Electronic Research Archive

ERA, 30 (9): 3452–3462.
DOI: 10.3934/era.2022176
Received: 14 June 2022
Revised: 05 July 2022
Accepted: 05 July 2022
Published: 22 July 2022

http://www.aimspress.com/journal/ERA

Research article

An innovative approach of determining the sample data size for machine learning models: a case study on health and safety management for infrastructure workers

Haoqing Wang[1], Wen Yi[2,*] and Yannick Liu[1]

[1] Faculty of Business, The Hong Kong Polytechnic University, Hung Hom, Hong Kong
[2] Department of Building and Real Estate, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

* **Correspondence:** Email: wenyi0906@gmail.com.

Abstract: Numerical experiment is an essential part of academic studies in the field of transportation management. Using the appropriate sample size to conduct experiments can save both the data collecting cost and computing time. However, few studies have paid attention to determining the sample size. In this research, we use four typical regression models in machine learning and a dataset from transport infrastructure workers to explore the appropriate sample size. By observing 12 learning curves, we conclude that a sample size of 250 can balance model performance with the cost of data collection. Our study can provide a reference when deciding on the sample size to collect in advance.

Keywords: transportation infrastructure; learning curve; sample size; machine learning; health and safety management

1. Introduction

Conducting experiments is an essential part of academic research in the field of transportation and construction management (see [1–3]). Scholars need to use samples to conduct analysis or validate models. Especially with the application of machine learning algorithms [4,5], an increasing number of studies have to collect data to test the used algorithms. Choosing the sample size of the experiment is an essential first step and a vital decision problem. If the sample size is too small, the results of

experiments will be not convincing. As the data collection process takes time and money, it is not necessary to collect too many samples. Thus, determining the appropriate sample size is worth investigating.

Hu et al. [6] explore the minimum training sample size for Bayesian network and find that Bayesian network is not sensitive to the training sample size. Ma et al. [7] pay attention to the influence of sample size on establishing reference intervals in clinical medicine. They report that the reference intervals are more consistent when the sample size is greater than or equal to 2000. Burmeiste and Aitken [8] also pay attention to the sample size problem in the field of clinical medicine and they believe this research topic is both clinically and statistically meaningful. Cui and Gong [9] use six regression models to study the effect of sample size when predicting personalized behavior. They find that the prediction accuracy increases with sample size. Taherdoost [10] solves the problem of survey sample size calculation in the social sciences. Lakens [11] propose six factors to determine the appropriate sample size in empirical study. With the widespread application of machine learning methods, which need to be experimentally verified and evaluated, the choice of sample size has become a problem that cannot be ignored. Specifically, there are many data-driven and algorithm-related studies in the field of transportation and construction management. In autonomous vehicle control, historical trajectories of vehicles are used to predict the future trajectories of vehicles (see [12–15]). In ship fuel consumption management, data on ships' static factors, voyage-dependent factors, weather information, and fuel consumption rate are used (see [16,17]). In ship inspection planning, data on ships' age, flag, and historical inspection results are used to developed machine learning models [18]. In health and safety management of construction workers, data on workers' physiological factors, environmental factors, work-related factors, and workers' fatigue are used [19]; Li et al. [20] use a dataset including 589 images of rebar to adopt deep learning in rebar counting. Shehadeh et al. [21] apply three machine algorithms to predict the residual value of heavy construction equipment. However, these mentioned studies in the field of transportation and construction management do not discuss how to determine the sample size used in their experiments.

The existing literature related to sample size are mainly in clinical medicine research and survey research. To the best of our knowledge, there are no studies focusing on sample size determination in transportation management or construction management. Therefore, our research chooses four typical regression models in machine learning—multiple linear regression, ridge regression, LASSO regression and support vector regression—to study the optimal sample size for machine learning models in health and safety management for transport infrastructure construction workers. Our research contributes at both theoretical and practical levels. First, the methodology used in our study can provide scholars and practitioners with a reference to determine the number of samples to collect. Even if they do not use the four methods tested in our experiments, the framework for starting with small sample sizes and increasing incrementally can still provide a reference to choose an appropriate sample size. Second, by observing learning curves, we give a suggestion of sample size for machine learning models using a case in health and safety management for urban transportation infrastructure workers. Studies in similar fields can directly apply our results.

We note that the collection of data and development of a machine learning model are usually not the end of solving a problem. The results of data collection and machine learning models are usually input for an optimization model, which produces decision that can be implemented. For instance, predicted numbers of commuters can be used for bus route design [22], predicted ship handling time can be used for pilot scheduling and shipping service design (see [23–25]), predicted amount of ship

emissions can be used for managing ship operations [26–31]. Moreover, a number of advanced techniques that solve problems involving prediction and optimization have been developed [32–34].

