Forecasting Road Transport Demand with Use Machine Learning

Artur Budzyński
*Silesian University
of Technology*
Katowice, Poland
artur.budzynski@polsl.pl

Aleksander Sładkowski
*Silesian University
of Technology*
Katowice, Poland
aleksander.sladkowski@polsl.pl

Abstract—This article deals with the topic of predicting the demand for road transport. Proposed is new term road transport demand basing on relation available freights and trucks on transport exchange. Data from Central Statistical Office in Poland and Transport Exchange are compared. Processing data for machine learning models are shown. 32 machine learning models are compared according to 3 success metrics. Implementation possibilities are presented. Models are shared in open access.

Keywords—forecasting, road transport demand, machine learning

I. INTRODUCTION

Knowing the dependencies that exist in the free road transport market helps to make better decisions in fleet management [1]. Knowledge what will be demand for road transport in future is beneficial. Demand forecast can be used as relevant data in forecasting prices for road transport services. Example [2] of road traffic forecasting was based on the correlation of the GDP coefficient and tonne-kilometres made by vehicles. In the publication [3], forecasting was based on the data of the Central Statistical Office in Poland, using an additive and multiplicative models. The paper [4] presents predictions the quantity of transported goods using 4 models: Holt-Winters double exponential smoothing, Holt-Winters double exponential smoothing assisted by artificial immune system, Bayesian networks and Bayesian networks with unemployment rate. The work [5] presents methods of forecasting the volume of road transport using the following methods: the seasonal Winter method, harmonic analysis and harmonic analysis supported by the artificial immune system. Scientists in [6] used Long-Short Term Memory Neural Networks (LSTM-NN) to predict the demand for transport during hurricanes "Irma" and "Matthew".

A. Useful Software

Python is very useful programming language for building machine learning models. It is attractive for exploratory data analysis and for the development of algorithms. A Jupyter Notebook [7] was used to write the code. It's a tool that allows you to write code and comments in your browser. It is convenient for project and document management. It also allows you to freely share the code like in this publication. There are many useful libraries for Python. Scikit-Learn [8] is a library that allows the implementation of many machine learning algorithms. The Pandas library [9] is useful for working with data sets. It enables data processing and creation of new features. The NumPy library [10] is useful for performing numerical calculations. Matplotlib [11] is a good library for data visualization. MapChart [12] allows you to create very precise data visualization in the form of maps. It is also possible to save the map configuration. This makes it easier to use in the future to achieve further goals. GitHub [13] is a platform that enables code sharing. It is beneficial for the advancement of science. Possible is easily create own version of a shared project or suggest changes. The code [14] for this publication is also available on the platform.

B. Machine Learning Models

32 machine learning models [15] were included in the study, divided into:

1) Linear Regression Models. The methods predict a value to be a linear combination of features.

a. Linear Classical Regression: LinearRegression, Ridge, RidgeCV. An example of an application is [16], where a model was used to predict the effort in a software development project.

b. Linear Regression with variable selection: Lasso, LassoCV, Lars, LarsCV, LassoLars, LassoLarsAI, ElasticNet, ElasticNetCV, OrthogonalMatchingPursuit, OrthogonalMatchingPursuitCV. Models have built-in variable selection fitting procedures.

c. Bayesian Linear Regression: BayesianRidge, ARDRegression. Bayesian regression techniques can be used to include regularization parameters in the estimation procedure.

d. Outlier-robust Linear Regression: TheilSenRegressor. A good model for corrupt data - either outliers or errors in the model.

e. Generalized Linear Models (GLM): TweedieRegressor. The models allow response variables to have a different error distribution than the normal distribution.

f. Miscellaneous Linear Regression: PassiveAggressiveRegressor. Models for large-scale learning.

2) Kernel Ridge Regression: KernelRidge. Models combines Ridge regression and classification.

3) Gaussian Processes: GaussianProcessRegressor. Method for generic supervised learning

4) Neural Network: MLPRegressor. Examples of the use of artificial neural networks in transport are automatic docking of ships [17], prediction of fuel consumption in ships [18] and price prediction for road transport [19].

5) Dummy Estimators: DummyRegressor. The model is based on a simple algorithm. It cannot be used for real problems. It is applicable as a model for comparison. If the model has a worse result than DummyRegressor then the prediction is useless.

