Hybrid Model of Random Forests and Genetic Algorithms for Commute Mode Choice Analysis

Hironobu HASEGAWA[a], Mikiharu ARIMURA[b], Tohru TAMURA[c]

[a]*Department of Civil and Environmental Engineering, Akita National College of Technology, 1-1 Iijimabunkyocho, Akita City, Akita 011-8585, JAPAN ; E-mail: hasegawa@ipc.akita-nct.ac.jp*
[b]*Department of Civil Engineering and Architecture, Muroran Institute of Technology, 27-1 Mizumoto, Muroran City, 050-8585, JAPAN ; E-mail: arimura@mmm.muroran-it.ac.jp*
[c]*Graduate School of Engineering, Hokkaido University, Kita 13 Nishi 8, Kita-ku, Sapporo City, Hokkaido, 060-8628, JAPAN ; E-mail: tamura-t@eng.hokudai.ac.jp*

Abstract: Accurate commute mode choice analysis is essential in transportation planning in urban areas. This study proposes a hybrid model of random forests with genetic algorithms (RFGA model) for commute mode choice analysis. The random forest (RF) model is one of the most efficient methods for classification and regression and a typical ensemble learning method based on the decision tree. We propose a practical method for optimizing the parameters of the RF model by metaheuristic optimization using genetic algorithms. This model is compared with conventional methods, i.e., normal RF model and multinomial logit (MNL) model. This demonstrates that the RFGA model has higher performance of classification than other models, thus establishing the efficiency of this model.
Keywords: Commute Mode Choice, Travel Behaviour Analysis, Random Forests, Genetic Algorithms, Optimization, Machine Learning

1. INTRODUCTION

Accurate commute mode choice analysis is essential in transportation planning; therefore, much research has been conducted in this respect. Typically, discrete choice models based on the random utility maximization theory were adopted to analyze commute mode choice behavior (e.g., Ben-Akiva and Lerman (1985); Committee of Infrastructure Planning and Management (1995); Kitamura et al. (2002)). Understanding of behavioral principle and a few sample size requirements to estimate models are considerable advantages of discrete choice models.

On the other hand, a lot of machine learning techniques for classification are being developed and improved in the computation research field in recent years (Bishop (2007)). Not only commute mode choice models but also other choice behavior models can be regarded as the classification problem and thus, can get the benefit from rapid progress of machine learning techniques. Nevertheless, the application of machine learning techniques to choice behavior models lacks its accumulation. For example, there are 11 issues of Transportation Research Record featured about travel behavior and available via online, however, only two papers (Lu et al. (2008); Lu and Kawamura (2010)) matched the search query 'machine learning'. If analysts have enough amount of data and regard not understanding of behavioral principle but the accuracy as important, machine learning techniques are considerable and competitive options.

This study is one of the application of machine learning techniques to choice behavior models and proposes a hybrid model of random forests with genetic algorithms (RFGA model) for commute mode choice analysis. The household travel survey collected for the centre of Hokkaido Prefecture, Japan, in 2006 has been used to investigate the commute mode choice analysis by using an ensemble learning technique. Our study reveals that the generalization performance would be more accurate as a result of optimizing the random forest (RF) model parameter. Furthermore, compared to our previous RF model for commute mode choice analysis

(Hasegawa et al. (2012)), the present RFGA model offers improved generalization performance. Therefore, the study indicates that the RFGA model performs much better than conventional models.

The rest of the paper is organized as follows: The data for commute mode analysis are explained in Section 2. The evaluation method of commute mode choice analysis and each of the models are explained in Section 3. The analysis results and discussion are explained in Section 4. Finally, a summary is provided in Section 5.

2. DATA

2.1 Data Source

The data source used in this study is the household travel survey collected for the centre of Hokkaido Prefecture, Japan, in 2006. Figure 1 shows the map of the survey area and table 1 shows city name. There are 233177 records (trips) and 162 attributes in the original data.

