IJAMEC
e-ISSN: 2147-8228

INTERNATIONAL JOURNAL OF APPLIED MATHEMATICS ELECTRONICS AND COMPUTERS

www.dergipark.org.tr/ijamec

International Open Access

Volume 08
Issue 04

December, 2020

Research Article

Analysis and detection of Titanic survivors using generalized linear models and decision tree algorithm

Burcu Durmuş [a], ***Öznur İşçi Güneri*** [a,*]

[a] *Mugla Sitki Kocman University, Rectorate Unit, Kotekli Campus, Mugla, Turkey*
[b] *Mugla Sitki Kocman University, Faculty of Science, Department of Statistics, Kotekli Campus, Mugla, Turkey*

ARTICLE INFO

Article history:
Received 25 August 2020
Accepted 9 October 2020

Keywords:
Decision tree,
Generalized linear models,
Logit regression,
Probit regression,
Random tree.

ABSTRACT

In the article, it is aimed to investigate the factors affecting survival in today's legendary giant accident with different methods. The analysis aims to find the method that best determines survival. For this purpose, logit and probit models from generalized linear models and random tree algorithm from decision tree methods were used. The study was carried out in two stages. Firstly; in the analysis made with generalized linear models, variables that did not contribute significantly to the model were determined. Classification accuracy was found to be 79.89% for the logit model and 79.04% for the probit model. In the second stage; classification analysis was performed with random tree decision trees. Classification accuracy was determined to be 77.21%. In addition; according to the results obtained from the generalized linear models, the classification analysis was repeated by removing the data that made meaningless contribution to the model. The classification rate increased by 4.36% and reached 81.57%. After all; It was determined that the decision tree analysis made with the variables extracted from the model gave better results than the analysis made with the original variables. These results are thought to be useful for researchers working on classification analysis. In addition, the results can be used for purposes such as data preprocessing, data cleaning.

1. Introduction

Titanic is a world-famous cruise ship that sank on its first voyage in the North Atlantic [1]. There is a lot of speculation in the literature about the legendary Titanic disaster, and research on this is still ongoing [2-3]. Over the years, a dataset containing information about survivors as well as dead passengers and crew has been created [4]. This data set is publicly available on Kaggle.com [5].

When the literature is reviewed, it stands out that the Titanic data have been examined for different purposes in recent years. In Barhoom et al.'s study, the prediction of survivors was determined by artificial neural networks. The algorithm has achieved 99.28% accuracy [6]. Singh et al. studied Titanic data on logistic regression, decision tree, decision tree with hypertuning, k-nearest neighbors and support vector machines. At the end of the study, they obtained the highest estimation with decision trees as 93.6% [7]. Kakde et al., on the other hand, performed the analysis with logistic regression, decision tree, random forest and support vector machines methods using data cleaning. They suggested that ideally the logistic regression and support vector machine gives a good level of accuracy when it comes to the classification problem [8]. In another study, Kshirsagar et al. showed that Titanic survivors could be predicted by logistic regression with 95% accuracy [4].

With the development of technology, data collection and storage has become quite easy. As a result, it became more important to discover new methods for analyzing data. Much progress has been made in this area in recent years. Important steps have been taken, especially in the field of data mining. Many new algorithms have been introduced and existing algorithms have also been improved. As a result of these developments, reaching different and new results by analyzing the data with different methods has become the goal, as is the case with

* Corresponding author. E-mail address: *oznur.isci@mu.edu.tr*
DOI: 10.18100/ijamec.785297

the researchers working on Titanic data.

In this study, unlike the literature, Titanic data were analyzed using Random Tree algorithm and generalized linear models. The main purpose of the study is to determine the characteristics of survivors of titanic disaster using different methods. In this direction, logit and probit regression models, which are generalized linear models, were examined in the first stage. At this stage, firstly, the significance test was applied to the data and variables that contributed significantly to the model were included in the analysis. In the second stage, analysis was made with the Random Tree algorithm, which is the decision tree learning algorithm. In order to increase the success of the model, random tree classification analysis was repeated with variables that significantly contributed to the model. The study was completed by comparing the results.

2. Methods and Material

In this study, logit and probit models from the generalized family of linear models and decision tree from data mining methods are discussed.