6) Nearest Neighbors: KNeighborsRegressor. The model predicts value from the most similar example he has learned.

7) Support Vector Machines: SVR, LinearSVR, NuSVR. Models depends only on a subset of the training data.

8) Ensemble Methods: AdaBoostRegressor, BaggingRegressor, ExtraTreesRegressor, GradientBoostingRegressor, RandomForestRegressor, HistGradientBoostingRegressor. Models combined with multiple estimators.

C. Metrics

It was decided to evaluate the models according to 3 success metrics. Smaller error is better model.

- Equation (1) describes Mean Square Error (MSE)

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2 \qquad (1)$$

where: $\hat{y}_t$ - predicted value of the dependent variable; y_t - actual value of the dependent variable.

- Equation (2) describes Root Mean Square Error (MSE). Smaller error is better model.

$$RMSE = \sqrt{MSE} \qquad (2)$$

- Equation (3) describes Coefficient of determination R2. It ranges from 0 to 1. Closer 1 is better model.

$$R^2 = \frac{SS_M}{SS_T} = \frac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \qquad (3)$$

II. DEMAND PREDICTION

It was decided to build demand prediction models based on data from one of the most important transport exchanges TimoCom [20]. Transport exchanges are one of the most promising solutions for transport logistics [21]. The data the barometer takes into account is not complicated. Equation (4) describes it as simply a percentage ratio reported cargos and vehicles. No scientific papers on predictions using data from freight exchanges were found.

$$D = QL/QT \qquad (4)$$

where: D – demand; QL – quantity loads, QT – quantity trucks

A. Bilateral Relation Demand Transport Poland and other countries

The data is for 2019 and comes from the TimoCom barometer. It was decided to use a bilateral visualization defining the demand for exports from Poland in relation to imports. Data concern relations with 44 countries from Europe and Asia. The method of calculating the coefficient is presented in the pattern (5).

$$RY = \frac{\sum DE_n = (DE_1 + DE_2 + \cdots + DE_{12})}{\sum DI_n = (DI_1 + DI_2 + \cdots + DI_{12})} \qquad (5)$$

where: RY – ratio year; DE – demand export; DI – demand import; n – month number

Fig. 1 shows the export-import relationship. The color green shows the countries to which the demand for the transport service is greater in export than in import, and the red is greater demand for import. The greatest advantage of demand in exports is in relation to the Balkan countries: Croatia, Bosnia and Herzegovina, and especially to Montenegro. This may be due to the political situation in Montenegro. The biggest advantage of demand in import occurs in the countries of: Asia: Azerbaijan, Armenia and Georgia. A particularly high disproportion is in relation to Luxembourg. This may be due to the small area of the country and the fact that it is very economically developed.
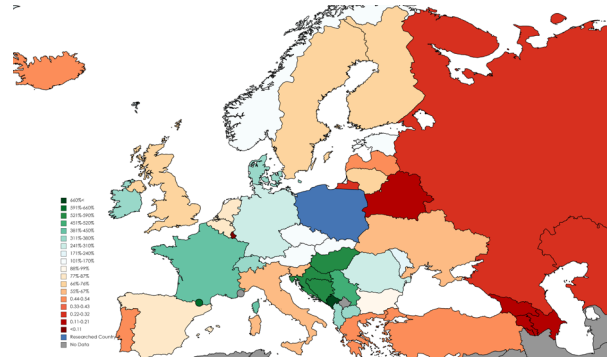

Fig. 1. Map of the relation of demand export to import Poland and other countries in 2019 based on barometer TimoCom.

Fig. 2 shows analogous data provided by the Central Statistical Office in Poland. The differences in the data presented may result from the fact that many goods are transported without the use of freight exchanges using permanent contracts.
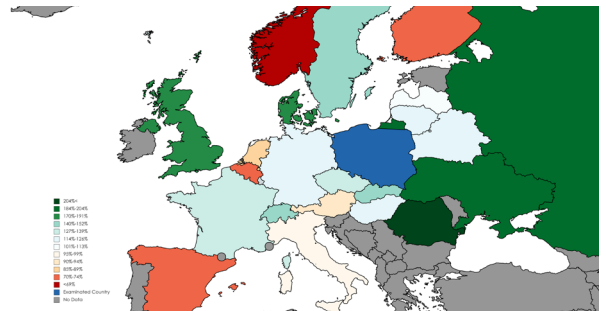