The remainder of this paper is organized as follows. Section 2 presents a detailed description for our material and methods. Section 3 reports the results. Conclusions are drawn in Section 4.

2. Dataset and methods

We have collected a dataset of 550 workers in construction sites in Hong Kong. The dataset contains 8 features: age, BMI (Body Mass Index, BMI), alcohol drinking habit, smoking habit, temperature, relative humidity, job nature, work duration, and one label, that is, RPE (Rating of Perceived Exertion, RPE). Chan et al. [35,36] have used this dataset in their study to explore construction workers' heat-stress problems. The detailed feature and label explanations can be found in Table 1. Descriptive statistics of the dataset are shown in Table 2.

Table 1. Feature and label explanations.

Feature name	**Feature explanation**
Age	Age of worker, e.g., 37
BMI	Body Mass Index, an indictor used to measure a person's degree of fat and thin (BMI = Body weight (kg) / Height^2 (m^2))
AH	Alcohol drinking habit (none = 0, occasionally = 1, usually = 2)
SH	Smoking habit (none = 0, occasionally = 1, usually = 2)
Temperature	Temperature of the construction sites (℃)
Relative humidity	Relative humidity of the construction sites (%)
Job nature	Binary variable that equals 0 if a worker engages in bar bending and 1 if a worker engages in bar fixing
Work duration	How long the work has worked (minute)
Label name	**Label explanation**
RPE	The rating of perceived exertion (RPE), an indictor to indicate the intensity of effort, stress, or discomfort during physical activities. RPE is an integer ranging from 1 to 10. The larger the value of RPE, the more tired the worker is.

We adopt four commonly used regression algorithms: multiple linear regression, ridge regression, LASSO regression and support vector regression (SVR) as these four models are typical and our problem is suitable for regression models. Multiple linear regression is a well-known method which fits feature variables and the label into a linear model by minimizing the residual sum of squares [37,38]:

$$f(\boldsymbol{x}_i) = \boldsymbol{\omega}^T \boldsymbol{x}_i + b \tag{1}$$

$$(\boldsymbol{\omega}^*, b^*) = \arg\min_{(\boldsymbol{\omega},b)} \sum_{i=1}^{N} (f(\boldsymbol{x}_i) - y_i)^2 \tag{2}$$

where subscript i means a sample, $i = 1, \ldots, N$, and N is the total number of samples; $\boldsymbol{x}_i$ and y_i represent feature variables and the label (i.e., RPE) of sample i, respectively. Multiple linear

regression aims to find the optimal $\boldsymbol{\omega}^*$ and b^* to predict unknown labels using the given feature variables. In our study, $\boldsymbol{x}_i$ is an 8-dimensional vector, as shown in Table 1, and $\boldsymbol{\omega}^*$ is an 8-dimensional coefficient vector.

Ridge regression introduces L2-norm in the objective function [39,40]:

$$(\boldsymbol{\omega}^*, b^*) = \arg\min_{(\boldsymbol{\omega},b)} \sum_{i=1}^{N} (f(\boldsymbol{x}_i) - y_i)^2 + \alpha \sum_{i=1}^{N} ||\boldsymbol{\omega}||^2 \tag{3}$$

where $||\boldsymbol{\omega}||^2$ represents the regularization term which is the sum of the squares of coefficients, and $\alpha \geq 0$ is a hyperparameter which makes a trade-off between prediction errors and the regular term. $\boldsymbol{x}_i$ is the 8-dimensional vector and y_i represents the label (i.e., RPE).

LASSO regression introduces L1-norm regularization in the objective function [41,42], i.e., $|\boldsymbol{\omega}|$ represents the sum of absolute value of coefficients:

$$(\boldsymbol{\omega}^*, b^*) = \arg\min_{(\boldsymbol{\omega},b)} \sum_{i=1}^{N} (f(\boldsymbol{x}_i) - y_i)^2 + \alpha \sum_{i=1}^{N} |\boldsymbol{\omega}|. \tag{4}$$

Both ridge regression and LASSO regression can effectively deal with the problem of overfitting as they add $\boldsymbol{\omega}$ into the minimization objective function. Moreover, LASSO regression can reduce the coefficients of some features to 0, i.e., LASSO regression can do feature selection [43]. We use cross validation to determine the value of α.