2.1. Titanic Dataset

The dataset contains variables given in Table 1. However, it was determined by binary logit and probit analysis that some of these variables (sibsp, parch, embarked) did not make a significant contribution to the model. Therefore, it has been removed from the dataset. The remaining variables were included in the logit and probit model as categorical data. Descriptive statistical analysis was done with variables in SPSS 22.0 packages program. In the continuation of the study, binary logistics and binary probit analyses were performed with Stata 11.0 program and decision tree classification analysis was performed. Decision tree analysis was done with both the remaining variables of the data set and the original version of the data set.

Table 1. Variables in the dataset

Variables	Definition
survived	no:0, yes:1
pclass	passenger class (1, 2, 3)
sex	female, male
age	age
sibsp	number of siblings or spouses aboard
parch	number of parents or children aboard
fare	passenger fare
embarked	port of embarkation

Binary logit and probit regression analyses were made by determining indicator variables. Indicator variables: pclass1, male, age0 (children), fare0.

2.2. Generalized Linear Models

Generalized linear models are obtained by extending the linear models due to the assumption distortions [9]. In many fields, these models are used if the data is categorical or discontinuous [10]. Generalized linear models consist of random component, systematic component and link function. The link function determines the name of the model used. If a logit link is used, the name of the model is called the logit regression model [11]. In this study, logit and probit models are discussed.

2.2.1. Logit Regression

If the canonical bond used in generalized linear models is logit, the model is logit regression [12]. Logistic regression is independent variables when the dependent variable is categorical, binary or multiple. In logit regression, there is no assumption of normality and continuity [13]. Therefore, it can be said to be more flexible than linear models.

The logit model is derived from the cumulative distribution function given by Equation 1 [14].

$$P_i = E(Y = \frac{1}{x_i}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} \quad (1)$$

In this model, Pi provides information about the argument xi while the first individual expresses the probability of making a particular choice [15]. Thus, Pi also takes values between "0" and "1" [16]. When the rate of realization of an event is divided by the rate of the event not realized, the odds ratio is obtained [17].

It becomes linear when the odds rate logarithm is taken. In this case, the model is called logit and the equation given by Equation 2 is called the logit link function.

$$L_i = log(\frac{P_i}{1 - P_i}) \quad (2)$$

There are 3 basic methods in logistic regression analysis:

- Binary
- Ordinal
- Nominal

2.2.2. Probit Regression

This model like the logit model is a model that ensures that the probabilities remain between 0 and 1. The probit model assumes that the dependent variable is normally distributed. Therefore, the graph of the logit model is wider than that of the probit model (Fig. 1). Logit and probit models can be compared with a coefficient proposed by Amemiya [18].

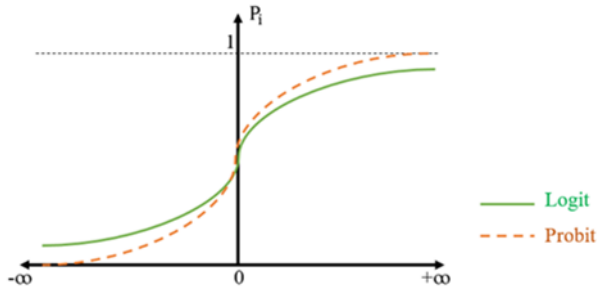

Figure 1. Logit and Probit distributions

When the error distribution is the standard normal cumulative distribution, the probit bond function is used

and the model is called the probit model [19]. The probit link function is defined by Equation 3.

$$Z = \varphi^{-1}(\mu) = \sum_{k=1}^{K} b_k x_k \quad (3)$$

Here φ^{-1} denotes the inverse of the standard normal distribution, b_k is the coefficient estimates and x_k is the explanatory variables. u to show the error for each eye; the standard cumulative distribution function is given by Equation 4.

$$\varphi(Z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} exp(\frac{-u^2}{2}) du \quad (4)$$

2.3. Random Tree Algorithm

Due to its many advantages, decision tree learning is often used in data mining studies [20]. A tree structure is created in decision tree learning. The tree starts from the root node and from there the structure is divided into inner nodes. The root node can be considered as the most determining feature of the differences between the data. It is divided into internal nodes after a series of operations applied to the data set. Each node is such that it can be divided into multiple internal nodes. The leaf node is reached by controlling all internal nodes. The leaf knot is where the decision is made. Each transition between nodes depends on a condition. The condition mentioned here is the theory on which the chosen algorithm is based [21-22]. Decision trees are very advantageous for reasons such as low calculation cost and ease of understanding. For this reason, as mentioned at the beginning, it is preferred in many data mining and especially classification studies [23-24].