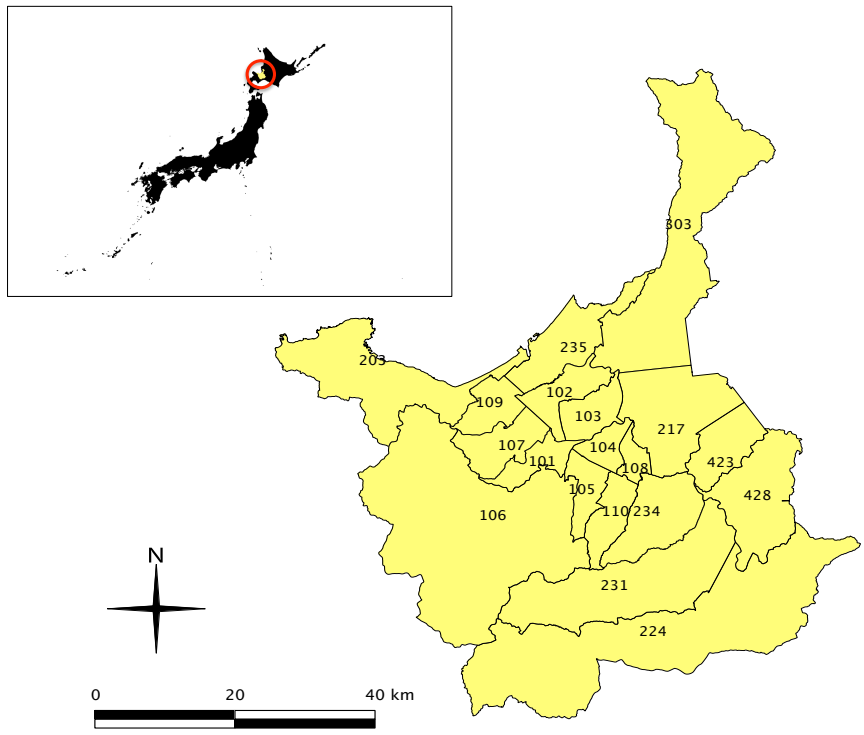

Figure 1. Survey area

2.2 Data Preparation

The data preparation processes are as follows:

1) Collect sample records meeting the following conditions:
   - Commute trips
   - First trip of trip chain
   - Departing from home
   - Completed by noon

Table 1. Survey area

No.	City name	No.	City name
101	Cyuo-ku (Sapporo)	203	Otaru
102	Kita-ku (Sapporo)	217	Ebetsu
103	Higashi-ku (Sapporo)	224	Chitose
104	Shiroishi-ku (Sapporo)	231	Eniwa
105	Toyohira-ku (Sapporo)	234	Kitahiroshima
106	Minami-ku (Sapporo)	235	Ishikari
107	Nishi-ku (Sapporo)	303	Tobetsu
108	Atsubetsu-ku (Sapporo)	423	Nanporo
109	Teine-ku (Sapporo)	428	Naganuma
110	Kiyota-ku (Sapporo)		

2) Remove mass transit trips where there is
- Use of special tickets for physically handicapped persons
- Use of an one day unlimited ticket

3) Select variables for commute mode choice analysis
- Dependent variable: typical mode of commute trip
  - Automobile
  - Mass transit (bus, train, tram)
  - Other
- Independent variables
  - Age
  - Gender (dummy variable)
  - Licence of auto mobile (dummy variable)
  - Automobile ownership (dummy variable)
  - Travel time
  - Travel cost of automobile
  - Travel cost of mass transit

4) Partition data into training and test datasets
- Training data (50 %): used to construct commute mode choice models
- Test data (50 %): used to validate the performance of the models

As a result of data preparation, a commute trip dataset with one dependent variable, seven independent variables and 33553 records was collected. Table 2 shows the number of trips by mode and use. In the table and hereinafter, auto, mass and other represent automobile, mass transit and other modes, respectively.

Table 2. Commute trips by type of transportation mode and use

	Training data	Test data	Total
Auto	9039	9039	18078
Mass	5157	5157	10314
Other	2580	2581	5161
Total	16776	16777	33553

3. METHODS

3.1 Performance Evaluation Methods

The criteria for evaluating a model are different for each domain. For the probabilistic transportation behaviour model, probability was used. Furthermore, the hit ratio for 'training data' was used to supplement it (Ben-Akiva and Lerman (1985)). However, if the model could classify 'training data' correctly, the use of 'test data' was not warranted. The accuracy in classifying 'test data' is called 'generalization performance' in the computation research field. This criterion is also important for mode choice analysis, which influences decision making for transportation policy.

The decision made by the classifier can be represented by a structure known as a confusion matrix or contingency table (Table 3).