SVR can tolerate an ϵ error between $f(\boldsymbol{x}_i)$ and y_i, i.e., the prediction error is 0 when $|f(\boldsymbol{x}_i) - y_i| \leq \epsilon$. SVR can be expressed by the formulas below [44,45]:

$$\min_{(\boldsymbol{\omega},b,\xi_i,\hat{\xi}_i)} \frac{1}{2}||\boldsymbol{\omega}||^2 + C \sum_{i=1}^{N} (\xi_i + \hat{\xi}_i) \tag{5}$$

subject to

$$f(\boldsymbol{x}_i) - y_i \leq \epsilon + \xi_i, i = 1, \ldots, N \tag{6}$$

$$y_i - f(\boldsymbol{x}_i) \leq \epsilon + \hat{\xi}_i, i = 1, \ldots, N \tag{7}$$

$$\xi_i \geq 0, \hat{\xi}_i \geq 0, i = 1, \ldots, N. \tag{8}$$

In the above formulation, C is a hyperparameter used to make a trade-off between bias and variance. $\boldsymbol{x}_i$ is an 8-dimensional vector that represents the eight feature variables and y_i is the label of sample i. We also use cross validation to determine its value. Our label is the rating of perceived exertion that ranges between 1 and 8. Therefore, we set ϵ to be 0.1 in our experiments.

Referring to previous studies [46,47], we use mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) to evaluate the performance of these four regression models with different sample sizes.

For multiple linear regression, we set the sample size to 50, 75, …, 500 in each experiment. For each sample size, we conduct 100 experiments by randomly splitting the dataset into the training dataset which has a total of "sample size" records and using the remaining samples as the testing dataset. We record the results of each prediction in the testing dataset and calculate the values of MSE, MAE, and MAPE. Finally, we measure the model's performance at each sample size by calculating the average of the three metrics in the testing dataset respectively.

Table 2. Descriptive statistics.

Variable name	Mean	Standard deviation	Min	Max	Percentile		
					25th	50th	75th
Age	43.85	11.54	20.00	63.00	37.00	48.00	52.00
BMI	21.92	2.68	17.01	26.85	19.76	22.29	24.2
AH	0.57	0.69	0.00	2.00	0.00	0.00	1.00
SH	0.88	0.40	0.00	2.00	1.00	1.00	1.00
Temperature	30.25	1.58	27.10	34.20	29.00	30.00	31.40
Relative humidity	75.23	8.74	20.00	63.00	19.76	22.29	24.20
Job nature	0.49	0.50	0.00	1.00	0.00	0.00	1.00
Work duration	45.69	27.03	5.00	125.00	25.00	45.00	5.00
RPE	4.64	1.44	1.00	8.00	4.00	5.00	6.00

For ridge regression, LASSO regression, and SVR, we first need to deal with hyperparameters α and C. We use K-fold cross validation to determine the value of hyperparameters. We set α to 0.001, 0.01, 0.1, 1, 5, and 10 respectively. We randomly divide our dataset into 10 parts, i.e., $K = 10$. We use 9 parts of the data to train the model, and the remaining 1 part is used as the testing dataset to evaluate the performance of the model. With each value of α, we conduct 10 experiments and also calculate the values of MSE, MAE, and MAPE. By comparing the average values of these three indictors, we choose the optimal α for ridge regression and LASSO regression, respectively. Then, we use the optimal α in ridge regression or LASSO regression to test the impact of different sample size. The process afterwards is the same as for multiple linear regression. The determination of C for SVR is similar. We set C to different values and then conduct K-fold cross validation to choose the optimal C.

3. Results

The learning curves of the forecasting of RPE using four regression methods are reported in Figures 1–4, respectively. The left axis of the graph represents the results of MSE and MAE and the right axis represents the results of MAPE. The results of multiple linear regression (see Figure 1) show that the prediction error decreases with increasing sample size at first and then the error lines go stable. The results of the other three methods (i.e., ridge regression, LASSO regression and SVR) also show consistent laws. The value of MSE, MAE, and MAPE of the four methods finally converged to approximately 1.05, 0.82 and 20.5, respectively. By observing these 12 learning curves, we find that the performance of the four models no longer improves significantly when the sample size exceeds 250. To be more specific, the values of MSE, MAE and MAPE almost keep decreasing until the sample size exceeds 250. After the sample size exceeds 250: MSE does not decrease by more than 2.5% and sometimes even increases; MAE does not decrease by more than 1.5% and sometimes even increases; MAPE does not decrease by more than 2.5% and sometimes even increases. Thus, 250 samples are enough to conduct experiments considering the cost of collecting data and the efficiency of computing.

