An Analysis of Vehicle Speed Distribution by Using Traffic Counter Big Data

Wataru YAMAMOTO [a] and Makoto TSUKAI [b]

[a,b] *Graduate School of Engineering University of Hiroshima, 1-4-1, Kagamiyama, Japan*
[a] *E-mail: wataruy9@hiroshima-u.ac.jp*
[b] *E-mail: mtukai@hiroshima-u.ac.jp*

Abstract: In order to find significant factors for speed decrease single lane for one way highway, we estimate a stochastic frontier model (SFM) to obtain possible maximum speed under the given traffic volume. A purpose of this study is to analyze more detail time and space. It is an issue to propose the method for reduction of driving speed using big data. As to find a plausible functional specification, we estimated four specifications in SFM as follows: log, cubic, squared, linear; and found that squared in traffic volume gave best fit in log likelihood. By a decision tree analysis, heavy vehicle ratio in traffic and weather condition would have significant influence on temporal decrease in driving speed.

Keywords: Stochastic frontier, Decision tree, Traffic counter, Heavy vehicle

1. INTRODUCTION

Progressive technology of information and communications allow us to get various big data, which is utilized for transportation planning, management, and evaluation. Tanishita *et al.* (2016) showed that changes in average speed affected on traffic accident rates per vehicle-kilometer by using a five-minute continuous monitoring data on an expressway. Weimin Zheng *et al.* (2017) focus on how to predict precisely the tourist's trip chain. They collected trip information from tourists using GPS tracking technology and proposed a method behavior model matching for GPS data. From a viewpoint to shorten transportation time on road traffic, prediction on congestion occurrence about the timming and location is considered issue. In order to detect congestion, GPS information can be directly utilized by calculating vehicle speed and location. Shi An *et al.* (2016) developed a method to measure the evolution patterns of urban daily congestion based on GPS-equipped vehicle mobility data. The datasets in this study was a Harbin digital map and taxi GPS data. Xiangjie Kong *et al.* (2016) proposed an approach to estimate and predict the urban traffic congestion using car trajectory on GPS data. Eleonora *et al.* (2016) proposed a system to detect traffic congestion and accidents from real-time GPS data. Gaetano *et al.* (2016) proposed a method to give precise traffic predictions by exploiting a number of probe vehicle data. Seungwoo *et al.* (2016) suggested a new statistical model to find the optimal data range according to various analyses on each link and provide better transportation time on a specific link by day of the week. As shown in the above, most of conventional studies focused on congestion. However, a minor congestion such as temporal decrease in driving speed has been neglected. In single lane for one way, the influence of decrease in driving speed by slowly driving car remains until the section of two lanes. Therefore, it is important to confirm how often the decrease occurs and what the condition for the decrease is. For example, heavy vehicle with slow speed is a key element of the road network condition. Heavy vehicle with low speed would be a key factor to influence on driving speed.

Over the past few decades, several research has conducted to find other factors to influence on driving speed, such as road geometry, functional classification, roadside interference, traffic, speed limits, weather condition and so on. The AASHTO Green Book (2011) recommended the 85th percentile of the driving speed distribution to be used as a standard to evaluate road performance. However, most of driving speed models do not care for driving speed distribution. Tarris *et al.* (1996) reported that the loss of information by speed data aggregation reduce the regression variance, which may cause a downward bias of road geometry influence. Tarris *et al.* proposed a model for the entire speed distribution in vehicle group in order to avoid a point estimation of driving speed. Figueroa and Tarko (2005) developed a speed estimation models with linear combination of the mean and standard deviation of the speed distribution. The models gave different factors to influence on average speed and speed dispersion. Furthermore, in another publication by Figueroa and Tarko (2004), percentile specific and site-specific random effects were introduced in order to avoid biased parameters. In the estimated model the random effects was significant. Lobo *et al.* (2014) applied a stochastic frontier model developed in econometrics to driving speed data. However, influence of traffic volume on driving speed was not considered.

The hypothesis in this study are as follows; even though the effect of traffic volume is removed, heavy vehicle significantly influence on speed. In order to test the hypothesis, we estimate a driving speed frontier model considering traffic volume and calculate a deviation from the frontier to observed speed, and finds the factor for the deviation.