Table 3. Confusion matrix for performance evaluation

	Auto	Mass	Other	Total
Auto	AA	AM	AO	AA+AM+AO
Mass	MA	MM	MO	MA+MM+MO
Other	OA	OM	OO	OA+OM+OO
Total	AA+MA+OA	AM+MM+OM	AO+MO+OO	AA+MA+OA+AM+MM +OM+AO+MO+OO

The confusion matrix has following nine categories:

- 'AA' refers to examples correctly labelled as automobile.
- 'AM' refers to mass transit examples incorrectly labelled as automobile.
- 'AO' refers to other mode examples incorrectly labelled as automobile.
- 'MA' refers to automobile examples incorrectly labelled as mass transit.
- 'MM' refers to examples correctly labelled as mass transit.
- 'MO' refers to other mode examples incorrectly labelled as mass transit.
- 'OA' refers to auto mobile examples incorrectly labelled as other mode.
- 'OM' refers to mass transit examples incorrectly labelled as other mode.
- 'OO' refers to examples correctly labelled as other mode.

There are several criteria that are calculated by the confusion matrix. Precision α, recall k and hit ratio H, the major criteria in the computation research field, are adopted. Precision evaluates the accuracy of the classification results by type of transportation mode, and is calculated

using one of Equations (1), (2) and (3).

$$\alpha_{Auto} = \frac{AA}{AA + AM + AO} \tag{1}$$

$$\alpha_{Mass} = \frac{MM}{MA + MM + MO} \tag{2}$$

$$\alpha_{Other} = \frac{OO}{OA + OM + OO} \tag{3}$$

Recall k evaluates the accuracy in reproducing the observations by type of transportation mode, and is calculated using one of Equations (4), (5) and (6).

$$k_{Auto} = \frac{AA}{AA + MA + OA} \tag{4}$$

$$k_{Mass} = \frac{MM}{AM + MM + OM} \tag{5}$$

$$k_{Other} = \frac{OO}{AO + MO + OO} \tag{6}$$

Hit ratio H evaluates the entire accuracy of the classification results, and is calculated using equation (7)

$$H = \frac{AA + MM + OO}{AA + AM + AO + MA + MM + MO + OA + OM + OO} \tag{7}$$

3.2 Random Forest Model

RF is one of the most efficient methods for classification and regression. It is also recognized as a typical ensemble learning method that is based on classification and regression trees.

The RF algorithm is summarized as follows:

1) Let L be the original dataset with M variables and N records, and let B be the total number of trees in the RF.
2) Growing each tree
   (a) Data sampling: Let L_k be the kth bootstrap sample created by randomly sampling N records with replacement from L. When L_k is drawn by sampling with replacement, about one-third of the cases are left out of the sample. This out-of-bag (OOB) data are used to obtain a running unbiased estimate of the classification error as trees are added to the forest, as well as to obtain estimates of variable importance.
   (b) Growing the tree: The kth tree T_k is growing using L_k set. When, T_k is growing, m variables are randomly selected from the M variable space and the best split on these m variables is used to split the node. The value of m is held constant during the forest growing.
   (c) Estimating the OOB error rate: The OOB error rate is calculated using the result of a majority of votes from each tree by using the OOB data. At each bootstrap iteration, each OOB dataset is used to obtain a OOB classification result. At the end of the run, the class that obtains most of the votes is the OOB predicted result. This is used to calculate the OOB error rate.
   (d) Iterate the above growing steps from $k = 1$ to $k = B$

3) Classifying a new object: Input the variable vector to each of the trees in the forest. Each tree gives a classification result, and we say that the tree 'votes' for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Interested readers may consult the original paper (Breiman (2001)) of RF for details on RF.

In the transportation research field, RF has been adopted primarily for traffic accident analysis. Haleem et al. (2010) has applied RF to select variables of the crash prediction model. Hossain and Muromachi (2011) has used RF to identify important factors and understand the crash mechanism on urban expressways. Pande et al. (2011) has used RF to select variables of the crash risk estimation model. On the other hand, we have applied RF to commute mode choice analysis (Hasegawa et al. (2012)), which has been the basis for the development of this study.

The randomForest package of the R program (R Core Team (2012); Liaw and Wiener (2002)) was used to implement the RF model. In this implementation, there are two adjustable parameters in the RF model. One is the number of variables randomly sampled as candidates at each split of tree m and the other is the number of trees to grow B. The default values, i.e., applied to the RF model, are as follows:

$$m = \sqrt{M} \quad , m \in \mathbf{N} \quad , 1 \leq m < M \tag{8}$$

$$B = 500 \quad , B \in \mathbf{N} \quad , 1 \leq B \tag{9}$$

When the value of m is not an integer, it is rounded to the nearest even number. The parameter m influences the accuracy of RF. Furthermore, the computation time and requirement memories increase in proportion as B increases. Hence, finding a pair of parameters is important for practical use in RF model.

