significantly improve the modeling results.

Figure 2: Smooth histograms of the optimal distributions after transformations compared with the normal curve



R1 [λ_1] R1 R1 [λ_1] R2 R2 [λ_2] R1 R2 [λ_2] R2

Figure 3: Box-and-whisker plots of standardized sequential transformations of Y. For comparison the last element in every group graph presents the box plot of the standardized normal distribution N(0,1).



3.2 Construction of GPSRR and Combined Models

A large number of models were built for untransformed dependent variable Y=Pout and for its different preliminary transformations from (3) – (11).

For each representation in (11), the GPSRR method and data mining techniques were applied to obtain combined models. For any of the input data for Y=Pout the pipeline of models comprises GPSRR models extracted by TreeNet, RuleLearner and ISLE techniques, and their combinatitions, with and without including the initial raw predictors. The models have been validated by 10-fold cross-validation technique (SPM 2013). This means that the input sample is randomly divided into 10 equal non-intersecting sub-samples. For each division, the 90% sample is the learn sample and the sub-sample is the test sample.

Normally, the main criterion for model performance is the coefficient of determination R^2, for the learn and test samples. Standard measures of goodness-of-fit of models such as MSE (mean squared error), RMSE (root MSE), MAD (mean absolute deviation) are also taken into account. The model performance of the obtained best 7 models M0, M1, …, M6 for GPSRR and combined GPSRR with the above mentioned data mining techniques for untransformed and transformed variables in (3) are given in Table 2.

Table 2: Summary performance of the best GPSRR and combined models of laser output power Pout[a]

Model	M0	M1	M2	M3	M4	M5	M6
Transformations	-	R1 [λ_1] R1	R1 [λ_1] R1	R1 [λ_1] R2	R2 [λ_2] R1	R2 [λ_2] R1	R2 [λ_2] R2
Method	RL_RP	RL_RP	ISLE_RL_RP	RL_RP	RL_RP	ISLE_RL_RP	RL_RP
Learn R^2	0.9900	0.9870	0.9882	0.9872	0.9870	0.9869	0.9871
Learn N Coef.	441	185	187	192	209	150	535
% Compression	76.5%	95.1%	95.0%	94.9%	94.4%	96.0%	85.8%
Learn RMSE	3.5023	0.1139	0.1084	0.0980	0.1140	0.1142	0.0980
Learn MSE	12.266	0.0130	0.0118	0.0096	0.0130	0.0130	0.0096
Learn MAD	2.718	0.0680	0.0635	0.0587	0.0686	0.0682	0.0590
Test R^2	0.989	0.9690	0.9695	0.9695	0.9700	0.9689	0.9701
Test N Coef.	441	185	187	192	532	177	535
% Compression	78.0%	96.3%	96.3%	96.2%	89.4%	96.7%	89.3%
Elasticity	(1.1)	(1.0)	(1.1)	(1.1)	(1.1)	(1.1)	(1.1)
Maximum Predicted Pout, W	110.84	118.73	115.24	118.61	119.09	115.97	119.48

[a] Short notations for the methods are: RL_RP (RuleLearner _RawPredictors), ISLE_RL_RP (ISLE_RuleLearner_RawPredictors).

In Table 2, the best model M0 obtained by initial data of Pout, without preliminary transformations, is of type RuleLearner_Raw Predictors and has R^2=99.0%. But all other measures as MSE, RMSE and MAD are very high with respect to the other models. M0 predicts only 111W for the maximum of experimental 120W Pout. This way the model is not satisfactory, especially in higher output powers. The maximum measured output laser power is Pout=120W (see Table 1).

It is also observed from Table 2, that all the best models M1 – M6 demonstrate very good statistical indices both for learn and test samples with R^2=98-99.0%. All other indices are relatively small and are almost equal. To note that the models M3 and M6 are characterized by minimal errors RMSE, MSE and MDA, and give the best performance in predictions. The predicted maximum values for Pout for models M2 and M5, obtained by ISLE_RuleLearner_RawPredictors method are slightly smaller than the other. This is valid for the last higher 20 observations of Pout. Since our goal is to get as more accurate predictions for higher values of Pout, these two models will be omitted from our further analysis. Figure 4 is an example of the output from the Salford SPM software environment, representing the model performance in the case of R2 [3.23] R1 data.

