TRANSNAV
http://www.transnav.eu
the International Journal
on Marine Navigation
and Safety of Sea Transportation
Volume 14
Number 2
June 2020
DOI: 10.12716/1001.14.02.23

Use of a Multiple Regression Model to Determine the Parameters of Vessel Traffic Flow in Port Areas

A. Nowy & L. Gucma
Maritime University of Szczecin, Szczecin, Poland

ABSTRACT: The paper presents the method of determining ships traffic stream parameters by means of regression method. The aim of the studies was to determine the correlation between the ship's parameters and the parameters of the fairway. Developing the presented model with information on the position of the vessel's antenna and information on the accuracy of position determination will allow creating a model for predicting the parameters of waterways.

1 INTRODUCTION

One of the basic problems of marine traffic engineering is to determine the optimal parameters of newly built and modernized elements of waterways. Depending on the type of waterway, these parameters may be, for example, the width of the waterway or the diameter of the turning circle. These parameters are usually determined by one of two methods: analytical method or more expensive and a more accurate simulation method. Also the statistical data from computer simulation model have been used to determine waterway parameters [Gucma L. 2005].

With AIS (Automatic Identification System) data accessibility, the input data for the model represents the actual navigator behavior has been received. It helps to better understand the ships movement in the waterway. The characteristics of the ship traffic from the AIS data analysis will be used to generate input parameters.

Nowadays AIS data are used in researches on the actual behavior of vessels. Number of traffic studies have been conducted in last years. A classical traffic flow theory was used in an initially developed mathematical model (Yip, 2013). BP neural network was used to forecast vessel traffic flow (Zhang et al., 2018). The automatic recognition of traffic flow based on kernel density estimation is proposed by (Li et al., 2018). Most studies focus on the determination of traffic parameters and their distribution. However, this work focuses on the use of AIS data to determine the relationship between traffic flow parameters, vessel dimensions and the width of the fairway.

The paper presents studies on traffic flow in Baltic Sea ports as a part of researches on a general mathematical model of vessel traffic streams. The calculations are performed partially with the mathematical software tool IWRAP MK2 recommended by IALA. The statistical analysis was carried out using Stististica10 software.

2 METHOD

2.1 *Analyzed area*

In the study straight waterway section was taken into account. To create a model the fairways of different

width have been chosen. The authors analyzed the movement of ships in the following area (Tab.1).

Table 1. Localization of the analyzed waterway.

Lp.	Localization	*D*- width of the dredged fairway [m]	D_{10m}- width between 10m isobaths [m]
1	Swinoujscie	170	245
2	Police-Raduń	90	132
3	Gdańsk Portowy Kanal	135	165
4	Gdańsk- Martwa Wisła	45	105
5	Gdynia Portowy Kanal	190	194
6	Kołobrzeg	50	50
7	Kaliningrad Approach 338 berth	50	150

Movements in the port are regulated by port regulations. It can affected the maximum speed of the ship or prohibition of any activities due to the bad weather conditions. In that reason for the studies only arrivals of ships with wind ≤10m/s were taken into account. This will also limit the applicability of the model. In the following data analysis, only data samples with incoming ships are studied. A vessel with incoming direction means that the ship comes into the waterway from the open sea.

2.2 *Data*

Researches have been conducted on the basis of data possessed from AIS obtained from Polish Maritime Administration. Vessel traffic was analyzed using data from 2015 to 2017. General cargo vessels (GC) of length L≥50m were considered.

AIS raw data was processed using IWRAP MK2 application. The statistical function can be found using historical AIS data. The traffic patterns are illustrated in a density plot, which helps to identify the location of navigational routes (legs). Making a cross-section of the leg and creating a histogram for each direction the mathematical representation using a number of probability functions is prepared.

AIS data was filtered. Only ships going ahead were considered. For that reason next position of the vessels was checked (1 km ahead). If the position was recorded the ship was included to database. This allowed to select only this group of vessels which actually moved in a given direction. Mooring and circulated vessels were excluded.

It should be added that all the considerations presented were carried out for one-way traffic. Ships were divided into group with the same dimensions (length L and width B) and similar maneuvering characteristics. Usually there were a sister ships.