3.3 Hybrid Model of Random Forests and Genetic Algorithms

In this subsection, we present a hybrid RFGA model to predict commute mode choice. As mentioned in the previous subsection, finding a pair of parameters is important for practical use in RFs. However, no definitive method for finding the optimum pair of parameters exists. A simple method is trial and error, but there are many combinations of parameters, and it requires many iterations to evaluate the options.

Therefore, we propose a practical method for optimizing the parameters of RFs by meta-heuristic optimization using genetic algorithms (GAs). GA is an one of the most popular meta-heuristic optimization methods that emulates the evolutionary process of life (Mitchell (1996)). The rgenoud package of the R program (R Core Team (2012); Mebane and Sekhon (2011)) was used to implement the optimizing process of RF parameters m and B. A flowchart of the overall calculating process is shown in Figure 2. The figure shows that input parameters of the RFGA model are subjected to the GA-based parameter optimization process. Only that pair of parameters that minimizes the OOB error rate in this step is used as input to the RFGA model.

The objective function of GA is as follows:

$$Minimize(object) \tag{10}$$

Here,

$$(object) = (OOBerrorratio) \tag{11}$$

$$= f(m, B) \tag{12}$$

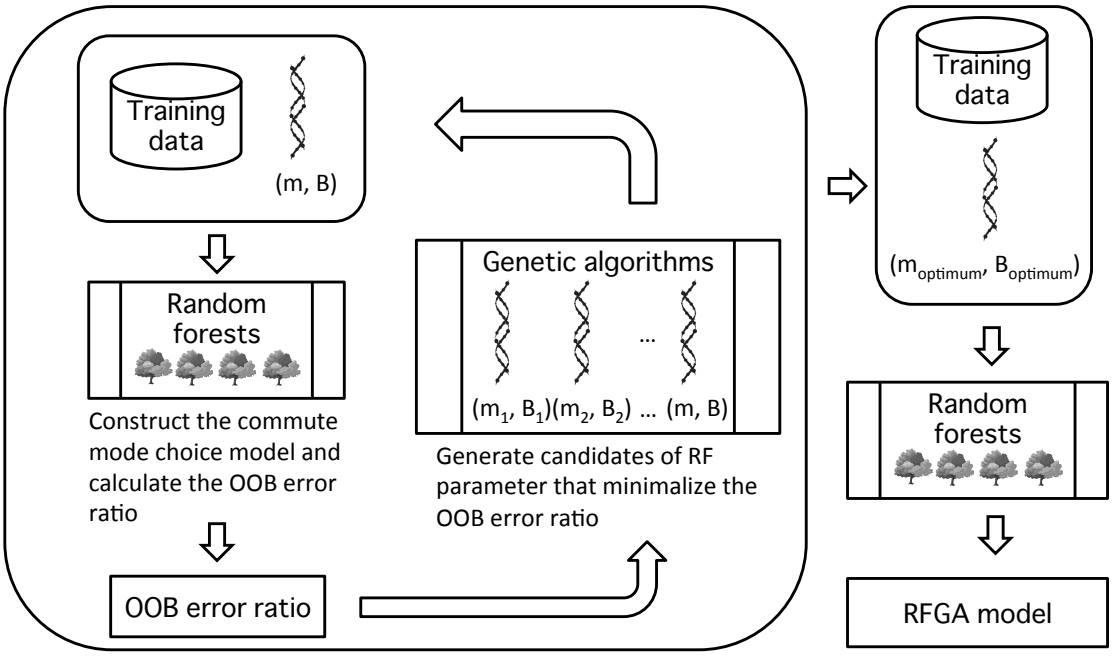

Figure 2. Flowchart of RFGA

Note that $f(m, B)$ is calculated by the procedure that growing each tree in the RF algorithms (see the previous subsection). The computational conditions for the GA-based parameter optimization process are as follows:

- The maximum number of generations is 100.
- The population size is 300.
- The domain of allowable values for each parameter of the function being optimized are $1 \leq m \leq 7$ ($m \in \mathbf{N}$) and $1 \leq B \leq 2000$ ($B \in \mathbf{N}$).