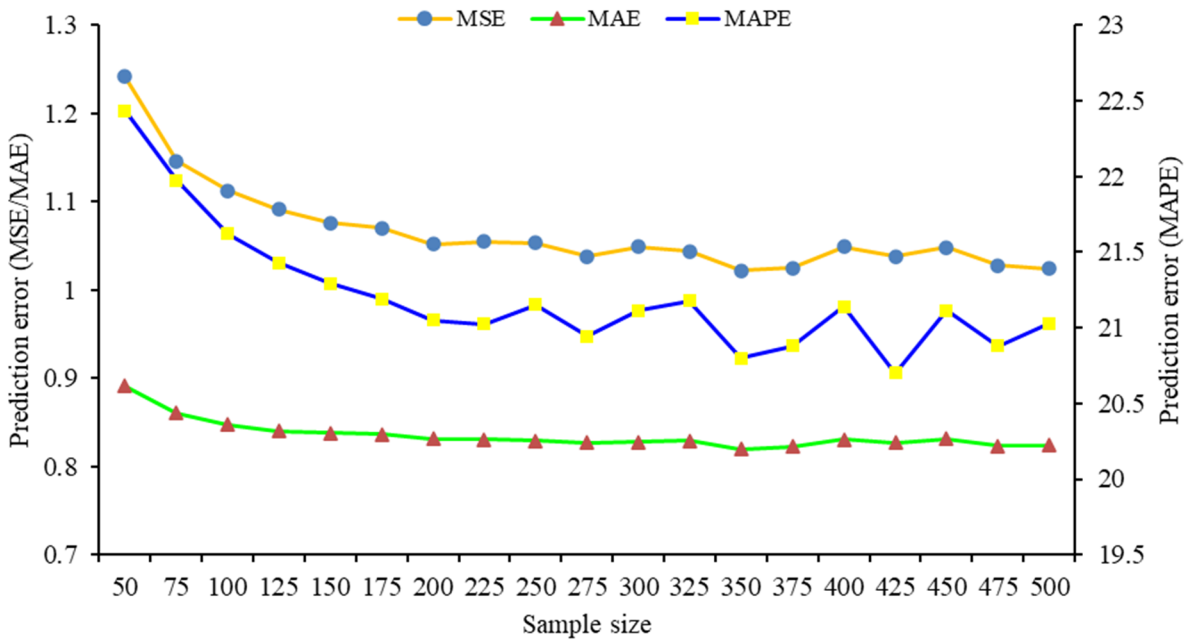

Figure 1. The results of multiple linear regression.

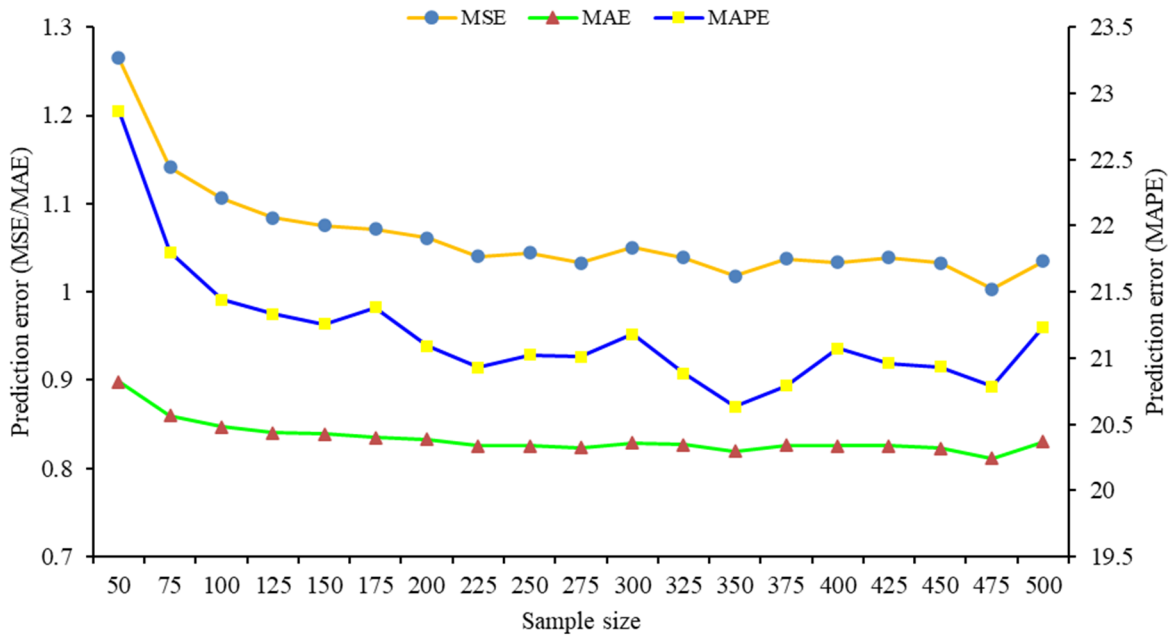

Figure 2. The results of ridge regression.