2.3 *Method*

This paper presents the methods of determining the parameters of traffic flow on straight waterway using a classic model of multiple regression supported by the analysis of residuals. In the model the introduction of hydrometeorological conditions and maneuverability features of ships was omitted. It is obvious that such assumptions considerably simplify the model.

For each waterway center of the traffic lane was established. Crossing-line perpendicular to the channel has been selected to derive the data for the behavior of ship traffic. For all sections, lateral distributions were determined by analyzing the number of ship crossings of report lines. In further step mean and standard deviation of lateral distribution for each section and each group of ships was determined.

The aim of the study is to find a relation between traffic stream parameters and width of the waterway. Multiple regression method was used to build models of mean and standard deviation of ship's distance from the center of the fairway. After implementation the position of the AIS antenna such models can be used to determine the probability of collision of the ship with hydrotechnical structures in the analyzed areas.

2.4 *Multiple regression model*

The model based on multiple regression describes the relationship between the dependent variable y and n independent variables formulated as follows:

$$y = b_0 + b_1 x_1 + \ldots + b_n x_n \quad (1)$$

where:
b_1 - model coefficient

The following parameters of vessel traffic flow such as mean m and standard deviation σ of vessels' position in relation to the center of the track were selected as dependent variables. It is assumed that the center of the track is located symmetrically in relation to the mean width of the dragged waterway D. In the regression model following independent variables were considered:

- width of the ship B [m],
- length of the ship L[m],
- width of the dragged fairway D [m],
- width between 10m isobaths D10m [m]

The basic problem occurring during the building of multiple regression models is the internal correlation between independent variables. In the proposed model it is obvious and occurs between the length and width of ships. Despite the correlation is very strong authors not decided to remove the independent variable the ship's length because it has a theoretical effect on the width of the traffic lane.

A very important independent variable in the model is the width of the fairway. The more difficult (the narrower) area for navigation the more accurate the steering of the vessel is performed. Tolerance for errors is less and the probability of a collision increases. The freedom of maneuver choice is also reduced and only some maneuvering methods are effective and safe. Analyzing the fairway area and draught of the ships authors decided to add variable D_{10m} width between 10m isobaths.

3 RESULTS

3.1 *Parameter estimation*

Models of two dependent variables: mean m and standard deviation σ of vessels' position in relation to the center of the track at a certain level of significance can be defined as:

$$m = b_0 + b_B B + b_D D + b_{D_{10m}} D_{10m} + b_L L \quad (2)$$

$$\sigma = b_0 + b_B B + b_D D + b_{D_{10m}} D_{10m} + b_L L \quad (3)$$

Table 2 presents the multiple regression coefficients of the model obtained by the least squares method. In addition, a coefficient of determination R^2 is presented, which determines the percentage of variation of the dependent variable explained by the model and the standard errors of the estimation s are interpreted as the average deviation of the dependent variable in the sample from the theoretical value. The significance of regression models was studied by means of F statistics.

3.2 *Model of variable m*

Tab. 3 shows parameters for the first model where dependent variable is mean of vessels' position in relation to the center of the track.

Estimating the parameters we obtain the regression function of:

$$m = -24.268 - 0.809 * B - 0.1165 * D + 0.474 * D_{10m} - 0.042 * L \quad (4)$$

Standard estimation errors of parameters are small in the case of the independent variable variables (≈ 0.03 for D and D_{10m}; ≈ 0.38 for B; ≈ 0.06 for L) for and the variable and acceptable in the case of intercept (≈ 4.00). The significance of the whole variable model is p<0.000.

To verify the statistical significance of one variable, the t-Student test was performed. The test is designed to determine whether an explanatory variable has a significant effect on a dependent variable. In model "mean" only variable L is not statistically significant. However, it was used in a model to determine changes in the mean position of vessels from the center of the track due to the length of the vessel. Standard error of estimate equal s=14.067. This means that the predicted values of the dependent variable differ from the empirical values on average by 14.067%. The equation (4) can therefore be written as:

$$m = -24.268 - 0.809 * B - 0.1165 * D + 0.474 * D_{10m} - 0.042 * L \pm 14.067 \quad (5)$$

Table 2. Coefficients of multiple regression model.