Figure 4: Comparative plot of model performance (in terms of R^2) for R2 [3.23] R1 data. The first model (noted as Orig.) corresponds to the pure TreeNet model.

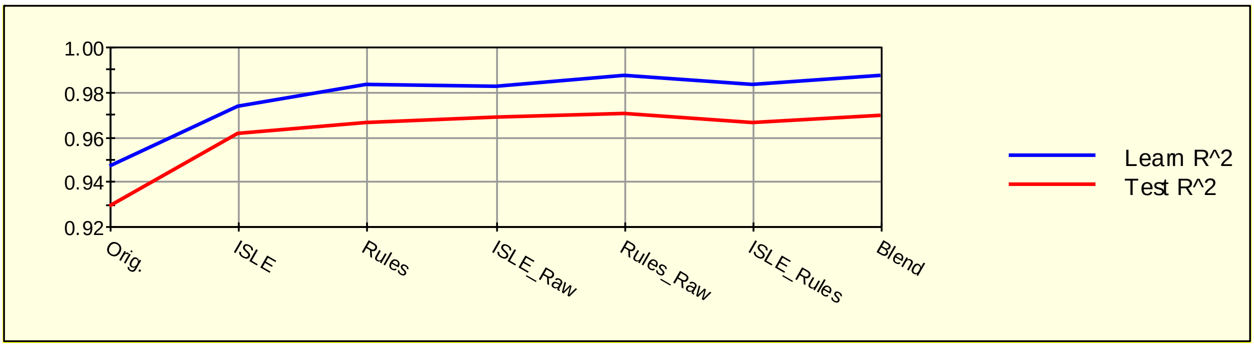


Table 3: Variable importance for the best models of laser output power Pout[a]