As a result, $m = 4$ and $B = 74$ are obtained. The run time of this process till the calculation is complete is approximately 10 h 12 min, because of no significant improvement in 10 iterations using a machine with 2 GHz Intel Core i7 CPU and 16 GB 1333 MHz DDR3 memory.

3.4 Multinomial Logit Model

The multinomial logit (MNL) model, based on the theory of random utility maximization, is one of the most popular methods for discrete choice analysis in the transportation research field (Ben-Akiva and Lerman (1985)). Hence, we chose the MNL model to be compared with the RFGA model. The systematic component of the utility V_{in} is as follows:

$$V_{in} = \sum_{k=1}^{K} \theta_k X_{ink} \tag{13}$$

where,
θ_k : the coefficient of k-th variable
X_{ink} : variables

The probability of mode choice P_{in} is as follows:

$$P_{in} = \frac{\exp(V_{in})}{\sum_{k=1}^{K} \exp(V_{jn})}, i = 1, \ldots, J \tag{14}$$

Table 4 shows estimated results of MNL using Nelder-Mead method.

Table 4. Multinomial logit model for commute mode choice

	Estimated coefficients	t statistics
Alternative-specific constant (specific to automobile)	−1.004	$-1.39e+01$
Alternative-specific constant (specific to mass transit)	−0.279	$-9.31e+00$
Age (specific to automobile)	−7.059	$-1.10e+02$
Gender dummy	−125.161	$-1.33e+01$
Licence of automobile dummy (specific to automobile)	1.282	$1.64e+01$
Automobile ownership dummy (specific to automobile)	2.001	$4.42e+01$
Travel time	−14.071	$-4.57e-05$
Travel cost for automobile (specific to automobile)	11.260	$1.19e+00$
Travel cost for mass transit (specific to mass transit)	14.507	$3.86e+02$

$\rho^2 = 0.413$
Adjusted $\rho^2 = 0.413$
Hit ratio = 0.760

In terms of model fit, ρ^2 and the adjusted ρ^2, as well as the hit ratio, were reasonably good. Nevertheless, estimated coefficients of 'Travel time' and 'Travel cost for automobile' are not significant at the 95 % confidence level. Therefore, to estimate more reasonable model, 'Travel time' and 'Travel cost for automobile' were removed from the model, and the model was estimated again (table 5).

Table 5. Revised multinomial logit model for commute mode choice

	Estimated coefficients	t statistics
Alternative-specific constant (specific to automobile)	−1.014	$-1.40e+01$
Alternative-specific constant (specific to masstransit)	−0.286	$-9.52e+00$
Age (specific to automobile)	−8.763	$-3.47e+09$
Gender dummy	−4.531	$-1.60e+09$
Licence of automobile dummy (specific to automobile)	1.331	$1.70e+01$
Automobile ownership dummy (specific to automobile)	2.037	$4.53e+01$
Travel cost for masstransit (specific to masstransit)	12.118	$2.01e+00$

$\rho^2 = 0.404$
Adjusted $\rho^2 = 0.404$
Hit ratio = 0.760

'Travel cost' is commonly known as a negative utility, although the estimated coefficient and t-statistic are positive in table 4 and table 5. It can be explained that the almost commuting allowance is provided by the employer in Japan, therefore, commuters don't have to bear their 'Travel cost'.

4. RESULTS AND DISCUSSION

4.1 Results

The RF model, RFGA model and MNL model have been inputted with training data and test data. Table 6 shows the classification results of the RF model using test data that include confusion matrix, precision, recall and hit ratio. Table 7 shows the classification results of the RF model using training data that include confusion matrix, precision, recall and hit ratio. From these results, following points were clarified:

- The test data tending to show higher accuracy than the training data were used for clarification.
- The accuracy of the other mode was lower in both training and test data.
- The minimum accuracy using test data was $k_{Other} = 0.474$
- The maximum accuracy using test data was $k_{Auto} = 0.913$

Table 6. Classification results of RF model using test data

	Auto	Mass	Other	Total
Auto	8250	735	1179	10164
Mass	319	4359	178	4856
Other	470	63	1224	1757
Total	9039	5157	2581	16777
α	0.812	0.898	0.697	
k	0.913	0.845	0.474	
H				0.825

