THE TOKYO FOUNDATION
FOR POLICY RESEARCH

TKFD Working Paper Series No.20-05

Forecasting Japanese Elections by Applying Ensemble Learning Methods

Budrul Ahsan
Sota Kato
Shun Ibaragi

March 2021

106-6234 東京都港区六本木3-2-1 六本木グランドタワー34階
The Roppongi Grand Tower 34th Floor
3-2-1 Roppongi, Minato-ku, Tokyo 1066234 Japan
政策研究 Policy Research | Tel: 03-5797-8401 Fax: 03-5570-6031

Forecasting Japanese Elections by Applying Ensemble Learning Methods

Budrul Ahsan
Sota Kato
Shun Ibaragi

Abstract

This paper applies machine learning methods, namely, the decision tree algorithm and ensemble learning methods to forecast Japanese lower house elections. By applying these two machine learning methods, we developed several non-linear forecasting models. We then compared our forecasting results with those generated by Lewis-Beck and Tien (2012), a pioneering model on this topic, using the same data and the same explanatory variables. We found that, even without tuning, all of our ensemble-learning-based forecasting models exceeded Lewis-Beck and Tien's model in mean accuracy. All of them also provided better explanations for mean variance. Despite small sample size inherent for country-specific forecasting models, our non-linear forecasting models of Japanese election generated better performance than linear models. We believe, by combining with substantive theories of each country's political situation, our methodological approach can improve predictability of country-specific forecasting models of other countries.

I. Introduction

Various types of election forecasting have been carried out in democracies around the world. Lewis-Beck and Tien (2011) divided them to three leading approaches: polls, election forecasting markets (e.g., Arrow et al. 2008), and modeling. In this paper, we take up the third, the modeling approach for electoral forecasting of Japanese elections. Lewis-Beck and Tien (LBT) (2012) claimed that, before their pioneering attempt, no one had constructed forecasting models for Japan.[1] Even after their attempt, very few have followed up with additional endeavors (for an exception, see Nasuno 2015).

The modeling approach combines substantive theory and methodological theory. Standard theories of electoral behavior are usually used for the former; regression theory is usually used for the latter, including the work of LBT on Japanese elections. In this paper, we focus on the latter, methodological theory. Instead of the regression theory used in LBT, we introduce a new methodological approach, namely, the decision tree algorithm and ensemble learning methods, developed in the field of machine learning. We use the same Japanese data and variables as LBT to comparatively show our model's improved performance over their pioneering work. We further argue that our new methodological approach can, by combining substantive theory that fits each country's political environment and institutions, improve the predictivity of electoral forecasting models of not only Japan but other countries.

One of the reasons for the underdevelopment of Japanese electoral forecasting models is a serious small-sample problem, which is inherent in electoral forecasting of any single country. Only 26 lower house elections have been held in Japan following the country's full democratization after the end of World War II. Such sample-size constraints hang especially heavily on constructing sophisticated, non-linear forecasting models. LBT thus used a very simple linear model to forecast the percentage of seats to be won in lower house election by the Liberal Democratic Party (LDP), the party that dominated Japan's postwar political landscape.[2] The larger sample sizes of cross-national electoral forecasting models, on the other hand, have recently enabled the development of non-linear forecasting models utilizing machine learning techniques (e.g., Kennedy et al. 2017). However, as the no-free-lunch-theorem (Wolpert & Macready 1995) implies, incorporating country-specific institutional and behavioral aspects of elections might substantially enhance the models' accuracy.

---

[1] There are a few possible exceptions to the Lewis-Beck and Tien's claim, including Inoguchi (1981).

[2] Since 1955 when the LDP was established, it has been the ruling party of Japan until the present except for about 5 years.

Cross-national forecasting models obviously face difficulties in accounting for such country-specific factors.

We developed non-linear forecasting models for Japanese elections by combining machine learning techniques, namely, the decision tree method and ensemble learning method. The decision tree method, unlike other non-linear learning algorithms such as neural networks, specifies its decision-making logic. It thus is compatible with modeling-type electoral forecasting where the model substantively specifies the causal relations between variables. The decision tree method, however, is prone to overfitting problems. It works well with training data but not with new data. To overcome the overfitting problem, we ensembled decision trees and created an ensemble model of decision trees to forecast election results. By combining these methods, we attempted to develop non-linear forecasting models that specify causal relations between variables and can more accurately make predictions. Since our goal is to show how our methods improved the performance of forecasting models methodologically, we fixed the substantive part of our forecasting model by using the same variable and the same dataset as LBT's pioneering work on Japanese electoral forecasting.