2. METHODOLOGY

2.1 Stochastic Frontier Analysis

The stochastic frontier analysis (SFA) was proposed in 1977 by Aigner *et al.* (1977) and by Meeusen and van den Broeck (1977), which is widely applied to measure efficiency of samples in econometric analysis (2008). Recent applications are found in many other fields as follows; agriculture, (Baten *et al.* (2009), Ali and Samad (2013)), finance (Wang (2003), Neffati *et al.* (2011)), public utility (Hattori (2002), Vishwakarma and Kulshrestha (2010)), and transportation (Pendyala *et al.* (2002), Cullinane *et al.* (2002), Holmgren (2013)). There are two different approaches in sample efficiency measurement; parametric and non-parametric approaches. SFA is a parametric approach to evaluate efficiency of each sample, to estimate maximum output by use of available inputs.

SFA is specified as follows

$$Y_i = \beta X_i + v_i - u_i \tag{1}$$

where,

i : sample in observation,

Y_i : output in sample i,

β : vector of input coefficients,

X_i : vector of inputs of sample i,

v_i : error term,

u_i : one-sided inefficiency term.

The model consists of deterministic component (this type is called “production function”) and two disturbance components. One of disturbance term is a random to give stochastic variation of samples as observation error. Other disturbance term measures a deviation from the stochastic frontier to each sample (i.e.u is positive). In SFA,$u$is specified in half-normal, truncated normal, exponential, and gamma distributions have been suggested as possible distributions. This paper specified as truncated normal distribution following to Aigner’s original specification. Since $E[v_i]=0$, an expected value of inefficiency for each sample is obtained as follows;

$$E[u_i]=E[y_i-\hat{\beta}X_i-v_i]=E[y_i-\hat{\beta}X_i] \tag{2}$$

Note that $E[u_i]$ is an expected value with stochastic term μ, so it can be negative due to the variation of u_iTherefore, depending on the error term, the stochastic frontier output can lie below the deterministic component. In this paper, the output below the frontier level lead to inefficiency.

2.2 Decision Tree Analysis

Decision tree analysis gives a hierarchical tree structure to classify the sample set in terms of reference attribute distribution which is called target attribute in decision tree analysis, Ilyes Jenhani *et al.* (2008). The tree structure is composed of following three basic elements, nodes corresponding to an attribute to give different distribution for target attribute, edges corresponding to connecting another possible attribute. And leaves with homogeneous samples against the other leaves in same hierarchy. Such representation allows us to induce decision rules to classify new samples. In fact, each path from the root to a leaf corresponds to a conjunction of attributes and the tree is considered as a disjunction of these conjunctions. The majority of decision trees is made up of two major procedures: the building (induction) and the classification (inference) procedures.

Building procedure: Given a training set, building a decision tree is done by starting with an empty tree and selecting for each decision node by repeatedly applying appropriateness test about the candidate attribute. The principle to select an attribute is to maximally diminish the mixture of classes between each training subset created by the test, thus, making easier the determination of object’s classes. The process continues for each sub decision tree until reaching leaves and fixing their corresponding classes. Note that the above procedure can not give on unique optimized decision tree, because some trials are made at node section. However, the obtained tree is considerably robust for slightly different dataset or parameter setting.

Classification procedure: To classify a new sample, having only values of all its attributes, we start with the root of the constructed tree and follow the path corresponding to the observed value of the attribute in the interior node of the tree. This process is continued until a leaf is encountered. Finally, we use the associated label to obtain the predicted class value of the sample at hand.

In this study, we apply decision tree analysis for the dataset about the deviation from estimated frontier in order to find some leaves to include highly deviated, i.e. low speed samples with the combination of attributes appearing on the path from the root to the leaf.

Since our purpose is not to classify the new sample, the latter classification procedure is skipped.

3. DATA

In stochastic frontier analysis, objective variable is speed of vehicle. In order to remove structure effect, we used a data of traffic volume as an explanatory variable. The data is observed on Tottori-Himeji line between April 2015 and March 2016. Tottori-Himeji line was constructed as local highway to connect Tottori city (Capital in Tottori Prefecture) to Chugoku expressway at Sayo Interchange through Himeji-city to connect Sanyo expressway. Tottori-Himeji line is basically single lane for one way, except for some interchange sites. Some sections around Tottori and Sayo are tolled, but the intermediate sections are untolled, hence the speed limit is not constant over the section. The data is collected by four traffic counters located in different site and the speed is averaged over every hour from am 7 to pm 6. In decision tree, the target attribute is inefficiency (i.e. index of driving speed) estimated from the stochastic frontier model, and other attributes are heavy vehicle ratio (HVR) by traffic counter and weather information. Weather is an observed at monitoring site, and we made matching the closest monitoring site with each traffic counter.