Dependent variables	b_0	b_B	b_L	b_D	$b_{D_{10m}}$	R^2	s [m]	Significance of regression	p
m	-24.2679	-0.8088	-0.0420	-0.1165	0.4741	0.6433	14.067	F=138.46	0.000
σ	12.4513	-0.0271	-0.0112	0.1264	-0.043	0.5285	4.1523	F=88.156	0.000

Table 3. Regression summary for dependent variable: Mean m

	R= .80209949 R2= .64336358 Adjusted R2= .63871686 F(4,307)=138.46 p< 0.0000. Std.Error of estimate: 14.067					
	b*	Std.Err.of b*	b	Std.Err.of b	t(307)	p-value
Intercept			-24.2679	4.005164	-6.05915	0.000000
L[m]	-0.074105	0.097887	-0.0420	0.055484	-0.75705	0.449600
B[m]	-0.208607	0.098607	-0.8088	0.382290	-2.11555	0.035188
D_{10m}	0.943881	0.064613	0.4741	0.032456	14.608250	0.000000
D	-0.242209	0.066555	-0.1165	0.032013	-3.63925	0.000321

Table 4. Regression summary for dependent variable: Std.Dev. σ

	R= .73115299 R2= .53458470 Adjusted R2= .52852065 F(4,307)=88.156; p<0.0000 Std.Error of estimate: 4.1523					
	b*	Std.Err.of b*	b	Std.Err.of b	t(307)	p-value
Intercept			12.45138	1.182242	10.53200	0.000000
L[m]	-0.076701	0.111823	-0.01123	0.016378	-0.68591	0.493285
B[m]	-0.270152	0.112645	-0.27063	0.112844	-2.39825	0.017070
D	1.016896	0.076030	0.12639	0.009450	13.374890	0.000000
D_{10m}	-0.335100	0.073812	-0.04349	0.009580	-4.53992	0.000008

3.3 *Model of variable* σ

For the second model where standard deviation of vessels distance from the center is obtained regression function is as follows:

$$\sigma = 12.451 - 0.271 * B + 0.126 * D - 0.043 * D_{10m} - 0.011 * L \quad (6)$$

Tab.4 shows the results for the model of standard deviation. As in the first model standard estimation errors of parameters are small in the case of the independent variable variables (≈ 0.001 for D and D_{10m}; ≈ 0.11 for B; ≈ 0.02 for L) for and the variable and acceptable in the case of intercept (≈ 1.18). The significance of the whole variable model is p<0.000. Again, variable L hasn't got any effect on the model (p=0.49). Standard error of estimate equal s=4.1523. This means that the predicted values of the dependent variable differ from the empirical values on average by 4.1523%. The equation (6) can therefore be written as:

$$\sigma = 12.451 - 0.271 * B + 0.126 * D - 0.043 * D_{10m} - 0.011 * L \pm 4.15 \quad (7)$$

3.4 *Verification of the model*

The statistical validity of the model was tested with the use of several indicators. The first one is the determination coefficient $R2$. The coefficient R^2 for the first model is 0.6387 which means that the model explains 64% of the variability of the response data around its mean. The second model explains 53% (R^2=0.5285). The coefficient of determination R^2 is satisfactory according to the accepted interpretation which leads to further researches on the topic. The coefficient of determination can be low cause the model actually predicts navigator behavior. Humans are harder to predict that for example physical process. It should be remembered that we are not always able to achieve the very high value of the coefficient R^2. The aim of the evaluation of the existing model is not to obtain the highest possible level of R^2, but to determine a relationship between the consider variables and reliable parameter assessments.

In order to obtain a reasonably correct regression model, the obtained residuals values must always be analyzed after estimation and verification of this model. The analysis of residual according to [1], should begin from the most important matter i.e. checking presumptions of the classic method of the smallest squares. This is because the correctly constructed model is characterized by certain desirable properties of the residuals (such as normality, constancy of variance, lack of the autocorrelation).