Overall, our non-linear machine-learning approach to Japanese electoral forecasting showed promising results. We first divided the dataset into a training dataset and testing dataset; using the former to train model and adjust its hyper-parameters by cross validation; and then evaluating generalization performance using the latter. As for indicators to evaluate the models, we used maximum absolute error and explained variance. All our five models obtained better scores for the two indicators compared to LBT's model. Among them, the models using random forest and Gradient boosting performed the well; they reduced maximum absolute error by 0.50 to 0.76 points and increased explained variance by 0.10 to 0.12 points. Since we assumed the same substantive relations between variables and electoral results as LBT, we argue that the enhanced predictability of our model can be explained by the difference between our methodological approach using non-linear ensemble learning methods and LBT's methodological approach using linear regression models.

We believe the advantage of our approach is not confined to a Japanese forecasting model; it can be applied to other country-specific forecasting models and can substantially enhance their predictability. One of the aims of LBT's paper was to show the theoretical compatibility of their model across borders. If the substantive part of their model is theoretically compatible, our model which only changed the methodological part of theirs' should also be compatible. Our approach also allows researchers to incorporate country-specific institutional and behavioral aspects into non-linear

forecasting models. Inclusion of such aspects should further increase predictability of country-specific forecasting models regardless of country.

II. Data and Methods

1. Data

To rigidly compare our model with LBT's (2012) forecasting model, we used the same dataset with theirs: Japanese lower house election results for the Liberal Democratic Party (LDP) from November 1955, when the LDP was established. Since several lower house elections were held since LBT's paper,[3] we have added them to evaluate both our model and LBT's model.

The outcome (dependent variable) is the LDP's seat occupancy rate (LDP seats). It is defined as the number of seats the LDP won in the lower house election divided by the total number of seats in the lower house. As for elections after the comprehensive electoral reform of 1993, the number of seats includes the number of seats the LDP won in single-member districts as well as proportional representation constituencies.

We also used the three explanatory variables of LBT's model. The first is real GDP growth rate (calendar year, GDP). As LBT pointed out, in addition to theoretical claims, several electoral studies empirically found significant relationships between pre-election key economic indicators such as GDP growth rate and the percentage of votes the incumbent party earned in the election (e.g., Fair 1978; Inoguchi 1981; Lewis-Beck & Rice 1984; Hirano 2007, 2012). Replicating LBT's model, we added real GDP growth rate a year before the election year. The second is the cabinet approval rate (PM approval). The positive relation between the approval rate and electoral results is theoretically straightforward and empirically reported in several studies (e.g., Sigelman 1979; Lewis-Beck and Rice 1984). We again copied LBT's model and included the cabinet approval rate surveyed by Jiji Press one month before the election. The third is the number of days between two consecutive elections (Days). Japanese prime minister who concurrently holds the leadership position of the incumbent party (i.e., LDP for this paper), can strategically call an early election before his term expires in the lower house (Kato and Inui 2013). An earlier election implies the situation is advantageous for the prime minister and the incumbent party because if the situation was not favorable, the prime minister could have waited until the situation gets better.

Although there were 21 lower house elections between 1958 and 2017, we excluded the

[3] We added all the lower house elections since then through the most recent one, held in 2017.

1958 and 2012 elections from our study. As for the 1958 election, it was the first one since the founding of the LDP, so we could not measure the explanatory variable "days between two consecutive elections (Days)." As for the 2012 election, it was the only one called by a non-LDP party, so we could not measure the dependent variable "LDP's seat occupancy rate." Table 1 shows the descriptive statistics of each variable used in this paper.