The location of traffic counters and weather monitoring site is shown in Figure 1, and summary of data are shown in Table 1. Sample size is 4,382 at Shimoajino, 4,359 at Katayama, 4,346 at Takatsuhara, 4,358 at Minari bridge because some missing or erroneous observation occurred with zero traffic but positive speed record.

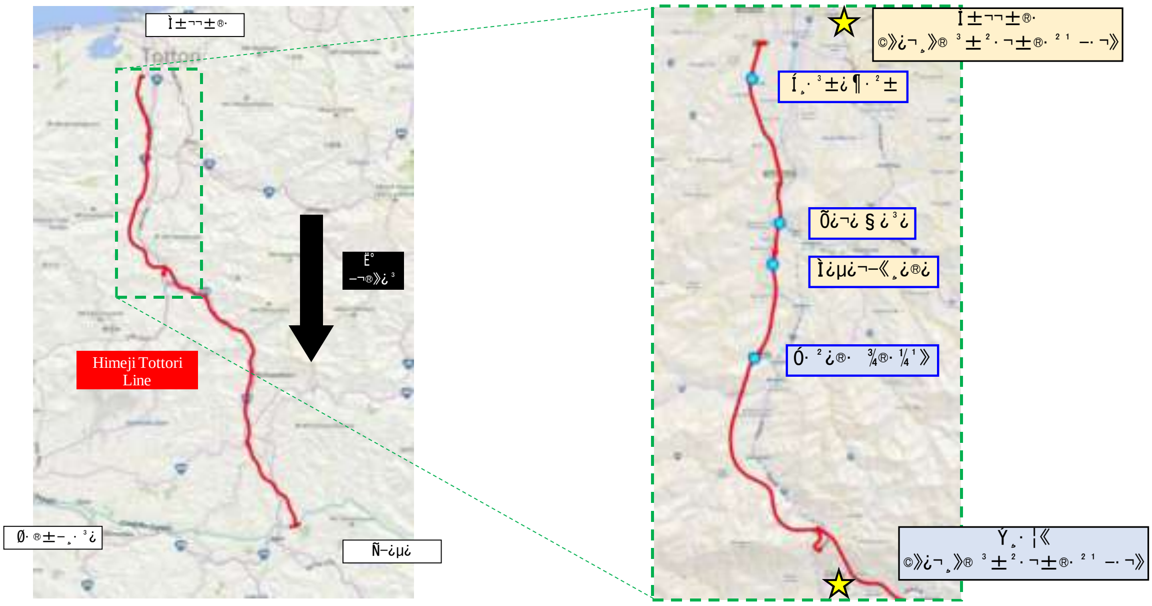

Figure 1. The location of traffic counters and weather monitoring site

Table 1. Summary of data

Variable description	Average	Standard Deviation	Minimum	Maximum
Traffic volume				
Shimoajino	443.3	126.1	1	1105
Katayama	493.5	131.5	1	1152
Takatsuhara	413.8	121.4	1	1131
Minari bridge	360.6	114.6	1	1104
Speed(km/h)				
Shimoajino	70.6	8.2	0	82
Katayama	70.1	5.9	0	77
Takatsuhara	77.6	7.5	0	87
Minari bridge	75.2	8.6	0	85
HVR (%)				
Shimoajino	10.4	0.1	0	100
Katayama	10.1	0.1	0	79
Takatsuhara	14.7	0.1	0	75
Minari bridge	17	0.1	0	100
Rain (dummy)				
Tottori	0.1	0.3	0	1
Chizu	0.1	0.3	0	1
Snow (dummy)				
Tottori	0.02	0.1	0	1
Chizu	0.02	0.1	0	1

- Not relevant.

4. EMPIRICAL ANALYSIS

4.1 Estimation of Speed Frontier

The explanatory variables in stochastic frontier model (SFM) are linear combination of traffic volume with differently transformed as follows: log, cubic, squared and linear. In order to find a plausible specification, we estimated the following equations in (3) to (6).