3.5 *Normality of residuals values*

In order to obtain normality checking of regression model, graph of residual normality was created (Fig.1, Fig 2). It enables a visual examination of residuals compliance with normal distribution. If points are situated along the straight line that confirm the normality of residual distribution. Some objection can relate to the first and lasts observations, because it is a bit off from the line, but this distance has not influenced significantly the normality of residuals values. The same information give as the histogram of residuals (Fig.3, Fig.4). It can be noticed that this is a good situation because the normal line (red line on the graph) crosses the column upper edge centers (especially for second model).

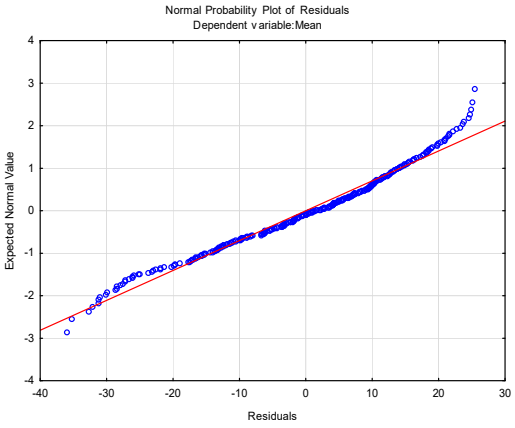

Figure 1. Normality graph of residuals values for „Mean"

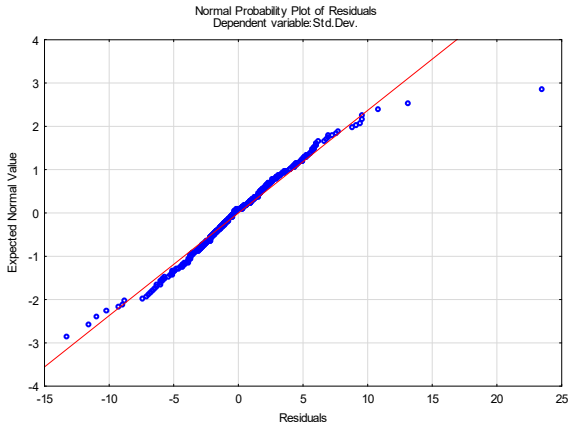

Figure 2. Normality graph of residuals values for „Standard deviation"

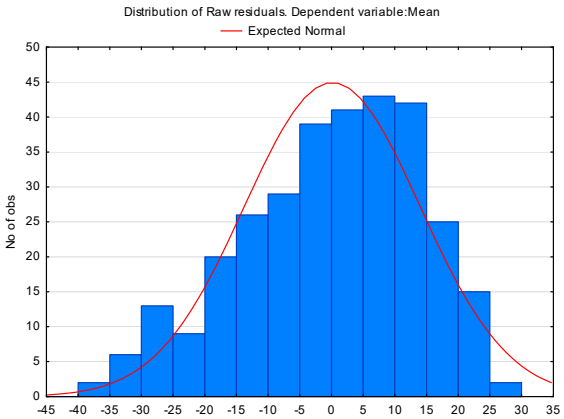

Figure 3. Histogram of residuals for "Mean"

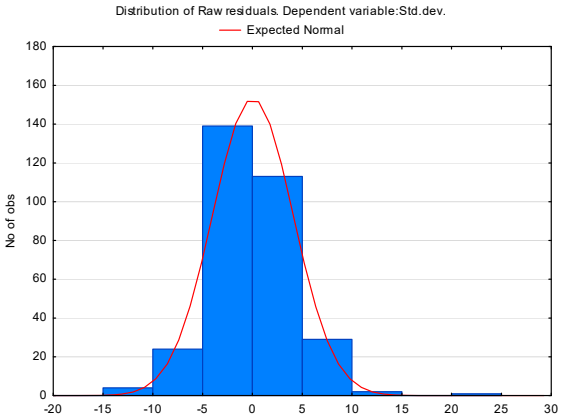

Figure 4. Histogram of residuals for "Standard deviation"

3.6 *Autocorrelation of the residual values*

The assumption of autocorrelation was not verified due to the fact that observations are not ordered.