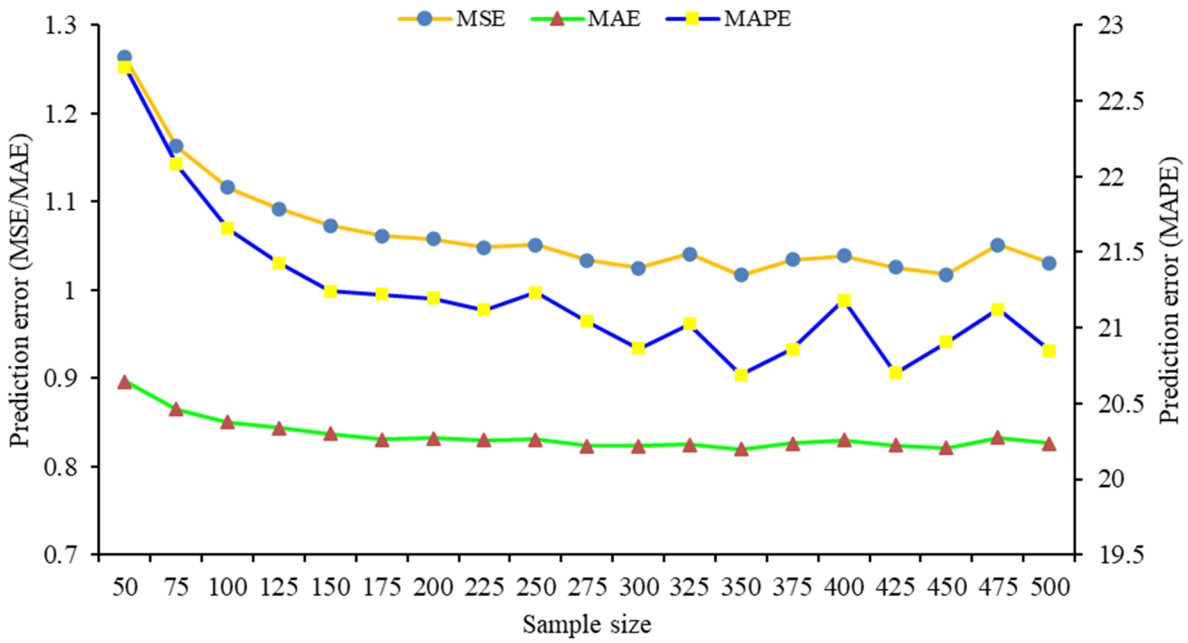

Figure 3. The results of LASSO regression.

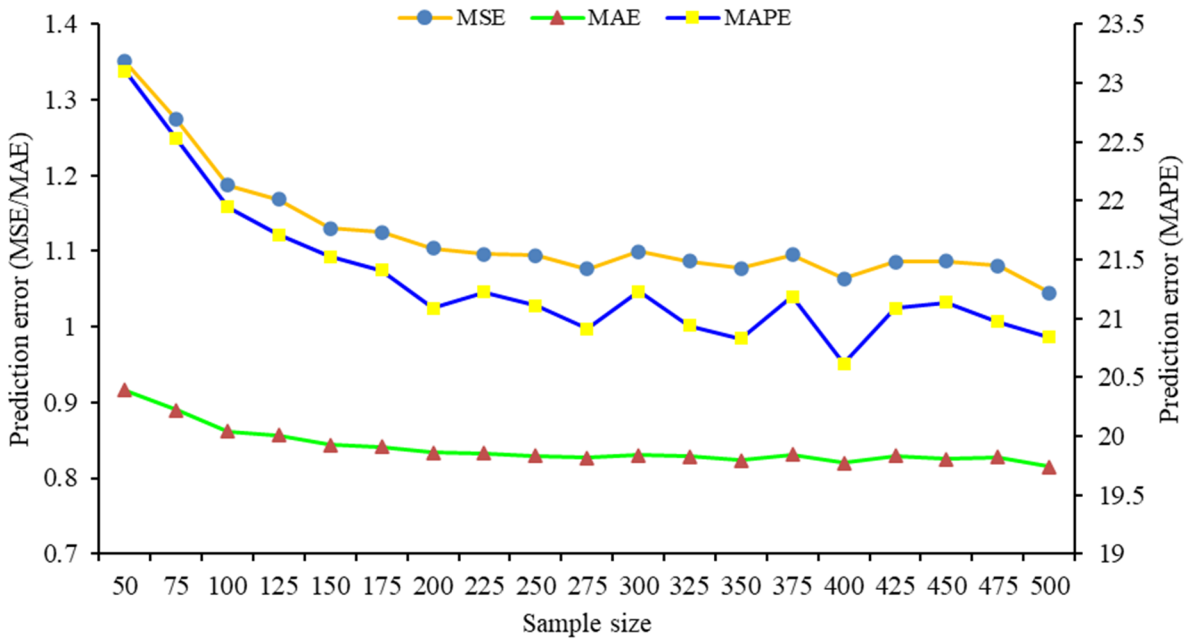

Figure 4. The results of SVR.

