ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 143-149
doi: 10.18178/wcse.2019.06.021

Proposed Forest Prediction System based on Large-scale Adaptive Boosting Support Vector Regression Method

Li-Li Wang[+] and Matthew R Evans

The University of Hong Kong, Faculty of Science, Hong Kong, China

Abstract: In this paper, a forest prediction system for incorporating large-scale data on individual trees into one hybrid model is proposed. The proposed algorithm incorporates both forest biometry and statistical information, and constructs the hybrid model through combining adaptive boosting classification and support vector regression learning from large-scale forest data. More specifically, the species of a tree is firstly identified based on its measured features by using the adaptive boosting method. Subsequently, for each tree species the system relates the height of trees to the diameter at breast height and annual mean temperature for each tree species through a Support Vector Regression technique. This allows the tree's height in the future to be well predicted. Experimental results show that the proposed algorithm has the capability to identify the species of trees and further predict tree growth through valid statistical inference.

Keywords: forest prediction system, large-scale forest data, adaptive boosting method, Support Vector Machines, growth height of trees, statistical inference

1. Introduction

Forest modelling involves the biometric description of the dynamics of real forest stands through computer programming. The degree of the description depends on the existing knowledge about the structure and behaviour of the real system. It is meaningful and significant to construct simulated models because the lifespan of trees is long (up to many centuries) and it is difficult to perform experiments covering these lifespans. Through these models to simulate the forest stand dynamics, one can obtain projections of the future state of forests, such as tree growth (shape, height, and diameter), propagation, survival of trees, the geographical distribution and so on. This information plays an important role in inferring the influence of large-scale factors such as global climatic change on the ecological environment. Climatic changes are likely to have effects on tree populations and geographical distribution of some species [1-2]. Projections of the future state of the forest is meaningful in tree planting and forestation for environmental protection, to ensure that the natural ecosystem along with its associated goods and services is sustainable.

One approach to predict tree growth is to consider the response to temperature. This approach is necessary if we wish to take into account the potential effects of climatic change on global forests. It is expected that the climate change would affect tree growth and survival. Generally speaking, approaches include: parabolic degree-day response method [3], relative temperature response method [4], carbon assimilation method [5], empirical optimum temperature day method [6], and empirical cumulative degree-day method [7]. However, these modelling methods lack consideration of genetic variability in tree growth. The other method is to consider non-external-related aspects, such as diameter at breast height (DBH), and crown size, to predict tree growth [8, 9]. Note that temperature is not only a factor used to calculate the potential effects of climatic change on forests. It also directly influences chemical and physiological process. In this research, we present one mixed model that considers both DBH and temperature to make predictions of tree growth. How to achieve an accurate relationship that relates the tree height to some known parameters is significant for a forest forecasting system. In the previous forest

[+] Corresponding author. Tel.: + 852 67066932.
E-mail address: llwang@hku.hk.

forecasting system, the model is directly assumed to be linear or other relationship. In this paper, we obtain the model for tree growth forecasting through artificial intelligence learning from large-scale data.

For a learning based method, it is necessary to have large-scale data. However, it can be difficult to collect sufficient data due to the long lifespan of forest stands. To strike a balance, we propose a method that embeds Gaussian noise under a constraint of a high signal-noise-ratio (SNR) to generate large amount of simulated data for the purpose of training and testing. Note that it can be difficult to identify the species of trees seen in forests or gardens for most people. It is necessary first to acquire the species of trees, and further to forecast their growth. To achieve these objectives, we propose a hybrid method by taking advantages of both Adaptive Boosting (AdaBoost) [10] and Support Vector Regression [11] (AB-SVR) techniques in this paper. More specifically, an AdaBoost method is firstly used to identify the species of one tree. Adaptive boosting is a particular machine learning method used to train a series of weak classifiers. During the training process for AdaBoost, the weights of the training samples are adaptively updated after each boosting iteration. The weights of the training samples which are misclassified by the current component classifier are increased, while the weights of the training samples with correct classification are decreased. Finally, the weak classifiers are combined linearly to form a strong classifier. Secondly, the SVR method is adopted to predict the tree growth of each species. Through observation of the collected data, we can see that there exists non-linear statistical relationships among height, DBH and temperature. It is critical to transform DBH and temperature to a higher dimensional space to make the relationship linearly expressed. The proposed algorithm is to train a series of SVR models with non-linear kernel to perform prediction. Making full use of the acquired information, the future height of tree is well predicted according to the experimental results.