Fig. 2. Map of the relation of demand export to import Poland and other countries in 2019 based on data from Central Statistical Office in Poland

It is evident that in some relationships these data correlate more and in others less. The R-Person's correlation calculated from the pattern (6) is approximately 0.22.

$$r(x, y) = \frac{cov(x,y)}{\sigma x * \sigma y} \qquad (6)$$

where: r(x,y) – Pearson correlation coefficient; cov – coviarance; σ – standard deviation of the population

The correlation is positive which is expected. Figure 3 shows the breakdown of the demand based on the data from the barometer in relation to the actually transported goods in 2019. Perfectly even distribution is 1. In this case, it would mean that in 2019 the ratio of demand on the stock exchange coincides with the actually transported goods.

The most similar results were found in relation to Spain, Belgium, Finland, the Czech Republic, the Netherlands and

INPUTS
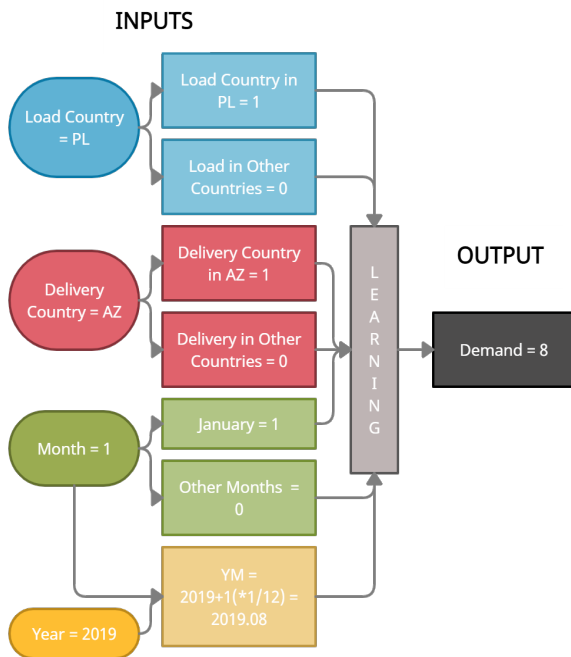
Romania. The greatest disproportion in favor of the demand on the stock exchange occurred in relation to Hungary. This proves that much more is offered on the stock exchange than it is actually transported. The greatest disproportion in favor of the actual transported goods occurred in relation to Russia. This proves that more loads are transported without the exchange. This method of ordering may take place through standing contracts. The situation is similar in the case of Belarus and Ukraine. The reason for this may be the difficult access to the market due to the required permits.

Fig. 3. Ratio demand on transport exchange and real transported goods in the 2019 year.

B. Data Preparing

Data was collected from publicly available data made available on the Internet by the Timo-Com barometer. Data concern 2019 and relations between Poland and 44 countries from Europe and Asia. Considering that a year has 12 months and each relation is bilateral, this equals 1056 rows with data. Calculations are made in Jupyter Notebook using the Scikit-Learn library. Fig. 4 shows a model diagram on the example of the demand for road transport services from Poland to Azerbaijan in January 2019. The diagram was prepared using the creately.com online wizard.

Fig. 4. Model scheme

C. Training and Testing

The collected data was divided into: 70% training data and 30% test data. Fig. 5 shows the results of testing according the RMSE metric. The chart on the MSE metric is not presented because it is a square RMSE metric. RMSE is better at showing errors than MSE.
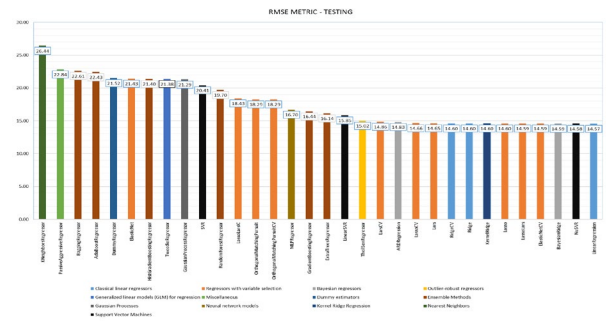