4. Conclusions

In this work, we adopt four typical regression models in machine learning—multiple linear regression, ridge regression, LASSO regression and SVR—to determine the optimal sample size in

the field of transportation and construction management. We find that a sample size of 250 is a good choice for scholars by the observation of the percentage decline of the three indicators. The observing process can be automated through calculating the error variation between two experiments and designing the experiment stopping rules. Error changes and the corresponding sample size can be easily recorded in a computer program. For example, we can set that we have found the optimal sample size when the percentage of error changes in five consecutive experiments do not exceed 3%.

As collecting data and running programs are costly and time-consuming, our study makes contributions in helping scholars and practitioners determine the used sample size. However, this study is not without limitations. First, the used dataset comes from transport infrastructure workers. Therefore, the findings may not be widely applicable to other fields. Second, we only use regression models to test impact of sample size. Other complicated machine learning models, e.g., Long Short-Term Memory network (LSTM), maybe a different story. Third, our dataset contains 8 feature variables and 1 label. If more feature variables are available to forecast the label, the optimal sample size may reduce as more variables provide more information. Similarly, if the value of the label is more diverse, the optimal sample size may increase for better prediction performance. Therefore, the used feature variables and labels will affect the results, which leads to a limitation of the study.

Acknowledgments

The authors thank the three referees for their constructive comments that significantly improve the quality of the paper. This study is supported by the start-up grant of The Hong Kong Polytechnic University (Project ID P0040224).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. H. Ding, N. N. Sze, Effects of road network characteristics on bicycle safety: a multivariate Poisson-lognormal model, *Multimodal Transp.*, **1** (2022), 1–9. https://doi.org/10.1016/j.multra.2022.100020
2. Z. Ma, P. Zhang, Individual mobility prediction review: data, problem, method and application, *Multimodal Transp.*, **1** (2022), 1–11. https://doi.org/10.1016/j.multra.2022.100002
3. X. Z. Simon, Q. Cheng, X. Wu, P. Li, B. Belezamo, J. Lu, et al., A meso-to-macro cross-resolution performance approach for connecting polynomial arrival queue model to volume-delay function with inflow demand-to-capacity ratio, *Multimodal Transp.*, **1** (2022), 1–28. https://doi.org/10.1016/j.multra.2022.100017
4. W. Yi, H. Wang, Y. Jin, J. Cao, Integrated computer vision algorithms and drone scheduling, *Commun. Transp. Res.*, **1** (2021), 1–4. https://doi.org/10.1016/j.commtr.2021.100002
5. X. Lang, D. Wu, W. Mao, Comparison of supervised machine learning methods to predict ship propulsion power at sea, *Ocean Eng.*, **245** (2022), 110387. https://doi.org/10.1016/j.oceaneng.2021.110387