Table 7. Classification results of RF model using training data

	Auto	Mass	Other	Total
Auto	8183	739	1209	10131
Mass	342	4363	171	4876
Other	514	55	1200	1769
Total	9039	5157	2580	16776
α	0.808	0.895	0.678	
k	0.905	0.846	0.465	
H				0.819

Table 8 shows the classification results of the RFGA model using test data that include confusion matrix, precision, recall and hit ratio. Table 9 shows the classification results of the RFGA model using training data that include confusion matrix, precision, recall and hit ratio. From these results, following points were clarified:

- The training data tending to show higher accuracy than the test data were used for clarification.
- The accuracy of the other mode was lower in both training and test data.
- The minimum accuracy using test data was $k_{Other} = 0.537$
- The maximum accuracy using test data was $\alpha_{Mass} = 0.930$

Table 10 shows the classification results of the revised MNL model (table 5) using test data that include confusion matrix, precision, recall and hit ratio. Table 11 shows the classification results of the MNL model using training data that include confusion matrix, precision, recall and hit ratio. From these results, following points were clarified:

- Each accuracy using test data and training data were approximately same.
- The accuracy of the other mode was lower in both training and test data.
- The minimum accuracy using test data was $k_{Other} = 0.295$
- The maximum accuracy using test data was $\alpha_{Mass} = 1.000$

Table 8. Classification results of RFGA model using test data

	Auto	Mass	Other	Total
Auto	8246	667	1113	10026
Mass	250	4381	82	4713
Other	543	109	1386	2038
Total	9039	5157	2581	16777
α	0.822	0.930	0.680	
k	0.912	0.850	0.537	
H				0.835

Table 9. Classification results of RFGA model using training data

	Auto	Mass	Other	Total
Auto	8340	585	1078	10003
Mass	193	4478	62	4733
Other	506	94	1440	2040
Total	9039	5157	2580	16776
α	0.834	0.946	0.706	
k	0.923	0.868	0.558	
H				0.850

Note that 'MA' and 'MO' are zero in table 10 and table 11 . It suggests that the obtained utility function of mass transit responds discretely to input values.

Table 10. Classification results of MNL model using test data

	Auto	Mass	Other	Total
Auto	8798	1546	1820	12164
Mass	0	3190	0	3190
Other	241	421	761	1423
Total	9039	5157	2581	16777
α	0.723	1.000	0.535	
k	0.973	0.619	0.295	
H				0.760

Table 11. Classification results of MNL model using training data

	Auto	Mass	Other	Total
Auto	8800	1552	1840	12192
Mass	0	3205	0	3205
Other	239	400	740	1379
Total	9039	5157	2580	16776
α	0.722	1.000	0.537	
k	0.974	0.621	0.287	
H				0.760

4.2 Discussion

This study proposes a hybrid RFGA model for a practical and accurate commute mode choice analysis. We propose a practical method for optimizing the parameter for the RF model by metaheuristic optimization using GAs.

Our study reveals that the generalization performance would be more accurate through the optimization of the RF model parameter. Furthermore, compared to conventional model on commute mode choice analysis, the present RFGA model offers improved generalization performance (Table 6, Table 8, Table 10). The results obtained from the comparison can be summarized as follows:

- The hit ratio H becomes more accurately. Hence, in terms of the entire accuracy of the classification results, RFGA model is the most accuracy.
- The precision α_{Auto} becomes more accurately, furthermore, α_{Mass} and α_{Other} are second best. Hence, in terms of the accuracy of the classification results by type of transportation mode, RFGA model is accuracy enough.
- The recall k_{Mass} and k_{Other} become more accurately, furthermore, k_{Other} is third by a narrow margin. Hence, in terms of the accuracy in reproducing the observations by type of transportation mode, RFGA model is accuracy enough.

- The minimum accuracy using test data is $k_{Other} = 0.537$, and the maximum is $\alpha_{Mass} = 0.930$ Hence, the generalization performance of RFGA model is more stable.

Therefore, the study indicates that the RFGA model performs much better than conventional models.

In the future, we will examine the extension of the suggested methods to followings:

- Applying to other survey data to confirm the model transferability.
- Applying to large scale and continuous datasets (The World Economic Forum (2012)), i.e., probe car data, probe person data, and public transport smart card data.
- Fusing into the random utility maximization theory to understand travel behaviour deeply.

5. CONCLUSIONS

This study proposes a hybrid RFGA model for a practical and accurate commute mode choice analysis.