Random tree algorithm is a method in which multiple decision trees are created [25]. Algorithm steps:

- The feature that provides the best classification is selected and the starting node is created.
- A training set is formed with a part of the data set. The remaining data is the test set.
- Trees are created with the number of variables to be used in each node and the numbers of trees in N. Variables are selected randomly at each node.
- When N trees are produced, the model is completed and the class of the new member is estimated [25-26].

2.4. Confusion Matrix

Confusion matrix is an analysis tool that explains correctly classified observations and incorrectly classified observations. The confusion matrix is the state of a data set and the number of correct and incorrect predictions of our classification model converted into a table. The general form of the mess matrix is given in Table 2.

Table 2. General confusion matrix

Actual Class	Predicted Class	
	Positives	Negatives
Positives	TP (True Positives)	FN (False Negatives)
Negatives	FP (False Positives)	TN (True Negatives)

3. Results

3.1. Descriptive Statistics

When the relationship between the variables of survived and sex is examined in Table 3, it is seen that 359 people died and 93 survived, 64 women died and 195 survived.

In another comment; of the 423 people who died, 359 were men and 64 were women. Similarly, out of the 288 survivors in the accident, 93 are men and 195 are women.

Table 3. Relationship between survived and sex variables

Survived	**Sex**		Total
	0	1	
0	359	64	423
1	93	195	288
Total	452	259	711

In addition, it can be said that there is a significant agreement between the survived and sex variables in Table 4.

Table 4. Harmony between survived and sex variables

	Value	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	204.540	0.00	0.00	
Likelihood Ratio	210.756	0.00		
Fisher's Exact Test				

When the relationship between survived and fare variables is examined in Table 5, it is seen that 286 of 429 people who paid a low fare died and 143 lived; of the 282 people who paid a high fare, 137 survived and 145 died.

Table 5. Relationship between survived and fare variables

Survived	**Fare**		Total
	0	1	
0	286	137	423
1	143	145	288
Total	429	282	711

It can be said that there is a statistically significant ($p=0.00<0.05$) relationship between survival and fare variables in Table 6.

Table 6. Harmony between survived and fare variables

	Value	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	23.093	0.00	0.00	
Likelihood Ratio	23.029	0.00		
Fisher's Exact Test				

3.2. Logit Regression Results

When Table 7 is examined, it can be said that the predicted model is significant at 5% error level since p=0.00<0.05.

When the significance values for the variables are examined, it is seen that all the variables have a significant contribution to the model (Those who did not have any meaningful contribution were already removed).

Table 7. Results of binary logit regression model

		Number of obs		=	711	
		LR Chi2 (8)		=	317.89	
		Prob > chi2		=	0.000	
		Pseudo R2		=	0.3312	
Survived	Odds Ratio	Std. Err.	z	P > |z|	[95% Conf. Interval]	
pclass2	0.253	0.074	-4.67	0.000	0.14	0.45
pclass3	0.066	0.021	-8.46	0.000	0.03	0.12
female	12.90	2.732	12.07	0.000	8.51	19.5
age1	0.217	0.099	-3.32	0.000	0.08	0.53
age2	0.225	0.082	-4.09	0.000	0.11	0.46
age3	0.204	0.075	-4.29	0.000	0.09	0.42
age4	0.090	0.044	-4.92	0.000	0.03	0.23
fare1	0.610	0.150	-2.01	0.045	0.37	0.98
_cons	6.561	3.160	3.91	0.000	2.55	16.8

It can be said that there is a statistically significant (p=0.00<0.05) relationship between survival and fare variables in Table 5.