Fig. 5. Testing results according RMSE metric

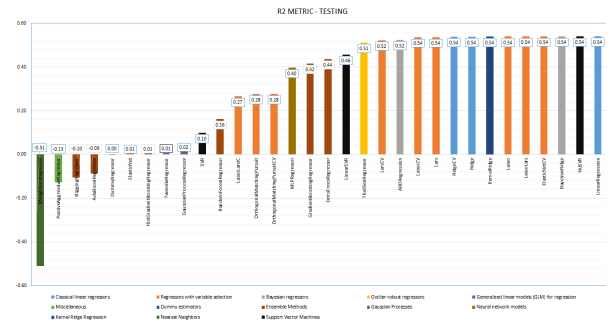Fig. 6 shows the results of testing according the R2 metric.

Fig. 6. Testing results according R2 metric

Fig. 7 shows the relationship between the forecast and real values using the TheilSenRegressor as an example.
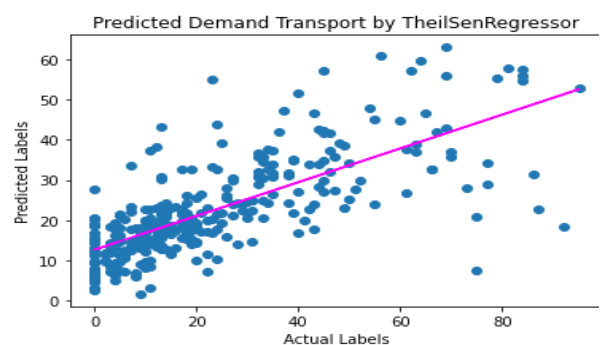

Fig. 7. Relationships between forecasted and real values on the example of TheilSenRegressor.

D. Validation

The form of validation data is the same as for training and testing. The difference is that the models are validated in the next year 2020. There were some changes in the market compared to the previous year. Fig. 8 shows these changes. The increase in exports in relation to imports with a given country is marked in green. The reduction is marked in red.

The more intense the color, the greater the change. This is described in detail in the legend
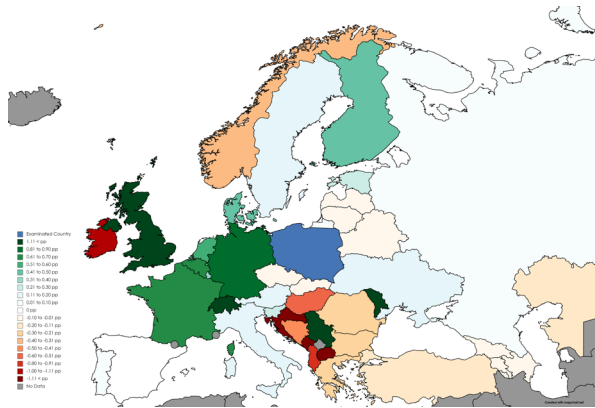

Fig. 8. Changes in the demand-supply relationship between 2019 and 2020 in percentage points

Fig. 9 shows the results of the validation according to the RMSE metric. The validation according to the MSE metric is not presented, because this is the square of the RMSE metric. RMSE metric better capture errors.
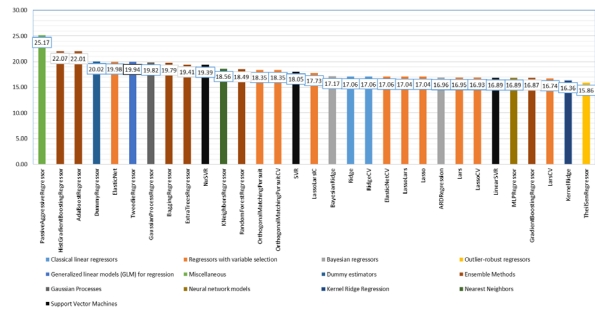