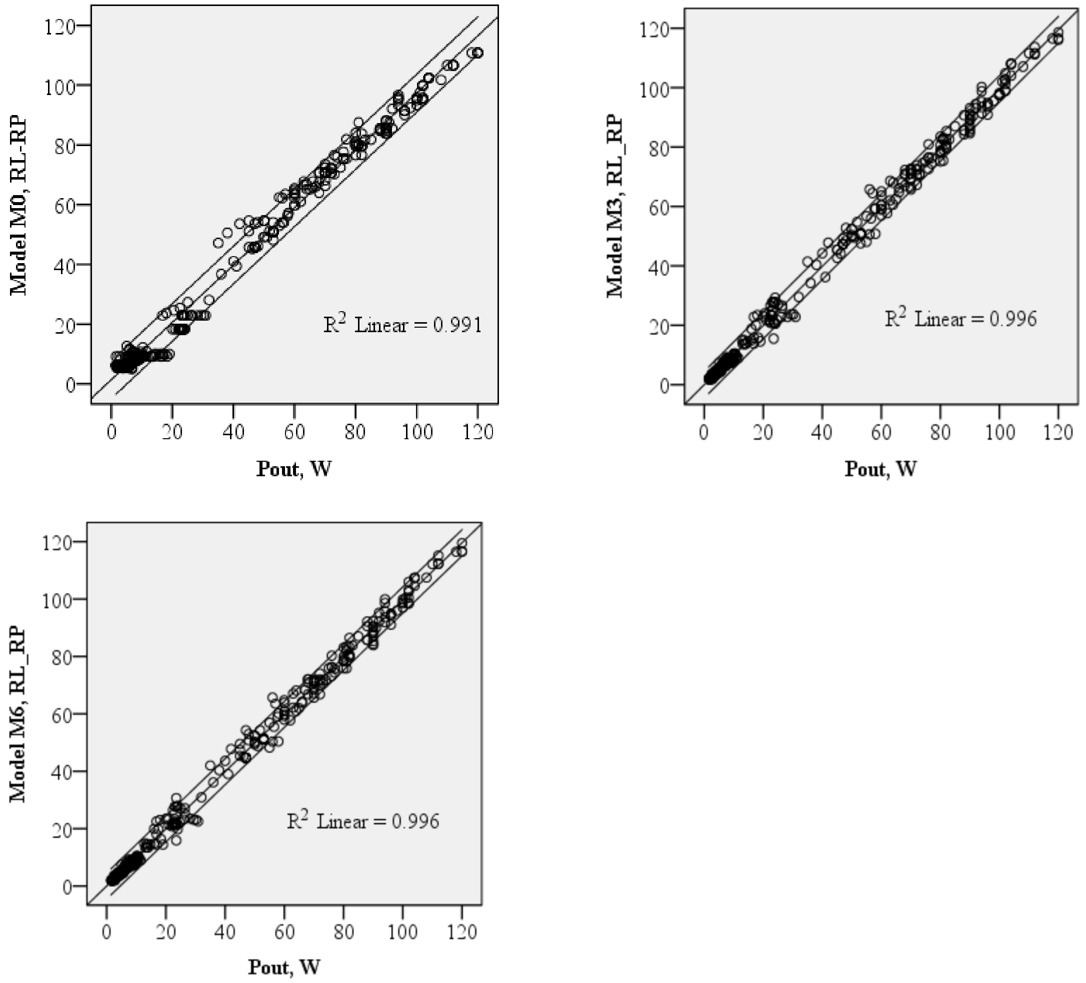
Model	Relative importance
M0	PIN(100), DR(23), PRF(18), C(17), DR(15), PH2(6), L(4), PL(2), PNE(2), TR(0)
M1	PIN(100), DR(48), PH2(36), D(20), PRF(19), L(17), PNE(16), C(14), TR(14), PL(5)
M2	PIN(100), DR(48), PH2(36), D(20), PRF(19), L(17), PNE(16), C(14), TR(14), PL(5)
M3	PIN(100), DR(48), PH2(36), D(20), PRF(19), L(17), PNE(16), C(14), TR(14), PL(4)
M4	PIN(100), DR(52), PH2(36), D(20), L(20), PRF(19), PNE(17), C(14), TR(14), PL(5)
M5	PIN(100), DR(52), PH2(36), D(20), L(20), PRF(19), PNE(17), C(14), TR(14), PL(5)
M6	PIN(100), DR(52), PH2(36), D(20), L(20), PRF(19), PNE(17), C(14), TR(14), PL(4)

[a]The maximum value is taken to be equal to 100 absolute units.

Another important point is the evaluation of the impact of the variables in the model. Table 3 shows the importance of the ten predictors in the derived models. It is observed similar values for models M1-M6. One can conclude that the influence of the predictors is stable within the 3-4 relative units. Big difference is observed for M0 and other models, which explains its lower performance.

Scatter plots in Figure 5 illustrate the comparison between Pout and the predicted values by the models M0 (without transformation), M3 and M6, respectively. The experimental measurements are very well replicated by the last two models. It can be observed that the models M3 and M6 give much better prediction, especially for the last 50 cases.

Figure 5: Comparison of all experimental data versus the predicted values with a 5% confidence intervals from: model M0, model M3 and model M6 (see also Table 2)



3.3 Using the Models for Prediction of Future Experiments

As it is mentioned above from a practical point-of-view, the predictions of the models in the region of high laser output powers are more important. Moreover, the ultimate goal is the ability of models to predict future experimental results (extreme experiment).

To this end, some cases (sets) of given values for the input laser characteristics have been selected as "future" experiment. Based on the best models, the corresponding outcomes were calculated. The obtained predicted values are given in Table 4. The model M0, obtained by the method GPSRR/RuleLearner_RawPredictors shows only 1% increase. For the other models derived by the same method for the transformed data, it is observed up to 15% possible increase in the output laser power.