The rest of the paper is organized as follows. In Section 2, large-scale data preparation is discussed. The proposed method to identify tree species and predict forest growth is presented in Section 3. Experimental results are given and discussed in Section 4. Section 5 concludes the paper.

2. Large-scale Forest Data Preparation

To develop the tree growth model, the datasets from two woodland sites: United Kingdom woodlands-Wytham Woods and Alice Holt [12] are used. These two woodland sites are monitored by the Environmental Change Network (ECN) in the United Kingdom. Each individual tree in the ECN Wytham Wood (ECN-W) dataset had height measured three times (1993, 2002 and 2012). For each individual tree in the ECN Alice Holt (ECN-AH) dataset, height was measured on three occasions (1994, 2002, and 2011). These height, DBH and temperature [13] information form the dataset for the regression model formulation. In total, 1122 individuals of eight categories are adopted in this research to develop ecological models that are capable of projection into the future. Table 1 lists the details of 8 species of deciduous trees in the two sites including taxonomic code, common name, the number (Num1.) of each species, the maximum and the minimal information of both height and DBH. In Table 1, minHeight (minDBH) and maxHeight (maxDBH) denote the minimal and the maximal height (DBH) values of trees in each species.

Table 1 Detailed information of trees in the dataset (units are m)

Taxonomic code	Common name	Num1.	minHeight	maxHeight	minDBH	maxDBH	Num2.
ACERPS	Sycamore	156	3.5	30.5	0.05	1.00	17
ACERCA	Field maple	51	6.0	19.0	0.08	0.59	12
BETUSP	Birch	94	4.0	22.5	0.04	0.45	\
CORYAV	Common hazel	118	3.0	19.5	0.04	0.25	9
CRATMO	Common hawthorn	91	3.0	15.0	0.05	0.39	5
FAGUSY	European beech	63	3.5	39.0	0.06	1.62	7
FRAXEX	European ash	229	4.0	37.5	0.05	0.79	17
QUERRO	Pedunculate oak	320	3.5	35.5	0.05	1.54	9

Besides DBH and height, other information, such as height to bottom crown, crown radius, crown height and so on, were measured. To identify the species of trees, DBH, height, crown height and mean crown radius at the same time point are used as input features of one tree. Since some trees were dead or their information was missing during the observed period, as a result, the number of trees (Num2.) with these four features at the same time point is insufficient for a learning based classification. On the other hand, it is very

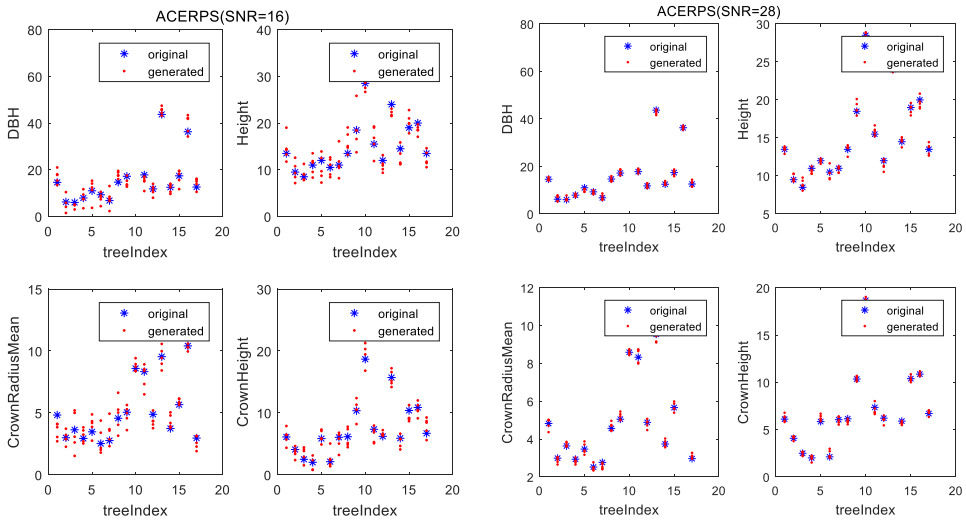
time-consuming to collect massive amount of trees' data. To handle this problem, we generate a large amount of simulated data through embedding Gaussian white noise under a constraint of a high signal-noise-ratio (SNR) for the purpose of training and testing. Fig. 1 shows an example of the simulated trees' DBH, height, mean of crown radius and crown height with different values of SNR for ACERPS species. Due to an introduction of higher noise with the value of SNR equal to 16, we can see from the figure that the fluctuation of ST signals is large compared that produced with a higher value of SNR equal to 28. Through this method, a quantity of 657 simulated trees are generated from the original 76 trees. The detailed procedures are concluded as follows:

Step 1: read the information {I_i| i = 1, 2, 3 and 4 for DBH, height, mean of crown radius and crown height} for each tree in 2012 from the collected dataset.

Step 2: add Gaussian white noise N_i with an optimal value of SNR to each I_i.

$$x_i' = I_i + N_i, \quad i = 1, \ldots, 4 \tag{1}$$

Step 3: Repeat Steps 1 and 2 until the number of generated data is equal to the predefined value.



(a) (b)

Fig. 1: Generated data with different values of SNR based on the additive model:
(a) ACERPS (SNR=16), and (b) ACERPS (SNR=28)

3. Proposed Forest Growth Modelling

Having a uniform method with adjustable model parameters is promising for forests in different geographic areas since forest growth can be influenced by their environmental conditions. In this paper, the AB-SVR is developed for classification of tree species by training a strong classifier, and further to predict the future growth based on a trained support vector regression method. Figure 2 depicts the flowchart of the AB-SVR model. More specifically, large-scale data are first collected as depicted in Section 2. Data are split into training and testing sets. Features are then extracted. In the AB-SVR framework, adaptive boosting technique is adopted for training a strong classifier to identify trees species, and SVR is employed to obtain a regression model to predict the tree growing behaviour of each species based on the current information.

The procedures of the proposed hybrid AB-SVR model are described as follows:

Step 1: large-scale data generation based on the method described in Section 2

Step 2: split data into training set and testing set

Step 3: for each training sample, the feature vector [DBH, Height, CrownHeight, CrownRadiusMean] is extracted

Step 4: train a strong classifier based on AdaBoost method to predict the tree species

$$F(x) = \max_c \sum_{t=1}^{T} h_t(v) = \max_c \sum_{t=1}^{T} h(v, f_t, \theta_t, c) \tag{2}$$

where h_t is the t^{th} weak learner, θ is a threshold, v denotes a feature vector, and f_t means that the f^{th} component of v is used as input feature in weak learner $h_t(v)$. The optimal AdaBoost classifier for tree species recognition is obtained through a training process [10]. Once one tree species is identified based on the AdaBoost classifier, the achieved SVR model is subsequently employed to predict the future behaviour of this tree.

Step 5: train a regression model to predict the tree height of each species

The main purpose of the ecological models is to make more accurate long-term predictions for tree growth, survival, reproduction etc and use these to understand changes in the structure of the forest over time. Note that the size of trees depends on age, light, water, access to minerals, stand structure and so on. In this investigation, we relate the height of a tree to its DBH (diameter at breast height) and temperature through training SVR models. SVR includes linear and non-linear regressions [11] by selecting different kernel function, which reflects similarity between data points. Training an SVR model involves finding an optimal function that is consistent with as many of the training data points as possible, such that the function can predict the height information of trees more accurately. Figure 3 shows the distribution of samples in the (DBH, temperature, height) space. From this figure, we can observe that different species of trees have inconsistent distributions. Intra-specific SVR model is necessary to be trained for each species of trees. Another observation from Figure 3 is that the relationship among them is non-linear. To find this non-linear regression function, SVR adopts a kernel function [11] to map the training data points from the original data space to a higher dimension data space where an optimal function may exist. The radial basis function (RBF) was chosen as the kernel function to transform DBH and temperature to a higher dimensional space for accurate prediction. RBF kernel is used in the SVR model to relate height to DBH as in eqn. (3)