6. J. Hu, W. Zou, J. Wang, L. Pang, Minimum training sample size requirements for achieving high prediction accuracy with the BN model: a case study regarding seismic liquefaction, *Expert Syst. Appl.*, **185** (2021), 1–13. https://doi.org/10.1016/j.eswa.2021.115702
7. C. Ma, X. Wang, L. Xia, X. Cheng, L. Qiu, Effect of sample size and the traditional parametric, nonparametric, and robust methods on the establishment of reference intervals: evidence from real world data. *Clin. Biochem.*, **92** (2021), 67–70. https://doi.org/10.1016/j.clinbiochem.2021.03.006
8. E. Burmeister, L. M. Aitken, Sample size: How many is enough? *Aust. Crit. Care*, **25** (2012), 271–274. https://doi.org/10.1016/j.aucc.2012.07.002
9. Z. Cui, G. Gong, The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features, *NeuroImage*, **178** (2018), 622–637. https://doi.org/10.1016/j.neuroimage.2018.06.001
10. H. Taherdoost, Determining sample size; how to calculate survey sample size, *Int. J. Econ. Manage. Syst.*, **2** (2017), 237–239. https://ssrn.com/abstract=3224205
11. D. Lakens, Sample size justification, *Collabra: Psychol.*, **8** (2022), 1–28. https://doi.org/10.1525/collabra.33267
12. S. Mao, G. Xiao, J. Lee, L. Wang, Z. Wang, H. Huang, Safety effects of work zone advisory systems under the intelligent connected vehicle environment: a microsimulation approach, *J. Intell. Connected Veh.*, **4** (2021), 16–27. https://doi.org/10.1108/JICV-07-2020-0006
13. L. Yue, M. Abdel-Aty, Z. Wang, Effects of connected and autonomous vehicle merging behavior on mainline human-driven vehicle, *J. Intell. Connected Veh.*, **5** (2022), 36–45. https://doi.org/10.1108/JICV-08-2021-0013
14. J. Zhu, S. Easa, K. Gao, Merging control strategies of connected and autonomous vehicles at freeway on-ramps: a comprehensive review, *J. Intell. Connected Veh.*, **5** (2022), 99–111. https://doi.org/10.1108/JICV-02-2022-0005
15. J. Zhu, I. Tasic, X. Qu, Flow-level coordination of connected and autonomous vehicles in multilane freeway ramp merging areas, *Multimodal Transp.*, **1** (2022), 1–13.
16. Y. Du, Q. Meng, S. Wang, H. Kuang, Two-phase optimal solutions for ship speed and trim optimization over a voyage using voyage report data, *Transp. Res. Part B Methodol.*, **122** (2019), 88–114. https://doi.org/10.1016/j.trb.2019.02.004
17. R. Yan, S. Wang, Y. Du, Development of a two-stage ship fuel consumption prediction and reduction model for a dry bulk ship, *Transp. Res. Part E Logist. Transp. Rev.*, **138** (2020), 1–22. https://doi.org/10.1016/j.tre.2020.101930
18. R. Yan, S. Wang, J. Cao, D. Sun, Shipping domain knowledge informed prediction and optimization in port state control, *Transp. Res. Part B Methodol.*, **149** (2021), 52–78. https://doi.org/10.1016/j.trb.2021.05.003
19. W. Yi, S. Wang, Mixed-integer linear programming on work-rest schedule design for construction sites in hot weather, *Comput.-Aided Civ. Infrastruct. Eng.*, **32** (2017), 429–439. https://doi.org/10.1111/mice.12267
20. Y. Li, Y. Lu, J. Chen, A deep learning approach for real-time rebar counting on the construction site based on YOLOv3 detector, *Autom. Constr.*, **124** (2021), 1–14. https://doi.org/10.1016/j.autcon.2021.103602

21. A. Shehadeh, O. Alshboul, R. Mamlook, O. Hamedat, Machine learning models for predicting the residual value of heavy construction equipment: an evaluation of modified decision tree, LightGBM, and XGBoost regression, *Autom. Constr.*, **129** (2021), 1–16. https://doi.org/10.1016/j.autcon.2021.103827
22. X. Qu, S. Wang, D. Niemeier, On the urban-rural bus transit system with passenger-freight mixed flow, *Commun. Transp. Res.*, **2** (2022), 1–3. https://doi.org/10.1016/j.commtr.2022.100054
23. K. Wang, S. Wang, L. Zhen, X. Qu, Cruise service planning considering berth availability and decreasing marginal profit, *Transp. Res. Part B Methodol.*, **95** (2017), 1–18. https://doi.org/10.1016/j.trb.2016.10.020
24. L. Zhen, Y. Hu, S. Wang, G. Laporte, Y. Wu, Fleet deployment and demand fulfillment for container shipping liners, *Transp. Res. Part B Methodol.*, **120** (2019), 15–32. https://doi.org/10.1016/j.trb.2018.11.011
25. L. Wu, Y. Adulyasak, J. F. Cordeau, S. Wang, Vessel service planning in seaports, *Oper. Res.*, 2022. https://doi.org/10.1287/opre.2021.2228.
26. L. Zhen, Y. Wu, S. Wang, G. Laporte, Green technology adoption for fleet deployment in a shipping network, *Transp. Res. Part B Methodol.*, **139** (2020), 388–410. https://doi.org/10.1016/j.trb.2020.06.004
27. J. Qi, S. Wang, H. Psaraftis, Bi-level optimization model applications in managing air emissions from ships: a review, *Commun. Transp. Res.*, **1** (2021), 1–5. https://doi.org/10.1016/j.commtr.2021.100020
28. S. Wang, H. N. Psaraftis, J. Qi, Paradox of international maritime organization's carbon intensity indicator, *Commun. Transp. Res.*, **1** (2021), 1–5. https://doi.org/10.1016/j.commtr.2021.100005
29. S. Wang, L. Zhen, D. Zhuge, Dynamic programming algorithms for selection of waste disposal ports in cruise shipping, *Transp. Res. Part B Methodol.*, **108** (2018), 235–248. https://doi.org/10.1016/j.trb.2017.12.016
30. S. Wang, D. Zhuge, L. Zhen, C. Y. Lee, Liner shipping service planning under sulfur emission regulations, *Transp. Sci.*, **55** (2021), 491–509. https://doi.org/10.1287/trsc.2020.1010
31. S. Wang, J. Qi, G. Laporte, Optimal subsidy design for shore power usage in ship berthing operations, *Nav. Res. Logist.*, **69** (2022), 566–580. https://doi.org/10.1002/nav.22029
32. S. Wang, R. Yan, A global method from predictive to prescriptive analytics considering prediction error for "Predict, then optimize" with an example of low-carbon logistics, *Cleaner Logist. Supply Chain*, **4** (2022), 1–3. https://doi.org/10.1016/j.clscn.2022.100062
33. R. Yan, S. Wang, Integrating prediction with optimization: models and applications in transportation management, *Multimodal Transp.*, **1** (2022), 1–5. https://doi.org/10.1016/j.multra.2022.100018
34. R. Yan, S. Wang, L. Zhen, G. Laporte, Emerging approaches applied to maritime transport research: past and future, *Commun. Transp. Res.*, **1** (2021), 1–14. https://doi.org/10.1016/j.commtr.2021.100011
35. A. P. Chan, W. Yi, D. W. Chan, D. P. Wong, Using the thermal work limit as an environmental determinant of heat stress for construction workers, *J. Manage. Eng.*, **29** (2013), 414–423.
36. A. P. Chan, W. Yi, D. P. Wong, M. C. Yam, D. W. Chan, Determining an optimal recovery time for construction rebar workers after working to exhaustion in a hot and humid environment, *Build. Environ.*, **58** (2012), 163–171. https://doi.org/10.1016/j.buildenv.2012.07.006