Table 4: Prediction for future experiments with increased output power[a]

Case	Laser characteristics								Model				
	D, mm	DR, mm	L, cm	PIN, Kw	PL, Kw/cm	PH2, Torr	PNE, Torr	TR, °C	M0	M1	M3	M4	M6
0	58	58	200	5	12.5	0.6	20	490	111	119	117	119	120
1	68	65	220	5.3	11.82	0.6	20	500	116	128	128	128	128
2	70	68	240	5.7	11.88	0.5	21	500	120	132	132	131	131
3	70	70	240	5.5	11.46	0.3	20	500	117	122	121	120	120
4	70	70	240	5.6	11.67	0.4	20	500	119	129	129	128	128
5	70	70	240	5.7	11.88	0.5	21	500	120	134	134	132	132
6	75	73	230	5.3	11.52	0.3	20	500	115	123	121	120	120
7	75	73	230	5.3	11.52	0.5	20	500	117	133	133	131	131
8	75	73	240	5.5	11.46	0.3	21	500	118	124	122	121	121
9	75	73	240	5.5	11.46	0.5	21	500	119	135	135	132	133
10	75	75	240	5.7	11.88	0.3	21	500	120	128	126	125	125
11	75	75	240	5.7	11.88	0.5	21	500	121	138	138	136	136
									1%	**15%**	**15%**	**13%**	**14%**

[a] The values of some operating laser characteristics are fixed as follows: *PRF*=17.5kHz and *C*=1.3Nf. The case 0 is the experiment with the highest measured Pout=120W.

4. Conclusion

Appropriate transformations of data for improving the distribution of the dependent variable closer to the normal distribution were proposed. Based on the new GPSRR method, enhanced by the machine learning statistical techniques, high performance models of laser generation of CuBr metal vapor lasers depending on 10 basic laser operating characteristics were built. The best models were selected with higher predictive ability in the region of the highest values of the output power. The models were applied for predicting of future experiment that showed up to 15% increasing of the output power of the laser device.

The derived models can be further used in estimation and prediction of current and future experimental outcomes in order to improve the output laser characteristics, in our case the very important one – the output laser power. Along with the importance of the new derived models for the considered problem in laser technology, the proposed methodology and results clearly show that preliminary transformations of data to improve the normality of the data could significantly increase the predictive ability of the new GPSRR technique. This is especially effective in practical applications for relatively small data samples, in presence of high multicollinearity, non-linear dependencies and non-normal distribution of the initial variables.

Acknowledgement

This study is supported by the project NPD-NI15 FMI-004.

References

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Belmont: Wadsworth Advanced Books and Software.

FOSTER, P. G. (2005). *Industrial Applications of Copper Bromide Laser Technology*. Adelaide: University of Adelaide, School of Chemistry and Physics, Department of Physics and Mathematical Physics. Ph.D. dissertation.

FRIEDMAN, J. H. (1991). Multivariate Adaptive Regression Splines (with Discussion). *Annals of Statistics*. Vol. 19, No. 1, pp. 1-141.

FRIEDMAN, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistic.* Vol. 29, No. 5, pp. 1189-1232.

FRIEDMAN, J. H. (2012). Fast Sparse Regression and Classification. *International Journal of Forecasting*. Vol. 28, No. 3, pp. 722-738.

FRIEDMAN, J. H. and POPESCU, B. E. (2003). *Importance Sampled Learning Ensembles*, Technical report. Stanford University. http://www-stat.stanford.edu/~jhf/ftp/isle.pdf.

FRIEDMAN, J. H. and POPESCU, B. E. (2005). *Predictive Learning via Rule Ensembles*. Technical report. Stanford University. http://www-stat.stanford.edu/~jhf/ftp/RuleFit.pdf.

GEL, Y. R. and GASTWIRTH, J. L. (2008). A Robust Modification of the Jarque–Bera Test of Normality. *Economics Letters*. Vol. 99, No. 1, pp. 30-32.