3.7 *The randomness test*

The randomness test is designed to examine the correctness of the analytical form of the model. This can be obtained by means of both a visual assessment of the distribution of residuals and statistical tests. In this paper authors decided to use a firs method. If the residuals of the model fulfill the assumption of randomness, then in the graph the residuals as observed values (for both the explanatory variables and the explained variable) should be arranged at random and should not show any regularity (e.g. subsequent series of positive and negative residuals).

In Fig.5 and Fig. 6 the residuals of the model diagram in relation to the empirical values of the explanatory variable is shown. The residuals are distributed irregularly, so we can assume that the assumption of randomness is fulfilled. It can be noticed that for mean *m* there is a lack of observation in range 30-50m. This is the results that need to be studied in the further researches. The same effect can be seen on Fig.7.

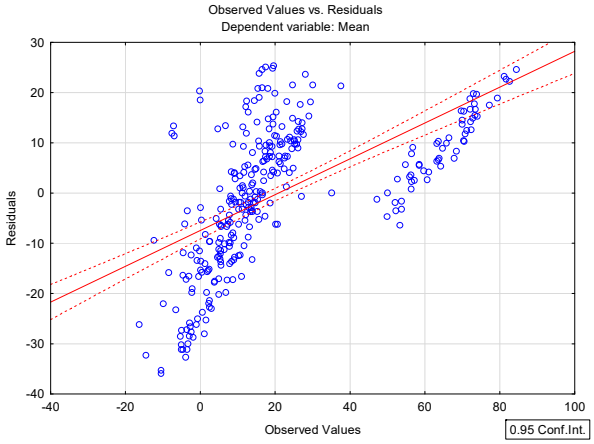

Figure 5. Residuals distribution in relation to observed values for variable "Mean"

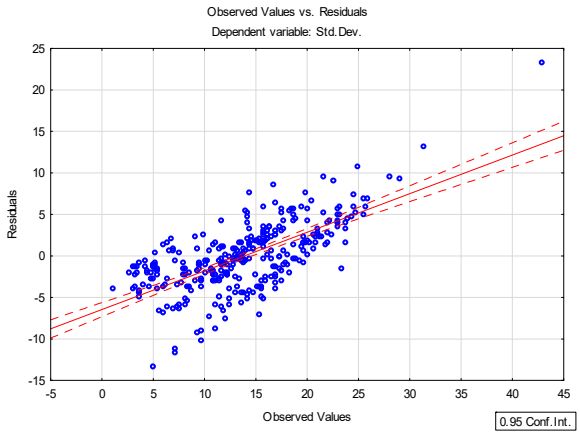

Figure 6. Residuals distribution in relation to observed values for variable "Std.Dev."

3.8 *Stability of residuals values variance*

The next desirable property of residuals values is a presumption about homoscedasticity of random component. For this purpose, a visual evaluation of the distribution of residuals in relation to predicted (theoretical) values was applied. The regular distribution of points on the residue scatterplot in relation to the predicted values (Fig.7, Fig.8) do not confirmed categorically the homoscedasticity of the variance of the random component. Further tests should be carried out (e.g. Goldfeld- Quandt test). The existence of heteroscedasticity does not always mean a bad choice of model or poor quality of statistical data. In that reason the model was not modified on this stage of the researches.

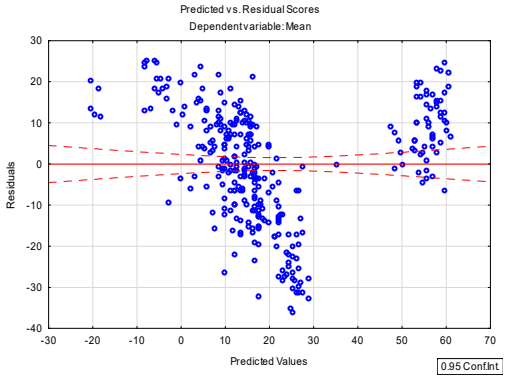

Figure 7. Residuals distribution in relations to predicted values for variable "Mean".