37. M. Flores-Sosa, E. León-Castro, J. M. Merigó, R. R. Yager, Forecasting the exchange rate with multiple linear regression and heavy ordered weighted average operators, *Knowl.-Based Syst.*, **248** (2022), 108863. https://doi.org/10.1016/j.knosys.2022.108863
38. Q. H. Luu, M. F. Lau, S. P. Ng, T. Y. Chen, Testing multiple linear regression systems with metamorphic testing, *J. Syst. Software*, **182** (2021), 1–21. https://doi.org/10.1016/j.jss.2021.111062
39. G. C. McDonald, Ridge regression, *Wiley Interdiscip. Rev. Comput. Stat.*, **1** (2009), 93–100. https://doi.org/10.1002/wics.14
40. G. Smith, F. Campbell, A critique of some ridge regression methods, *J. Am. Stat. Assoc.*, **75** (1980), 74–81. https://wwwtandfonline.53yu.com/doi/abs/10.1080/01621459.1980.10477428
41. C. R. Genovese, J. Jin, L. Wasserman, Z. Yao, A comparison of the lasso and marginal regression, *J. Mach. Learn. Res.*, **13** (2012), 2107–2143.
42. S. Wang, B. Ji, J. Zhao, W. Liu, T. Xu, Predicting ship fuel consumption based on LASSO regression, *Transp. Res. Part D: Transp. Environ.*, **65** (2018), 817–824. https://doi.org/10.1016/j.trd.2017.09.014
43. W. J. Fu, Penalized regressions: the bridge versus the lasso, *J. Comput. Graphical Stat.* , **7** (1998), 397–416. https://wwwtandfonline.53yu.com/doi/abs/10.1080/10618600.1998.10474784
44. V. Cherkassky, Y. Ma, Practical selection of SVM parameters and noise estimation for SVM regression, *Neural Networks*, **17** (2004), 113–126. https://doi.org/10.1016/S0893-6080(03)00169-2
45. W. C. Hong, Y. Dong, L. Y. Chen, S. Y. Wei, SVR with hybrid chaotic genetic algorithms for tourism demand forecasting, *Appl. Soft Comput.*, **11** (2011), 1881–1890. https://doi.org/10.1016/j.asoc.2010.06.003
46. D. Li, M. Qiu, J. Jiang, S. Yang, The application of an optimized fractional order accumulated grey model with variable parameters in the total energy consumption of Jiangsu Province and the consumption level of Chinese residents, *Electron. Res. Arch.*, **30** (2022), 798–812. https://doi.org/10.3934/era.2022042
47. X. Li, L. Kang, Y. Liu, Y. Wu, Distributed Bayesian posterior voting strategy for massive data, *Electron. Res. Arch.*, **30** (2022), 1936–1953. https://doi.org/10.3934/era.2022098

AIMS Press