Odds ratio is interpreted by reversing. Comments on odds ratios are as follows:

- Those in 2nd class are 6.55 times more likely to survive than those in 1st class.
- Those in the 1st class are 4 times more likely to survive than those in the 3rd class.
- Men are 12.80 times more likely to survive than women.
- Children are 4.76 times more likely to survive than the age-1 group.
- Children are 4.54 times more likely to survive than the age-2 group.
- Children are 5 times more likely to survive than the age-3 group.
- Children are 11.11 times more likely to survive than the age4 group.
- Low-payers are 1.63 times more likely to survive than high-payers.

Probability values for the data are calculated by Equation 4. Some sample probability values are given with Equations 5-6.

$$p = 1.88 - 1.37 * class2 - 2.70 * class3 + 2.55 * famale - 1.52 * age1 - 1.48age2 - 1.58 * age3 - 2.40 * age4 - 0.49fare1 \quad (4)$$

The probability of survival for a 1st class, woman, child, high fare person is $p = 0.98$.

The survival probability of a 3rd class, male, age2 group, low fare person is $p = 0.09$.

The marginal effect is the effect that a small change in the independent variable will cause in the dependent variable.

For the logit model given in Table 8, while the effect of other variables is fixed, 1 unit increase in age-1 variable decreases survival by an average of 0.22 units. This result is in line with the results of odds ratio.

Table 8. Marginal effect of binary logit regression model

Average marginal effects Model VCE: OIM Expression: Pr(Survived), predict () dy/dx w.r.t.: age1						
	dy/dx	Std. Err.	z	P > |z|	[95% Conf. Interval]	
age1	-0.22	0.065	-3.40	0.001	-0.34	-0.09

The classification results for the logit regression model are given in Table 9. The model performs with the classification accuracy of 79.89%.

Table 9. Classification results of binary logit regression model

Actual Value	Predicted Value		Total Actual
	0	1	
0	207	62	269
1	81	361	442
Total	288	423	711
Classification Accuracy: 79.89%			

3.3. Probit Regression Results

According to Table 10 the model is significant since it is p=0.00<0.05. At least one variable has an effect on the model. Coefficients are also important except for the fare1 variable.

Table 10. Results of binary logit regression model

		Number of obs	=	711	
		LR Chi2 (8)	=	314.42	
		Prob > chi2	=	0.000	
		Pseudo R2	=	0.3276	
Survived	Coef.	Std. Err.	z	P > |z|	[95% Conf. Interval]	
pclass2	-0.78	0.16	-4.61	0.000	-1.11	-0.45
pclass3	-1.50	0.17	-8.66	0.000	-1.84	-1.16
female	1.49	0.11	12.64	0.000	1.26	1.72
age1	-0.81	0.25	-3.16	0.000	-1.32	-0.30
age2	-0.77	0.19	-3.88	0.000	-1.16	-0.38
age3	-0.82	0.20	-4.07	0.000	-1.22	-0.42
age4	-1.31	0.27	-4.79	0.000	-1.85	-0.77
fare1	-0.25	0.13	-1.84	0.066	-0.52	0.01
_cons	0.96	0.26	3.64	0.000	0.44	1.49

The classification results for the probit regression model are given in Table 11. The model performs with the classification accuracy of 79.04%.

Table 11. Classification results of binary probit regression model

Actual Value	Predicted Value		Total Actual
	0	1	
0	201	62	263
1	87	361	448
Total	288	423	711
Classification Accuracy: 79.04%			

The results obtained with the probit model are parallel with the results obtained with the logit model.

3.4. Random Tree Algorithm Results

The random tree algorithm is discussed for logit and probit models and variables that contribute significantly to the model among the original variables. A section of the decision tree obtained by the random forest algorithm is given in Table 12. Looking at the structure of the tree, it can be seen how the variables affect survival.

Table 12. A section from the decision tree

```
sex = male
|  age = 0
|  |  pclass = 1 : 1 (3/0)
|  |  pclass = 2 : 1 (9/0)
|  |  pclass = 3
|  |  |  fare = 0 : 1 (9/1)
|  |  |  fare = 1 : 0 (15/1)
|  |  |  fare = 2 : 0 (0/0)
|  age = 1
|  |  pclass = 1 : 0 (2/1)
|  |  pclass = 2 : 0 (6/0)
|  |  pclass = 3
|  |  |  fare = 0 : 0 (22/2)
|  |  |  fare = 1 : 0 (4/0)
|  |  |  fare = 2 : 0 (0/0)
|  age = 2
|  |  pclass = 1
|  |  |  fare = 0 : 0 (0/0)

|  |  |  fare = 1 : 1 (17/8)
```

The classification results of the algorithm are given in Table 13. The classification result is more successful with about 2.5% compared to the result obtained in logit and probit models.