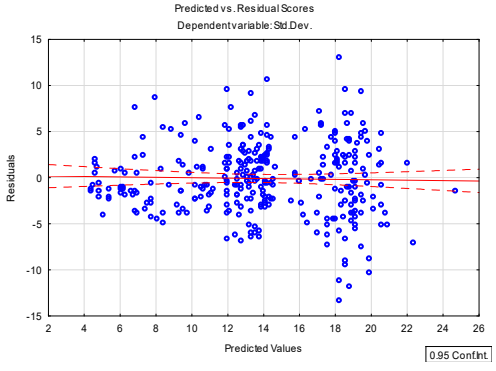

Figure 8. Residuals distribution in relations to predicted values for variable "Std.Dev."

3.9 *Atypical observation in regress analysis*

After adapting the regression equation on the basis of the observation results, it is always necessary to analyze the predicted values and residuals. In regression analysis, it is important that the model is not determined excessively by individual observations with values significantly different from those typical for a given sample. Such deviating values can significantly disturb the calculation results and lead to incorrect conclusions. Sometimes this one observation has to be deleted to prevent such case. However, observations that do not match to the model may indicate deficiencies in the model or a bad algebraic form of the model.

In order to detect such outliers graph of residuals distribution in relation to deleted residuals was generated (Fig.9 and Fig.10). It can be noticed that there is no coming off observation. It can be observed that there are some observations that can be removed after statistical analysis and identification of the source of this effect. In the presented models, however, no observations have been removed. In addition, it was noticed that the sample removal of some outliers did not have a significant impact on the quality of the examined models.

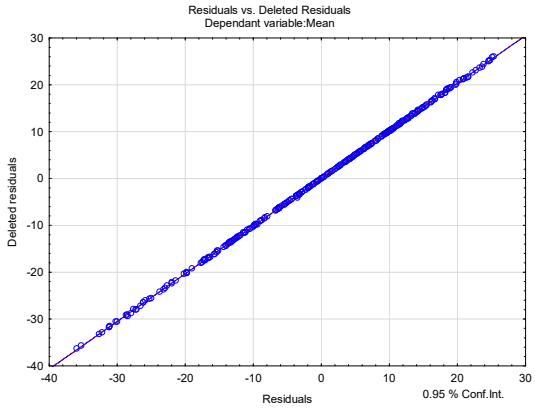

Figure 9. Residuals distribution in relations to deleted residuals for variable "Mean".

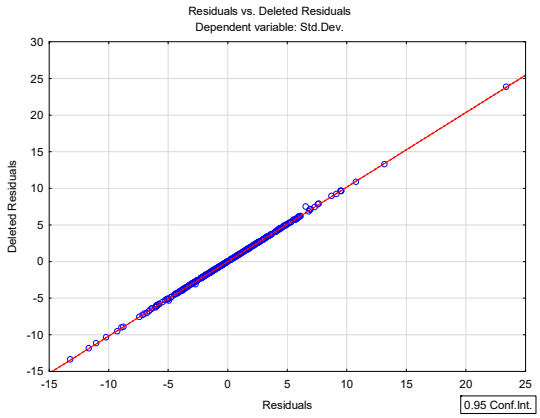

Figure 10. Residuals distribution in relations to deleted residuals for variable "Std.dev.".

3.10 *Prediction based on the regression model*

During the regression model building the possibility of prediction of variable values is taken into account i.e. what values will be assumed by a dependent variable with different values of an independent variable. The final stage of regression analysis is to use a verified regression model for prediction of a dependent variable. A graphical representation of the scatterplot can be used. Figure 11 shows the observed and predicted values of the mean position of the vessel's distance from center with a prediction interval of 95%. The limits of the prediction interval are shown with a dashed line. Fig. 12 shows the graph of observed and predicted values of standard deviations of the ship's distance from the center of the track on straight sections. The decrease of variance with the increase of mean as well as increase of variance with the increase of standard deviations should be noticed.