Table 1:Descriptive Statistics of Variables

	LDP_seats	GDP	PM_approval	Days
count	19	19	19	19
mean	52.99	4.27	36.12	1078.89
sd	8.77	3.63	9.34	283.98
min	24.80	-1.09	16.30	259.00
25%	48.60	1.52	29.30	983.50
50%	55.20	3.38	37.90	1096.00
75%	59.85	5.91	41.70	1260.00
max	63.40	11.91	54.80	1456.00

2. Methods

In this paper, we develop non-linear forecasting models to more accurately predict electoral results in Japan. Since the modeling approach to electoral forecasting combines substantive and methodological theories, it requires the causal interpretability of prediction results. Among the various machine learning techniques, we therefore decided to use decision tree algorithms that specify causal relations between variables. Furthermore, to avoid the overfitting problem often caused by the decision tree algorithms, we applied ensemble learning methods to ensemble decision trees, using the algorithms cheat sheet of python (Version 3.6) package scikit-learn.[4] In section 2.1 we describe the methods we use to create forecasting models in this paper. The next section 2.2. describes methods we used to evaluate the performance of each model.

2.1 Decision tree and ensemble methods

A decision tree is a non-parametric and non-linear supervised learning method. This algorithm is easy to intuitively grasp because it creates models that predict target value or class by a simple logical sequence. It works for both discrete and continuous target variables. Unlike other non-linear learning

[4] See https://scikit-learn.org/stable/tutorial/machine learning map/ for details regarding the python cheat sheet.

algorithms such as neural networks, a decision tree explicitly specifies its decision-making logic. Some scholars insist that this explainable attribute makes it the most preferable among machine learning algorithms (Seni and Elder 2010). However, models created by decision trees are likely to overfit. They work well for training data but not for testing or for new data. To overcome the overfitting problem, ensemble decision trees are known to be effective. We thus created ensemble models of decision trees to forecast election results.

Ensemble learning methods construct multiple prediction models for a single regression or classification problem and, through the modeling process, weak learners such as linear regressions and decision trees are ensembled into a strong learner (Zhou 2012, Freund and Schapire 1997). The methods have improved performance for a variety of machine learning tasks (Caruana and Niculescu-Mizil 2006).

Ensemble methods are broadly sorted into two categories; sequential techniques and parallel techniques. Sequential techniques train learners sequentially by correcting errors of the learners. AdaBoost (Freund and Schapire 1997) and gradient boosting (Friedman 2001) are typical techniques of this category. The AdaBoost (Adaptive boosting) algorithm generally uses a single split decision tree called a decision stump as a weak learner and assigns weights to observations that are difficult to predict by the weak learner. In each step of the sequential model building process, AdaBoost creates a new sample distribution by providing greater weights to misclassified observations and lesser weights to correctly classified observations. A targeted weak learner is trained on updated sample distributions and, at the end of sequential process, all the models are ensembled into a strong model. The strong model makes predictions by averaging the weights of each model that was ensembled.

Gradient boosting also generates ensemble models through a sequential process. First, it builds a model on a training dataset, and then it measures the residual loss of this model by calculating difference between original and predicted values. In the next step, another model is trained using a weak learner based on the residual loss of the previous step. This sequential model building continues until the residual loss reaches zero or a certain threshold. These newly added weak models are built on the observations where the previous models together are not performing well. In contrast to AdaBoost, gradient boosting does not create new sample distributions for training a weak learner in each step of the sequential process. eXtreme Gradient Boosting (XGBoost, Chen et al 2015), which has recently been frequently used in the field of machine learning, is also a sequential type method. XGBoost is an efficient implementation of a gradient boosting decision tree. XGBoost improves the

original gradient boosting using performance improvement measures, a novel tree building strategy, and an optimal treatment of missing values.

The parallel technique of ensemble methods, on the other hand, trains weak learners independently and integrate them to create a single ensemble model. Bagging (Breiman 1996) and random forest (Breiman 2001) are typical methods of this class. For example, bagging (Bootstrap Aggregating) builds multiple decision tree models using multiple bootstrap samples collected from an original single sample. Multiple decision models are then aggregated to create an ensemble model. The output of this ensemble model is calculated by averaging the results of all the multiple decision trees.

Random forest uses a bootstrapping resampling method and randomly selects a certain number of explanatory variables to generate multiple decision trees. The trees are combined to create the final model. It is known that if each decision tree is independent, the model will dramatically reduce the variance of output (Zhou et al 2015). In the following section, we use the AdaBoost, gradient boosting, bagging, random forest and XGBoost ensemble methods to forecast Japanese elections.