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 \ln x_i + v - u \tag{3}$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + v - u \tag{4}$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + v - u \tag{5}$$

$$y_i = \beta_0 + \beta_1 x_i + v - u \tag{6}$$

Hereafter, eq. (3) to (6) are called log, cubic, squared and linear function, respectively. Comparison of plot in estimated and measured value is shown in Figure 2 and comparison of log likelihood is shown in Table 2. In log and cubic function, estimated value of traffic volume appearing around the area close to zero. However, it is not acceptable because under the very low traffic volume such as the free flow condition, drivers can choose any preferable speed. Therefore in ordinary Q-V plot traffic volume in vertical axis, speed in horizontal axis, speed lies upper area around low traffic volume. Log and cubic function is not acceptable due to the above discussion. Cubic function is not good around higher traffic ($volume \geq 1{,}000$) in terms of data fitting. On the other hand, squared and linear function is better fitting than the other two functions. Furthermore, to compare with log likelihood, squared function is better than linear function. Therefore, we selected squared function as frontier curve. SFM parameter estimation results are shown in Table 3 and Table 4. All the parameter are significant and all λ are about 1.0.

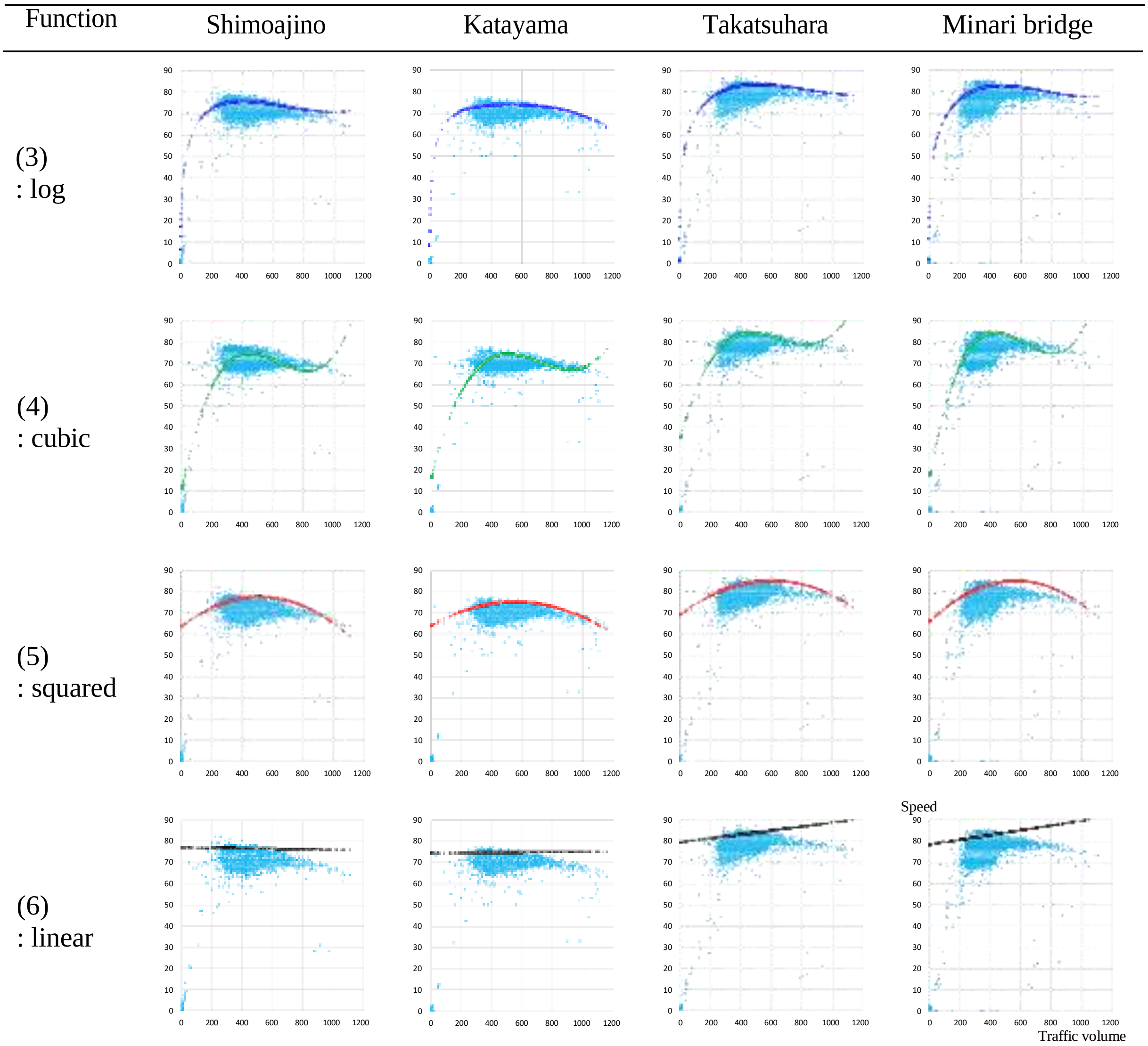