Fig. 9. Validation according RMSE metric

Fig. 10 shows the results of the validation according to the R2 metric. The model that performs best by all metrics is TheilsenRegressor. The LinearRegression model has been omitted from the graphs because it showed values with a very large error.
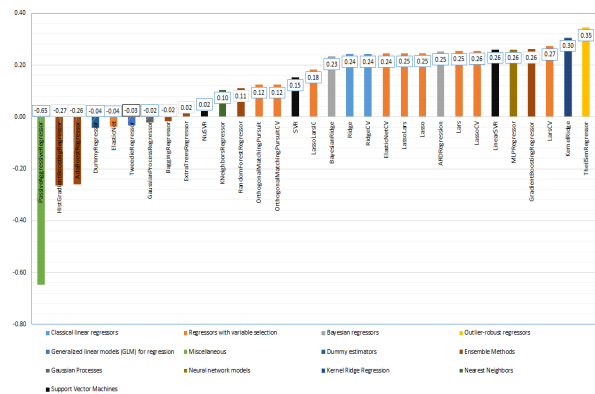

Fig. 10. Validation according R2 metric

III. IMPLEMENTATION

The model can be used to predict the demand. Fig. 11 shows an example of an application for demand prediction in relation to Poland to Azerbaijan.
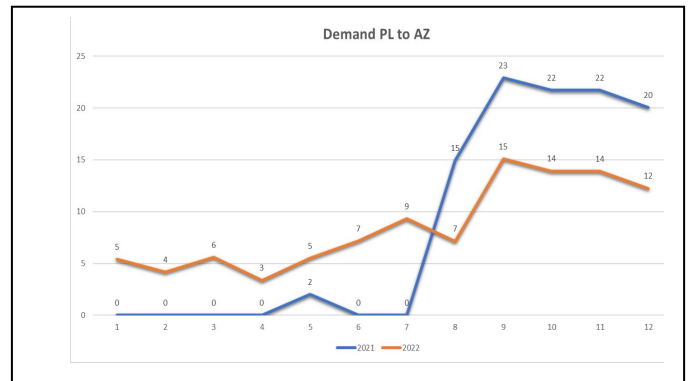

Fig. 11. Demand forecasting in relation from Poland to Azerbaijan

Fig. 12 shows a prediction in relation from Azerbaijan to Poland. Similarly, you can plot different relationships.
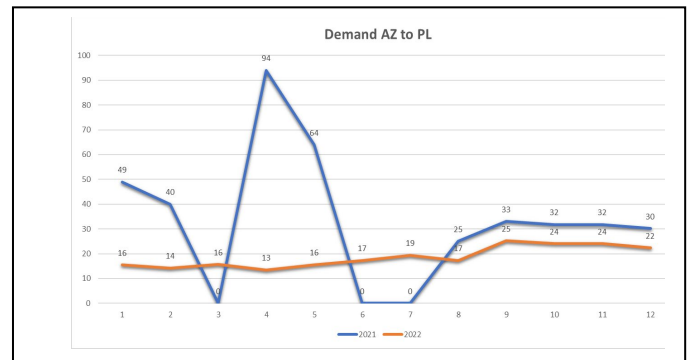

Fig. 12. Demand forecasting in relation from Azerbaijan to Poland

The introduction to this article covered the subject of the software language, tools, and software used. Creating the right environment to carry out such a project on your own is quite a complicated process. The solution to this problem is to use the ready-made Anaconda package. It should be downloaded from the manufacturer's website. The software is available for the most popular operating systems: Windows, Linux, and macOS. Once downloaded, the user gets access to Jupyter Notebook, which is a very convenient Python programming tool. All necessary libraries are also configured: Scikit-Learn, Pandas, and NumPy. The next step is to create your own "csv" file that will be used to train the models. Then download the ready code from the project's website and run it on your computer in Jupyter Notebook. In case of problems with implementation, you can use the "Step by Step" guide also on the project website. There, each of the steps above is presented with a detailed description of each click and a visualization of what it looks like in Windows. The tutorial is adapted to a person who had no contact with programming before.

IV. CONCLUSIONS

The transport barometer is an indicator that can be used to predict demand. Collecting and presenting data on the number of submitted offers is a favorable phenomenon for analyzes. Machine learning can be used to predict demand based on data from freight exchanges. The most effective model is TheilsenRegressor. The model presented in the text can be used by any company that uses road transport. Among them, the following can be distinguished: transport exchanges, transport, trade, production, and forwarding

companies. The presented model can be used as a base or act independently. Demand prediction models presented in this text can be used in road transport price-prediction models. The implementation of forecasting into the transport barometer may make the exchange's offer more attractive to its clients. All data used to present the research results in this publication are available on the project website [14]. The project page contains raw data, data processing methodology, training, and testing, validation, charts, maps, map configuration. All this creates favorable conditions for repeating this experiment. The process of building models for similar problems in future research is also facilitated.