2.2 Performance evaluation methods

Here, we describe how we evaluate the models we created by the methods described in the previous section (section 2.1) and evaluate their performance. We first shuffled the dataset and used 75% of them to train the models. In modeling, we used LOO (leave-one-out) cross-validation. LOO cross-validation is an effective validation method of K-division cross-validation and is suitable for verification of models with a small number of data. We performed LOO on the training data, and hyper-parameters were adjusted by grid search. For the validation metrics, considering possible excessive effects of outliers due to small sample size, we used average absolute error.

We used the remaining 25% of the dataset as testing data and evaluated generalization performance. Since either training or testing data may be biased depending on the size of data, we shuffled 10 times and calculated the average value. To make a proper comparison with the original LBT model, we went through the same modelling process with the linear model.

III. Performance Results

In this section, we first describe how we tuned hyperparameters when modeling the training data

(section 3.1). We then show and compare, using the training data, performance of our models with the original linear model (section 3.2). Finally, we test our models using the testing data to examine generalization performance and compare it with the original linear model (section 3.3).

3.1 Tuning hyper-parameters

We adjusted the following four types of hyperparameters for our modeling:

- Regularization: to prevent overfit-learning, we adjusted the depth of decision trees, number of leaves, and splits.
- Number of weak learners: adjusted the number of weak learners used in combination.
- Optimization parameters for sequential models: adjusted learning rate, eta, and others to tune the complexity of the model.
- Evaluation: the evaluation metrics are unified to the mean absolute error (MAE), as mentioned in Section 2.2.

3.2 Performance results using training data

We constructed five types of models using different ensemble methods namely, AdaBoost, bagging, gradient boosting, random forest and XGBoost, Table 2 shows the performance results of the ensemble methods using training data. It also shows the performance results of a linear model used in LBT's paper. In Table 2, the cross validation mean is the average of 10 shuffles of LOO cross validation. The smaller the cross validation mean, the better the performance of the model. The other scores are obtained by training with the training data and also measuring performance with the training data. If a certain model's MAE is closer to '0' and Explained Variance Score (EX) is closer to '1,' the model is performing better. The former (MAE) shows how accurate the model can predict the actual value and the latter (EX) shows how certain the predictions are.

Table 2: Performance Results of Models Using the Training Data

	Linear	Bagging	Random Forest	Adaboost	Gradient boosting	XGBoost (DART)
Cross Validation mean	5.40	3.98	4.10	3.08	3.61	4.66
MAE mean	3.61	2.36	2.05	0.51	0.33	0.15
EX mean	0.67	0.83	0.86	0.97	0.97	1.00
MAE max	4.22	3.59	3.61	1.60	1.48	0.72
EX min	0.53	0.66	0.56	0.91	0.75	0.99

Table 2 displays performance results of non-linear models using ensemble methods and the linear model used in LBT. In all categories of performance evaluation, non-linear models outperformed the linear model. Among the non-liner models using ensemble learning, the sequential type AdaBoost and gradient boosting showed superior performance. We believe the strong tendency of sequential construction methods to excessively adapt to the learning data resulted in better performance.

3.3 Generalization performance

To test generalization performance, we used the testing data to test our five models and the linear model of LBT. Table 3 shows the performance results. Non-linear models showed improved performance over the linear model in most of the categories. Among the five models, random forest for the parallel type ensemble method and gradient boosting for sequential type ensemble method in particular, showed superior results than the linear regression model. The MAE mean improved by 0.5 and 0.7, and the EX mean improved about 0.1. However, as for MAE max, all the non-linear models' performance worsened compare to the linear regression model.

Table 3: Performance Results of Generalization Performance

	Linear	Bagging	Random Forest	Adaboost	Gradient boosting	XGBoost (DART)
MAE mean	6.08	5.76	5.58	6.05	5.32	5.85
EX mean	0.28	0.38	0.38	0.28	0.40	0.31
MAE max	9.09	10.60	9.68	10.74	9.94	10.51
EX min	-2.69	-0.73	-0.44	-0.44	-0.08	-0.06

IV. Discussion: Importance of Explanatory Variables

To further investigate how our non-linear models are structually functioning, in this section, by using the random forest model for an example, we examined how important the three explanatory variables of the model are in predicting the forecasting results. Explanatory variables with high importance are drivers of the forecasting and their values have significant influence on the forcasting results. In contrast, explanatory variables with lower importance could be discarded from modeling.