Figure 2. Comparison in estimated and observed value
(horizontal: traffic volume, vertical: speed)

Table 2. Comparison of log likelihood

Function	Shimoajino	Katayama	Takatsuhara	Minari bridge
(3) : log	-11585.7	-10635.0	-11912.8	-12910.4
(4) : cubic	-12320.9	-11573.6	-12412.3	-13410.1
(5) : squared	-13322.3	-11848.2	-12783.3	-13599.9
(6) : linear	-13595.9	-12116.9	-13056.2	-13855.2

Table 3. SFM estimation results at Shimoajino and Katayama

Independent Variables	Shimoajino		Katayama	
	Parameter	t score	Parameter	t score
Constant	63.472***	83.527	63.843***	122.993
Traffic volume (linear)	0.054***	19.202	0.040***	22.588
Traffic volume (squared)	-5.E-05***	-21.762	-4.E-05***	-25.223
σ^2	79.898***	37.737	43.336***	40.029
λ	0.958***	206.391	0.967***	311.986
log-likelihood (at converged)	-13322.3		-11848.2	
Observations	4,382		4,359	

*** Significant at 0.1% level, $\sigma^2 = \sigma_v{}^2 + \sigma_u{}^2$, $\lambda = \sigma_u / \sigma_v$

Table 4. SFM estimation results at Takatsuhara and Minari bridge

Independent Variables	Takatsuhara		Minari bridge	
	Parameter	t score	Parameter	t score
Constant	69.136***	125.004	65.944***	95.636
Traffic volume (linear)	0.053***	25.717	0.067***	23.723
Traffic volume (squared)	-4.E-05***	-26.086	-6.E-05***	-23.851
σ^2	68.613***	40.960	95.394***	41.086
λ	0.971***	376.886	0.965***	273.671
log-likelihood (at converged)	-12783.3		-13599.9	
Observations	4,346		4,358	