$$f(\mathrm{x}) = \sum_{i=1}^{N} \boldsymbol{w_i} e^{(-\boldsymbol{\gamma} * \|\boldsymbol{\mu_i} - \mathbf{x}\|^2)} + b \tag{3}$$

where N is the number of support vectors, $\boldsymbol{\gamma}$ is gamma parameter, w and b are coefficients. The optimal $\boldsymbol{\gamma}$ is obtained through validation set. The regression coefficients $\mathbf{w}$ and b can be obtained by minimizing the Vapnik's $\epsilon-$insensitivity loss function [11]. The SVR model with the RBF kernel was trained to approximate the set of data $(y_1, \mathbf{x}_1), \ldots (y_m, \mathbf{x}_\mathrm{m}), \mathbf{x} \in R^n, y \in R$, and explore the relationships among tree height, DBH and temperature in a long time. Here y denotes the real tree height, *f* denotes the predicted tree height, $\mathbf{w}$ and b are the regression coefficients. During the training process, large-scale data is firstly clustered based on tree species. Each of them is considered as a dataset to train a SVR model for forecasting forest growth.

Step 6: given a tree, the tree species is firstly identified based on AdaBoost classifier, and then the future height could be predicted based on the obtained regression model.

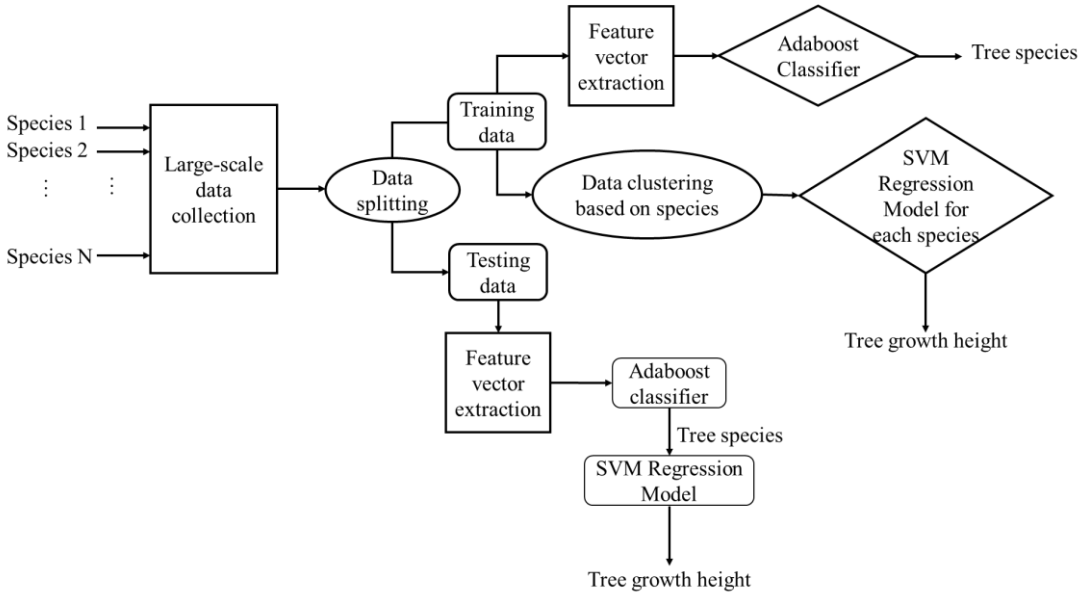


Fig. 2: Flowchart of the proposed hybrid AB-SVR model

4. Experimental results and discussion

In the simulation, ten-fold evaluation is performed on the 733 trees in 2012 for AdaBoost based classification method. The proposed algorithm is realized in MATLab. Table 2 list the classification accuracies by using AdaBoost based classification method. From the results, we can see that the proposed method has better classification accuracies with an average of 95%. Once the species of one tree is identified, subsequently, future height can be predicted based on SVR learning method.

Table 2. Classification results based on AdaBoost method (%)

Fold no.	1	2	3	4	5	6	7	8	9	10
Accuracy (%)	97	96	86	92	93	95	96	96	97	97
Average (%)	95									

Table 3 lists the confusion matrix for the 7 observed tree species, respectively. From the results, we can see that most of the species can be identified with high accuracies, such as 96% (ACERCA), 99% (CORYAV), 99% (CRATMO), 99% (FAGUSY), 91% (FRAXEX) and 95%(QUERRO). For the species of ACERPS, 14% is misclassified as other species (ACERCA, CORYAV, FAGUSY, FRAXEX or QUERRO). In the future, we will develop the image-based model to recognize the species of one tree. The classification accuracy could be further improved through deep learning [14, 15] from the view of visualization.