The household travel survey collected for the centre of Hokkaido Prefecture, Japan, in 2006 is used to investigate the commute mode choice analysis. This original data were prepared and partitioned into training data and test data that with one dependent variable and seven independent variables.

We propose a practical method for optimizing the RF model parameter by metaheuristic optimization using GAs. The input parameters of the RFGA model were subjected to the GA-based parameter optimization process. Only that pair of parameters that minimizes the OOB error rate in this step was used as input to the RFGA model. The run time of this optimization process till the calculation is complete is approximately 10 h 12 min using a machine with 2 GHz Intel Core i7 CPU and 16 GB 1333 MHz DDR3 memory.

Our study reveals that the optimization of the RF model parameter will lead to more accurate generalization performance. Furthermore, compared to conventional model on commute mode choice analysis, the present RFGA model offers improved generalization performance. The advantage of RFGA model can be summarized as follows:

- The highest entire accuracy of the classification results.
- High accuracy of the classification results by type of transportation mode.
- High accuracy in reproducing the observations by type of transportation mode.
- Stable generalization performance.

Therefore, the study indicates that the RFGA model performs much better than conventional models.

ACKNOWLEDGEMENTS

This paper was written in the term of the 1st author's sabbatical year in the School of Civil and Environmental Engineering, Faculty of Engineering, the University of New South Wales. I am deeply grateful to Professor John Black who offered continuing support and constant encouragement. I am also indebted to Professor Travis Waller and the other colleagues at the UNSW whose kindnesses were an enormous help to me. Finally, I would like to thank Institute

of National Colleges of Technology Japan (researcher overseas visit program) and Japan society for the promotion of science (KAKENHI Grant-in-Aid for Young Scientists (B) Number 23760469) for a subsidy those made it possible to complete this study.

REFERENCES

Ben-Akiva, M. E. and Lerman, S. R. (1985) *Discrete Choice Analysis: Theory and Applications to Travel Demand*, Vol. 6 of MIT Press Series in Transportation Studies: MIT Press, pp.412.

Committee of Infrastructure Planning and Management ed. (1995) *Theory and Practice of Disaggregate Behavioral Model*: Japan Society of Civil Engineering (in Japanese).

Kitamura, R., Morikawa, T., Sasaki, K., Fujii, S., and Yamamoto, T. (2002) *Modeling Travel Behavior*: GIHODO SHUPPAN Co., Ltd. (in Japanese).

Bishop, C. M. (2007) *Pattern Recognition and Machine Learning*: Springer, 1st edition, pp.738.

Lu, Y., Kawamura, K., and Zellner, M. L. (2008) Exploring the Influence of Urban Form on Work Travel Behavior with Agent-Based Modeling. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2082, pp. 132–140, December.

Lu, Y. and Kawamura, K. (2010) Data-Mining Approach to Work Trip Mode Choice Analysis in Chicago, Illinois, Area. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2156, pp. 73–80, December.

Hasegawa, H., Naito, T., Arimura, M., and Tamura, T. (2012) Modal choice analysis using ensemble learning methods. *Journal of Japan Society of Civil Engineering*, Vol. 68, No. 5, pp. 773–780, (in Japanese).

Breiman, L. (2001) Random forests. *Machine Learning*, Vol. 45, pp. 5–32.

Haleem, K., Abdel-Aty, M., and Santos, J. (2010) Multiple Applications of Multivariate Adaptive Regression Splines Technique to Predict Rear-End Crashes at Unsignalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2165, pp. 33–41, December.

Hossain, M. and Muromachi, Y. (2011) Understanding Crash Mechanisms and Selecting Interventions to Mitigate Real-Time Hazards on Urban Expressways. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2213, pp. 53–62, December.

Pande, A., Das, A., Abdel-Aty, M., and Hassan, H. (2011) Estimation of Real-Time Crash Risk. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2237, pp. 60–66, December.

R Core Team (2012) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Liaw, A. and Wiener, M. (2002) Classification and Regression by randomForest. *R News*, Vol. 2, No. 3, pp. 18–22.

Mitchell, M. (1996) *An Introduction To Genetic Algorithms*, Cambridge: MIT Press, pp.205.

Mebane, W. R. J. and Sekhon, J. S. (2011) Genetic Optimization Using Derivatives : The rgenoud Package for R. *Journal of Statistical Software*, Vol. 42, No. 11.

The World Economic Forum (2012) Big Data , Big Impact : New Possibilities for International Development. Technical report, The World Economic Forum, pp. 1–9.