There are two measures of variable importance in random forest. One is permutation importance and the other is Gini importance. In permutation importance method, at the time of

shuffles, variable importance is calculated by how much accuracy decreases when each variable is excluded.[5] In Gini importance method, Gini-impurity/Variance is used to decide which variable to split at each node in tree creation process. For each variable, the sum of the Gini-impurity/Variance decrease over all the trees of the random forest is calculated. The average score of it, the sum of Gini-impurity decrease divided by the number of trees, is variable importance value (Louppe et al 2013). We used Gini importance score in this analysis which is the default method for python package scikit-learn .The descriptive statistics of variable importance are displayed in Table 4 and Figure 1. Consistent with LBT's linear model, all three variables had significant effect in forecasting results. Compared to the linear model, however, "Days (number of days after the previous election)" had more substantial effect than the linear model and "GDP" had less effects.[6]

Table 4: Descriptive Statistics of Explanatory Variable Importance

Feature importance	GDP	PM approval	Days
mean	0.230	0.227	0.543
sd	0.059	0.095	0.125
min	0.153	0.036	0.394
max	0.318	0.395	0.811

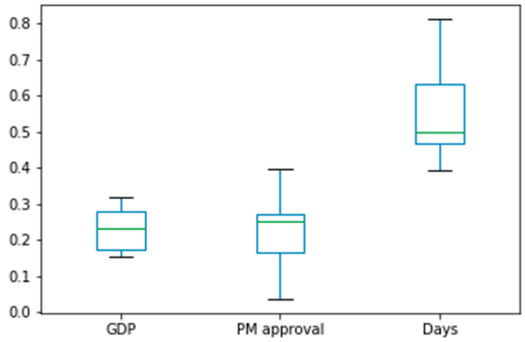

Figure 1: Descriptive Statistics of Explanatory Variable Importance

[5] This is a standard procedure often used in the field of machine learning. By going through this process, one can examine which variables are drivers of outcome and which are not.

[6] Since variable importance examines relative importance among explanatory variables, it takes values between 0 to 1 and the sign is always positive. "Days," however, affects the outcome negatively in this case.

Overall, the random forest model, one of our five models, seems to be structured properly and functioning well. There remains rooms for improvement, however. For example, as for the importance of explanatory variables, why did the results differ from LBT? Since this paper focuses on the methodological aspects of the Japanese electoral forecasting model, we will leave these substantive questions to be addressed in the future.

V. Conclusion

In this paper, we created non-linear forecasting models of Japanese elections using decision trees and ensemble learning methods. To assess methodological advantage over LBT's pioneering model of Japanese electoral forecasting, we replicated substantive theory and data of LBT's linear model. All of our non-linear models showed improved performance over LBT's linear model in almost all the evaluation categories. Despite small sample size inherent for country-specific forecasting models, we were able to develop non-linear forecasting models of Japanese election that generate better performance than the linear model.

Our study can be extended to several directions. First, by improving substantive theory of our electoral forecasting models, we can further improve the performance of the models. To do so, we need to examine both institutional and behavioral aspects of Japanese politics. Second, our methodological approach can be applied to, by combining with substantive theory of each country, other country-specific forecasting models. In fact, one of the aims of LBT was to show theoretical compatibility of their model across borders. Third, to overcome small sample size, we can utilize state of art data augmentation methods such as Generative Adversarial Networks (GANs) (Ahsan et al. 2019) to increase sample size of country-specific forecasting models.

References

Ahsan, Budrul, Sota Kato, Takafumi. Nakanishi, Hirokazu Shimauchi (2019) Augmenting Political Data through Generative Adversarial Networks (GANs), *APSA 2019 Annual Meeting Paper*.

Arrow, Kenneth J., Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O. Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D. Nelson, George R. Neumann, Marco Ottaviani, Thomas C. Schelling, Robert J. Shiller, Vernon L. Smith, Erik Snowberg, Cass R. Sunstein, Paul C. Tetlock, Philip E. Tetlock, Hal R. Varian, Justin Wolfers and Eric Zitzewitz (2008) The promise of prediction markets, *Science* 320 (5878), 877-878.