Table 13. Confusion matrix of dataset

Actual Value	Predicted Value		Total Actual
	0	1	
0	383	40	423
1	91	197	288
Total	474	237	711
Classification Accuracy: 81.57%			

A section from the decision tree of the original data set is presented in Table 14. All variables in the dataset are included in the decision tree. It is also seen that this tree structure starts with the sex variable as in Table 12.

Table 14. A section from the decision tree (for original dataset)

```
sex = male
|  fare = 0
|  |  embarked = C
|  |  |  age < 29.5
|  |  |  |  age < 5.5 : 1 (1/0)
|  |  |  |  age >= 5.5
|  |  |  |  |  sibsp = 0
|  |  |  |  |  |  pclass = 1 : 0 (0/0)
|  |  |  |  |  |  pclass = 2 : 0 (1/0)
|  |  |  |  |  |  pclass = 3
|  |  |  |  |  |  |  age < 25.5
|  |  |  |  |  |  |  |  age < 23
|  |  |  |  |  |  |  |  |  age < 15.5 : 0 (1/0)
|  |  |  |  |  |  |  |  |  age >= 15.5 : 0 (4/2)

|  |  |  |  |  |  |  |  age >= 23 : 0 (2/0)
```

The classification result of the original data set is shown in Table 15. The classification accuracy according to this table is 77.21%. This result is lower than the classification accuracy results obtained in Table 12.

Table 15. Confusion matrix of original dataset

Actual Value	Predicted Value		Total Actual
	0	1	
0	356	67	423
1	95	193	288
Total	451	260	711
Classification Accuracy: 77.21%			

4. Conclusion

In this study, estimation of survivors of titanic accident with different methods was investigated. Factors affecting survival were researched and survival rate was estimated by classification method.

In the first stage, logit and probit regression analyses were performed. With these analyses, variables that contribute significantly to survival were determined and the classification accuracy were found to be 79.89% and 79.04% respectively. In the second stage, two different analyses were done with the random tree algorithm. In the first analysis, variables used in logit and probit regressions that make a significant contribution to the model were used. Classification accuracy was found as 81.57%. The second analysis was done with the variables in the original data set and the classification accuracy fell to 77.21%.

When all the results are considered together, it is best to estimate the data that contributes significantly to the model with decision trees.

The study results reveal that, in addition to the expected results, doing decision tree analysis (data mining or machine learning analysis) with data that contributes significantly to the model yields more successful results. These results emphasize that decision-tree learning methods based on new technologies are more successful, but the results can still be enhanced by statistical methods.

References

[1] E. L. Rasor, "The Titanic: Historiography and Annotated Bibliography". Greenwood Publishing Group, London, 2001.

[2] A. Singh, S. Saraswat, N. Faujdar, "Analyzing Titanic Disaster using Machine Learning". International Conference on Computing, Communication and Automation, pp. 406-411, 2017.

[3] C. Dieckmann, "The Mystery of the Titanic: What Really Happened". Undergraduate Research Journal, vol. 13(1), pp. 243-248, 2020.

[4] V. Kshirsagar, N. Phalke, "Titanic Survival Analysis using Logistic Regression". International Research Journal of Engineering and Technology, vol. 6(8), pp. 89-91, 2019.

[5] Kaggle.com, 'Titanic Data Set', http://www.kaggle.com/, Accessed: Oct. 2020.

[6] A. M. Barhoom, A. J. Khalil, B. S. Abu-Nasser, M. M. Musleh, S. S. Abu-Naser, "Predicting Titanic Survivors using Artificial Neural Network". International Journal of Academic Engineering Research, vol. 3(9), pp. 8-12, 2019.

[7] K. Singh, R. Nagpal, R. Sehgal, "Exploratory Data Analysis and Machine Learning on Titanic Disaster Datase". 10th International Conference on Cloud Computing, Data Science & Engineerin, India, Jan. 2020.