REFERENCES

[1] A. Budzyński, "Use dependencies between freight prices and economic factor as a solution in improve efficiency work in road transport," in Proc. XII Int. Sci. Conf. & IX Int. Symposium of Young Researches „Transport Problems'2020", 2020.

[2] A.C. McKinnon and A. Woodburn, "Logistical restructuring and road freight traffic growth," Transportation 23, 1996, pp. 141–161.

[3] E. Macioszek and A. Kurek, „International road cargo transport in Poland and other EU countries," Scientific Journal of Silesian University of Technology. Series Transport 111, 2021, pp. 99–108.

[4] B. Mrowczynska, M. Ciesla, A. Krol and A. Sladkowski, "Application of Artificial Intelligence in Prediction of Road Freight Transportation", PROMET, vol. 29, no. 4, pp. 363-370, Aug. 2017.

[5] B. Mrowczyńska, K. Lachacz, T. Haniszewski and A. Sladkowski, "Acomparison of forecasting the results of road transportation needs.Transport," 2012.

[6] K. C. Roy, S. Hasan and A. Culotta and N. Eluru, "Predicting traffic demand during hurricane evacuation using Real-time data from transportation systems and social media," Transportation Research Part C: Emerging Technologies, Volume 131, 2021.

[7] T. Kluyver, B. Ragan-Kelley, F. Perez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila and S. Abdalla, "Jupyter notebooks – a publishing format for reproducible compu-tational workflows," International Conference on Electronic Publishing, 2016.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay and G. Louppe, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, 2012.

[9] W. McKinney, "Data structures for statistical computing in python," Prroceedings of the 9th Python in Science Conference, 2010, pp. 56–61.

[10] S. van der Walt, S. C. Colbert and G. Varoquaux, "The NumPy Array: A Structure for Efficient Numerical Computation," in Computing in Science & Engineering, vol. 13, no. 2, 2011, pp. 22-30.

[11] J. D. Hunter, "Matplotlib: A 2d graphics environment," Computing in Science & Engineering, 9(3), 2007, pp. 90–95.

[12] R. E. Voorrips, "MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs.," Journal of Heredity, 93(1), 2002, pp. 77–78.

[13] L. Dabbish, H. Stuart, J. Tsay and J. Herbsleb, "Social coding in github: Transparencyand collaboration in an open software repository," Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW, 2012, pp. 1277–1286.

[14] A. Budzyński, "BudzynskiA/DRTPS: Demand Road Transport Predicting system," GitHub. [Online]. Available: https://github.com/BudzynskiA/DRTPS. [Accessed: 27-Aug-2021].

[15] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt, G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," CoRR, 2013.

[16] M. Sadiq, A. Ali, S. U. Ullah, S. Khan and Q. Alam, "Prediction of softwareproject effort using linear regression model," International Journal of Information and ElectronicsEngineering, 2013.

[17] Y. Shuai, G. Li, X. Cheng, R. Skulstad, J. Xu, H. Liu and H. Zhang, "An efficient neural-network based approach to automatic ship docking," Ocean Engineering, Volume 191, 2019.

[18] E. Bal Beşikçi, O. Arslan, O. Turan and A.I. Ölçer, "An artificial neural network based decision support system for energy efficient ship operations," Computers & Operations Research, Volume 66, 2016, pp. 393-401.

[19] A. Budzyński and A. Sładkowski, in Polish "Predicting freight price in road transport with using artificial neural networks," in Proc. XIII Int. Sci. Conf. & X Int. Symposium of Young Researches „Transport Problems'2021, 2021, pp. 137-151.

[20] A Budzyński, in Polish: "Comparison the most important transport exchange in Europe," B.S. Thesis, Silesian University of Technology, Katowice, 2018.

[21] A. Sładkowski and A. Budzyński, in Russian "Transport exchanges, as one of the promising solutions for the problems of transport logistics," in Proc. of the IV Polish - Georgian Scientific-Technical Conference "Transport Bridge Europe - Asia", 2018.