Breiman, Leo (1996) Bagging predictors, *Machine Learning* 24, 123-140.

Breiman, Leo (2001) Random forests, *Machine Learning* 45, 5-32.

Caruana, Rich. and Alexandru Niculescu-Mizil, A. (2006) An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, June 2006, 161-168. http://dx.doi.org/10.1145/1143844.1143865

Chen, Tianqi, Sameer Singh, Ben Taskar and Carlos Guestrin (2015) Efficient second-order gradient boosting for conditional random fields. *In Proceeding of 18th Artificial Intelligence and Statistics Conference (AISTATS'15)* volume 1.

Fair, Ray C. (1978) The effect of economic events on votes for president, *The Review of Economics and Statistics* 60 (2), 159-173.

Freund, Yoav and Robert E. Schapire (1997) A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences* 55 (1), 119-139.

Friedman, Jerome H. (2001) Greedy function approximation: A gradient boosting machine, *The Annals of Statistics* 29 (5), 1189-1232.

Hirano, hiroshi (2007) henyō suru nihon no shakai to tōhyō-kōdō (Changes in the voting behavior of Japanese society), *Bokutakusha*.

Hirano, hiroshi (2012) Seiken-kōtai zengo ni okeru yūkensha no keizai-tōhyō: JES IV cyōsa dēta no bunseki kara (Voters' economic voting before and after the change of government: an analysis of JES IV survey data, Cyūō Cyōsa Hō 659, 1-5.

Inoguchi, Takashi (1981) Explaining and predicting Japanese general elections, 1960-1980, *The Journal of Japanese Studies* 7 (2), 285-318.

Kato, Sota and Masayuki Inui (2013) How valuable is Prime Minister's dissolution option? Black-Scholes approach to parliamentary dissolution. *APSA 2013 Annual Meeting Paper*. SSRN: https://ssrn.com/abstract=2303520.

Kennedy Ryan, Stefan Wojcik and David Lazer (2017) Improving election prediction internationally, *Science* 355(6324), 515-520.

Lewis-Beck, Michael S. and Charles Tien (2011). Election forecasting. In Michel P. Clements and David F. Hendry (Eds.), *The Oxford handbook of economic forecasting*, Chapter 24.

Lewis-Beck, Michael S. and Charles Tien (2012) Japanese election forecasting: Classic tests of a hard case, *International Journal of Forecasting* 28 (4), 797-803.

Lewis-Beck, Michael S. and Tom W. Rice (1984) Forecasting presidential elections: A comparison of Naive models, *Political Behavior* 6 (1), 9-21.

Louppe, Gilles, Louis Wehenkel, Antonio Sutera, Pierre Geurts (2013) Understanding variable importances in forests of randomized trees, In Christopher J. C. Burges, Léon Bottou, Max Welling, Zoubin Ghahramani, Kilian Q. Weinberger, editors, Advances in Neural Information Processing Systems, 431-439.

Nasuno, Kaoru, Shojiro Okuyama, Kyoko Nakanishi and Yutaka Matsuo (2015) Twitter ni okeru kōhosya no senkyo-jiban ni chakumoku sita kokusei-senkyo no tōsensya-yosoku (Predicting Japanese general election by focusing on candidates' constituency on Twitter), *IPSJ Journal* 56 (10), 2044-2053.

Seni, Giovanni and John F. Elder (2010) *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. Synthesis Lectures on Data Mining and Knowledge Discovery*, Morgan & Claypool Publishers.

Sigelman, Lee (1979) Presidential popularity and presidential elections, *The Public Opinion Quarterly* 43 (4), 532-534.

Wolpert, David H. and William. G. Macready (1995) No free lunch theorems for search, Technical Report SFI-TR-95-02-010, Santa Fe Institute.

Zhou, Qi-Feng, Hao Zhou, Yong-Peng Ning, Fan Yang, and Tao Li (2015) Two approaches for novelty detection using random forest, Expert Systems with Applications 42(10), pp. 4840-4850.

Zhou, Zhi-Hua (2012) *Ensemble Methods: Foundations and Algorithms*, Chapman and Hall/CRC.