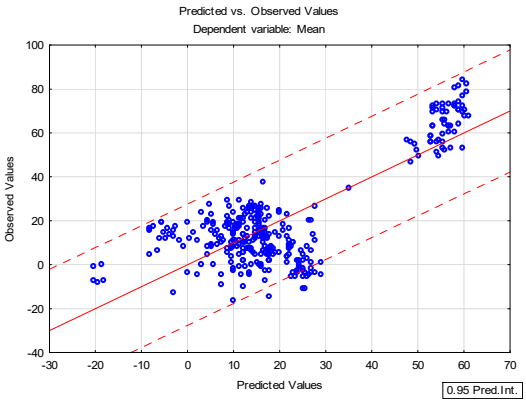

Figure 11. Comparison of predicted values of mean of the ship's distance from the center of the track using a multiple regression model and the observed values in straight sections.

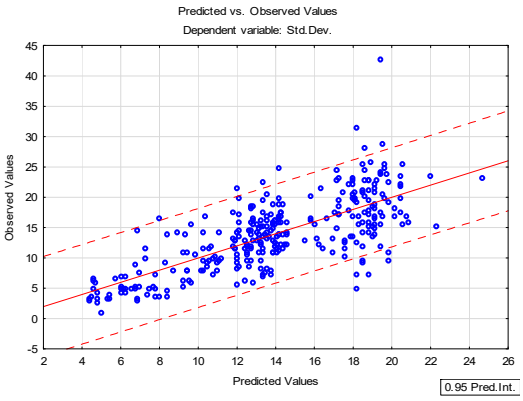

Figure 12 Comparison of predicted values of standard deviations of the ship's distance from the center of the track using a multiple regression model and the observed values in straight sections.

4 CONCLUSIONS

Multiple regression is used in prediction, i.e. determination of future values of a dependent variable on the basis of the equation. Used independent variables indicated a significant impact on the model which only confirmed the assumptions that with the increase in width of vessel B and length L, the mean and standard deviation of the vessel's position in relation to the center of the track

decreases. However, the larger the available width of the water area, these values increase. The aim of the studies was to build a model which describes the above mentioned dependencies in detail.

Despite the fact that the information from the AIS system, which was used to build the model, covered the whole range of variables L and B, the variables *m* and σ showed deficiencies in data continuity. From the histogram it can be seen that the values of mean and standard deviation of the position of vessels in relation to the center of the track in the range from 30 to 50 m practically do not occur. It is probably necessary to take a larger sample for tests of other waterway width D and D10m. Lack of data can be seen in the residuals plot. Therefore, further analyses should be carried out taking into account other fairways.

The presented models are based on AIS data. The position of the vessel in relation to the center of the track refers to the position of the antenna. Taking into account the position of ship's starboard and port extremities and the angle of drift, it will be possible to build a model allowing to determine the mean width of the safe maneuvering area of the ship. However, on the basis of the built regression models it is already possible to forecast the parameters of the vessel traffic flow in port areas. Further research should also take into account weather conditions and analyze the accuracy of the position obtained from the AIS system.

It is planned to build a model taking into account all relevant factors (including hydrometeorological conditions and maneuverability of the ship). Further work will also focus on the construction of regression models for different types of waterways such as port entrances and bends.

BIBLIOGRAPHY

[1] Gucma L. (2005), Modelowanie czynników ryzyka zderzenia jednostek pływających z konstrukcjami portowymi i pełnomorskimi. Wydawnictwo Naukowe Akademii Morskiej w Szczecinie.

[2] Li, Wei- Feng; Mei, Bin; Shi, Guo-You (2018): Automatic recognition of marine traffic flow regions based on Kernel Density Estimation. Journal of Marine Science and Technology 26, pp. 84–91.

[3] Stanisz A. (2007), *Przystępny kurs statystyki.* Statsoft Polska, Kraków 2007 r.

[4] Yip, T.L. (2013), A marine traffic flow model. TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation 7, 1, pp. 109–113.

[5] Zhang, Ze-guo; Yin, Jian-chuan; Wang, Ni-ni; Hui, Zi-gang (2018): Vessel traffic flow analysis and prediction by an improved PSO-BP mechanism based on AIS data. in: Evolving Systems (2018). https://doi.org/10.1007/s12530-018-9243-y.