[8] Y. Kakde, Agrawal, S., "Predicting Survival on Titanic by Applying Exploratory Data Analytics and Machine Learning Techniques", International Journal of Computer Applications, vol. 179(44), pp. 32-38, 2018.

[9] J. Garrido, J. Zhou, "Full Credibility with Generalized Linear and Mixed Models". ASTIN Bulletin, vol. 39(1), pp. 61-80, 2009.

[10] T. Koc, M. A. Cengiz, "Genelleştirilmiş Lineer Karma Modellerde Tahmin Yöntemlerinin Uygulamalı Karşılaştırılması". Karaelmas Science and Engineering Journal, vol. 2(2), pp. 47-52, 2012.

[11] Y. Kida, "Generalized Linear Models: Introduction to Advanced Statistical Modeling". Towards Data Science, Sep. 2019.

[12] B. Bozkurt, "Kredi ve Yurtlar Kurumunda Kalan Öğrencilerin Memnuniyet Derecelerinin Lojistik Regresyon Yöntemi ile Araştırılması: Edirne Ili Örneği". University of Trakya Social Sciences Institute Business Department Master Term Project, Aug. 2011.

[13] G. Çırak, Ö. Çokluk, "The Usage of Artifical Neural Network and Logistic Regresssion Methods in the Classification of Student Achievement in Higher Education". Mediterranean Journal of Humanities, vol. 3(2), pp. 71-79, 2013.

[14] D. N. Gujarati, N. C. Porter, "Temel Ekonometri". Ümit Şenesen ve Gülay Günlük Şenesen (çev.) İkinci Basım, Literatür Yayıncılık, İst. 2001.

[15] Ö. İ. Güneri, B. Durmuş, "Dependent Dummy Variable Models: An Application of Logit, Probit and Tobit Models on Survey Data". International Journal of Computational and Experimental Science and Engineering, vol. 6(1), pp. 63-74, 2020.

[16] M. Bilki, Ü. Aydın, "Konut Sahibi Olma Kararlarını Etkileyen Faktörler: Lojistik Regresyon ve Destek Vektör Makinelerinin Karşılaştırılması". Dumlupınar Üniversitesi Sosyal Bilimler Dergisi, vol. 62, pp. 184-199, 2019.

[17] S. Demirci, M. Astar, "Türkiye'de Özel Sigortayı Etkileyen Faktörler: Logit Modeli". Trakya Üniversitesi Sosyal Bilimler Dergisi, vol. 13 (2), pp. 119-130, Dec. 2011.

[18] T. Amemiya, "Qualitative Response Models: A Survey". Journal of Economic Literature, vol. 19(4), pp. 481-536, 1981.

[19] J. H. Aldric, F. D. Nelson, "Linear Probability, Logit and Probit Models", Sage Publications, USA, 1984.

[20] D. Bertsimas, J. Dunn, "Optimal Classification Trees". Mach Learn, vol. 106, pp. 1039–1082, 2017.

[21] J. Ali, R. Khan, N. Ahmad, L. Maqsood, "Random Forests and Decision Trees". International Journal of Computer Science Issues, vol. 9, pp. 5-3, Sep. 2012.

[22] G. Nuti, L. A. J. Rugama, "A Bayesian Decision Tree Algorithm". arXiv:1901.03214v2 [stat.ML], Jan. 2019.

[23] B. Gupta, A. Rawat, A. Jain, A. Arora, R. Dhami, "Analysis of Various Decision Tree Algorithms for Classification in Data Mining". International Journal of Computer Applications, vol. 163 (8), pp. 15-19, Apr. 2017.

[24] S. D. Jadhav, H. P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques". International Journal of Science and Research, vol. 5 (1), pp. 1842-1845, Jan. 2016.

[25] B. Durmus, Ö. İ. Güneri, "Data Mining with R: An Applied Study". International Journal of Computing Sciences Research, vol. 3(3), pp. 201-216, 2019.

[26] Ö. Akar, O. Güngör, "Rastgele Orman Algoritması Kullanılarak Çok Bantlı Görüntülerin Sınıflandırılması". Journal of Geodesy and Geoinformation, vol. 1(2), pp. 139-146, 2012.