Table 3. Confusion matrix of seven tree species (%)

	ACERPS	ACERCA	CORYAV	CRATMO	FAGUSY	FRAXEX	QUERRO
ACERPS	86	2	1	0	3	4	4
ACERCA	2	96	0	1	0	0	1
CORYAV	0	1	99	0	0	0	0
CRATMO	0	2	0	99	0	0	0
FAGUSY	0	0	0	0	99	1	0
FRAXEX	1	5	1	1	0	91	1
QUERRO	2	3	0	0	0	0	95

The predicted results of heights based on DBH and temperature for each category are shown in Figure 3. The blue points in Figure 3 denote the real samples, and the red points denote their predicted results. The results show that the tree height could be well predicted through the learned SVR models.

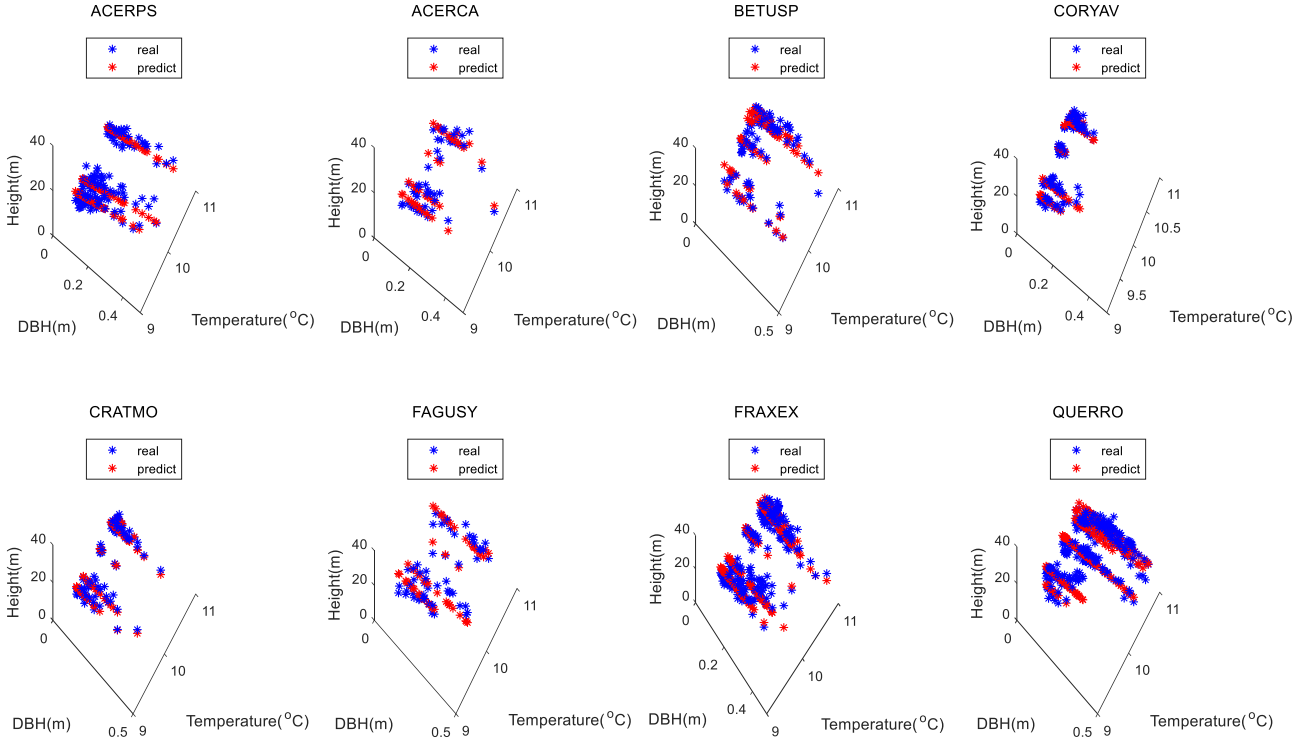


Fig. 3: regression results (red curve) with real samples (blue points) for each category