GOCHEVA-ILIEVA, S. G. (2014). Application of Generalized PathSeeker Regularized Regression. In: *Proceedings of the 43th Spring Conference of the Union of Bulgarian Mathematicians*, Borovets, Bulgaria, pp. 34-43. http://www.math.bas.bg/smb/2014_PK/tom_2014/pdf/034-043.pdf.

GOCHEVA-ILIEVA, S. G. and ILIEV, I. P. (2011). *Statistical Models of Characteristics of Metal Vapor Lasers*. New York: Nova Science Publishers.

HASTIE, T.; TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer.

HOERL, A. E. and KENNAR, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*. Vol. 42, No. 1, pp. 80-86.

ILIEV, I. P.; VOYNIKOVA, D. S. and GOCHEVA-ILIEVA, S. G. (2012). Simulation of the Output Power of Copper Bromide Lasers by the MARS Method. *Quantum Electronics*. Vol. 42, No. 4, pp. 298–303.

ILIEV, I. P.; VOYNIKOVA, D. S. and GOCHEVA-ILIEVA, S. G. (2013). Application of the Classification and Regression Trees for Modeling the Laser Output Power of a Copper Bromide Vapor Laser. *Mathematical Problems of Engineering*. Vol. 2013, Article ID 654845, pp. 1-10.

JARQUE, M. and BERA, A. K. (1980). Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals. *Economics Letters*. Vol. 6, No. 3, pp. 255-259.

JOHN, J. A. and DRAPER, N. R. (1980). An Alternative Family of Transformations. *Journal of the Royal Statistical Society: Series C*. Vol. 29, pp. 190-197.

KOEKEMOER, G. and SWANEPOEL, J. W. H. (2008). A Semi-parametric Method for Transforming Data to Normality. *Statistics and Computing*. Vol. 18, No. 3, pp. 241-257.

NISBET, R.; ELDER, J. and MINER, G. (2009). *Handbook of Statistical Analysis and Data Mining Applications*. Burlington: Elsevier Academic Press.

SABOTINOV, N. V. (2006). Metal vapor lasers, in ENDO, M. and WALTER, R. F. (Eds.) *Gas Lasers*. Boca Raton: CRC Press, pp. 449–494.

SHAPIRO, S. S. and WILK, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika.* Vol. 52, No. 3-4, pp. 591-611. doi:10.1093/biomet/52.3-4.591.

SON, I. P.; PARK, K. Y.; KIM, B. and KIM, M. N. (2014). Pilot Study of the Efficacy of 578 nm Copper Bromide Laser Combined with Intralesional Corticosteroid Injection for Treatment of Keloids and Hypertrophic Scars. *Annals of Dermatology*. Vol. 26, No. 2, pp. 156-61. http://dx.doi.org/10.5021/ad.2014.26.2.156.

SPM v7.0 User Guide (2013). San Diego: Salford Systems. http://www.salford-systems.com/products/spm/userguide.

TIBSHIRANI, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B*. Vol. 58, No. 1, pp. 267-288.

TIKHONOV, A. N. (1963). Solution of Incorrectly Formulated Problems and the Regularization Method. *Doklady Akademii Nauk SSSR.* Vol. 151, pp. 501-504. Translated in: *Soviet Mathematics.* Vol. 4, pp. 1035–1038.

VAN DER VEEKEN, S. (2010). *Robust and Nonparametric Methods for Skewed Data*. Leuven: Katholieke Universiteit Leuven, Arenberg Doctoral School of Science, Engineering and Technology, Faculty of Sciences, Department of Mathematics. Dissertation of Doctor in Sciences. https://lirias.kuleuven.be/bitstream/123456789/286471/1/finaleversiephdvanderveekenstephan.PDF.

ZOU, H. and HASTIE, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B*. Vol. 67, pp. 301-320.

ZURENG, X.; GUIYAN, Z. and FUCHENG, L. (1992). Applications of the CuBr Vapor Laser as an Image-brightness Amplifier in High-speed Photography and Photomicrography. *Applied Optics*. Vol. 31, pp. 3395-3397.