5. Conclusion

In summary, this paper presents a hybrid model for forest forecasting system. This model integrates both adaptive boosting and Support Vector Machine regression techniques to recognize the tree species and further predict its future growth. The SVR model considers both the potential effects of climatic change on forests and intrinsic growth characteristics. Through regression learning, this approach achieves the relationship that relate the tree growth height to some known parameters. It could make more accurate long-term predictions for tree growth. For practical applications, it is important for us to predict the future growth behaviour of forests. This would be beneficial to development of ecological environment. It is meaningful in tree planting and forestation for environmental protection. Note that deep learning is a very hot topic at present and the technique becomes more and more practical. For the future work, we will collect more visual data, such as images of tree leaves, as features to automatically identify the tree species based on deep learning with conventional neural networks. Once the species is identified from camera trap images or network images, the future growing behaviour could be forecasted based on the proposed regression model. On the other hand, we will focus on more elements which could affect the growth of trees to make the regression model more accurate.

6. Acknowledgments

This research was supported by The University of Hong Kong

7. Reference

[1] Gates, D.M., 1993. Climate Change and its Biological Consequences. Sinauer Associates, Sunderland, MA. 280 pp.

[2] Vance, E.d., 1995. Global change and forest responses: Theoretical basis, projections and uncertainties. NCASI Tech. Bull., 690:1-53.

[3] Sirois, L., Bonan, G.B. and Shugart, H.H., 1994. Development of a simulation model of the forest-tundra transition zone of northeastern Canada. Can. J. For. Res., 24: 697-706.

[4] Cumming, S.G. and Burton. P.J.. 1994. Zelig++ v0.9 Documentation. User Notes and Installation Guide. Department of Forest Sciences. University of British Columbia. Vanouver, B.C.. Canada, 29 pp.

[5] Prentice, I.C., Sykes, M.T. and Cramer, W., 1993. A simulation model for the transient effects of climate change on forest landscapes. Ecol. Model., 64: 51-70.

[6] Dale, V.H. and Hemstrom, M., 1984. CLIMACS: A Computer Model of Forest Stand Development for Western Oregon and Washington. Research Paper PNW-327. USDA Forest Service, Pacific Northwest Forest and Range Experiment Sation, Porland, OR, 60 pp.

[7] Reed, D.D., Jones, E.A., Holmes, M.J. and Fuller, L.G., 1992. Modeling diameter growth in local populations: A case study involving four North American deciduous species. For. Ecol. Manage., 54:95-114.

[8] Purves, D. W., J. W. Lichstein, N. Strigul, and S. W. Pacala. 2008. Predicting and understanding forest dynamics using a simple tractable model. Proc. Natl Acad. Sci. USA 105:17018–17022.

[9] Evans, Matthew R., and Aristides Moustakas. "A comparison between data requirements and availability for calibrating predictive ecological models for lowland UK woodlands: learning new tricks from old trees." Ecology and evolution6.14 (2016): 4812-4822.

[10] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection," in Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, 2004, pp. II-762-II-769 Vol. 2.

[11] Alex J Smola and Bernhard Scholkopf. “A tutorial on SVM regression. Statistics and computing”, 14(3):199–222, 2004.

[12] Matthew R. Evans, Aristides Moustakas, Gregory Carey, Yadvinder Malhi2, Nathalie Butt, Sue Benham, Denise Pallett & Stefanie Schäfer, “Allometry and growth of eight tree taxa in United Kingdom woodlands”, Scientific data, 1-9, 2015.

[13] Parker, D.E., T.P. Legg, and C.K. Folland. 1992. A new daily Central England Temperature Series, 1772-1991. Int. J. Clim., Vol 12, pp 317-342

[14] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffry. "ImageNet Classification with Deep Convolutional Neural Networks" (PDF). NIPS 2012: Neural Information Processing Systems, Lake Tahoe, Nevada.

[15] Schmidhuber, J.， "Deep Learning in Neural Networks: An Overview". Neural Networks. 2015， 61: 85–117. arXiv:1404.7828.