*** Significant at 0.1% level, $\sigma^2 = \sigma_v{}^2 + \sigma_u{}^2$, $\lambda = \sigma_u / \sigma_v$

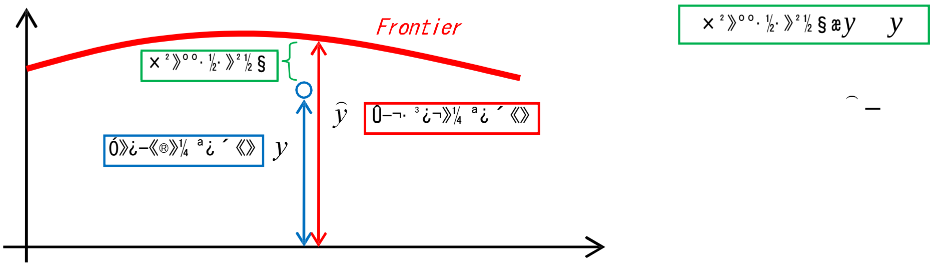

Figure 3. Definition of inefficiency

4.2 Factors for Inefficiency

Factors for the inefficiency are detected by decision tree. Inefficiency is calculated as the deviation from expected frontier under the given traffic volume to each observation. Definition of inefficiency is shown Figure 3. In decision tree, dependent variable (driving speed) is categorized by each 10percentile of its distribution in order to make it easy to find inefficiency deviation. By this data processing, driving speed index uniformly distributes over 0 to 90. The explanatory variables are as follows: weekday dummy, time in a day, season, site (at traffic counter point), heavy vehicle ratio (HVR), and weather (rain and snow dummy). The decision tree estimated by using statistical software R ("rpart" function), set for maxdepth:3 and cp (complexity parameter):0.0001. The result for all observations is shown in Figure 4. The nodes on decision tree reveals the factor to influence inefficiency distribution and those sample size is shown at the leaves (i.e. end node). Node 4, 5 and 14 seems minor, because the sample size of those group is less than 1% in whole sample. Site influences inefficiency at first, and whole sample are divided into Katayama and others. Then, HVR is the next factor following to Katayama. The traffic counter at Shimoajino locates behind a tunnel and at Takatsuhara and Minari bridge locates around a ramp. On the other hand, the traffic counter at Katayama locates a road without a tunnel and apart from ramp over 1km. Considering such difference in observation site Shimoajino, Takatsuhara, Minari bridge are less inefficiency (i.e. skewed to lower driving speed) than Katayama. In Katayama, HVR and the time in a day influences on speed which are inefficient from 8 to 10, from 13 to 15 and 18, or heavy vehicle ratio is larger than 0.076.

The decision tree using whole sample showed site specific factor is much stronger than other common factors. Hence, we apply the decision tree to site-wise data in Shimoajino, Katayama, Takatsuhara, and Minari bridge, respectively. The dependent variable is inefficiency categorized in 10 percentile for each site. Therefore, corresponding driving speed of percentile threshold results of analysis are shown in Figure 5 to 8, respectively. In Shimoajino, the value of 50 percentile, equal to the deviation from estimated frontier, is about -5.0km/h. The season influence on the speed and winter is more efficient than other seasons. In case of snow at Node 12 a trend with low speed in snowing is confirmed, although those node sample size is about 1.0% of the whole. In other seasons, it is more inefficient in time in a day with 13, 14, and 15 than others. The sample size of Node 5 is 18.3%. In Katayama, the value of 50 percentile is about -3.6km/h. As mentioned above, HVR and time in a day influence on the speed. Node 7 is more inefficient than Node 8. That is an effect of snow but those sample size is about 0.5% of the whole sample size. In Takatsuhara, the value of 50 percentile is about -4.4km/h. Time in a day is most fundamental factor on speed with more inefficient between 7 to 10 and 13 to 15 and 18 than the others. Node 3 shows speed decrease occurs at snowing, but those node sample size is minor 1.6% of the whole, as same as Shimoajino. In Minari bridge, the value of 50 percentile is about -5.5km/h. The factor on lower speed is not only season or HVR but also rain. Especially, in the case of rain (Node 7), the speed is much decreased and the sample size of Node 7 is 8.3%. In spring and rain, the time in a day exception 10, 13, 16 and 17 influence on the speed but not 1.7% of the whole sample size.

4.3. Discussion

Through above results, following two tendency are commonly observed over several sites. First, the existence of heavy vehicle influence on inefficiency of driving speed. That is confirmed Shimoajino, Katayama and Minari bridge except Takatsuhara. Especially in

Shimoajino, the sample influenced by heavy vehicle shares approximately 20% of whole sample size. Secondly, weather influence on driving speed decrease. Although those sample size is small, snow and rain on this route was significant impact on driving speed. However, the sample size influence by rain was related large in Minari bridge, such as 8.3%. According to results of Shimoajino, it is find that the degree of the influence of heavy vehicle on driving speed decrease is appears the time in a day between 13 to 15. Such the tendency imply the regular freight traffic uses the highway. If these heavy vehicle is an important local economic activity, it is difficult to sift the heavy vehicle into other route. Therefore, an expansion of width in this section is one of policy option. Weather influence on driving speed coupled with heavy vehicle. In the case of snow, an appropriate slip prevention by chemical substance is important. Moreover, for the rain, to improve drainage performance considering the condition of road surface is recommended.

5. CONCLUSION

This study proposed a statistical procedure to detect the potential factor to influence on speed decrease in single lane for one way expressway. We estimate a speed frontier to remove traffic volume influence. The inefficiency (i.e. an index of speed decrease) is analyzed by decision tree.

The decision tree found the factor for speed decrease. As expected initially, heavy vehicle significantly influence on driving speed decrease at 3sites. Furthermore, we found the influence of weather (rain and snow) on driving speed is also significant and the coupling with heavy vehicle and bad weather gives more decrease in driving speed.

Remaining and further discussed issues are follows. We use the data measured at fixed location in this paper. If we get a GPS data from a floating car probe data, it is possible to detect the location and the timing of driving speed decrease together with other conditions. The data handling procedure to deal with such space-time big data can be developed following to our study.

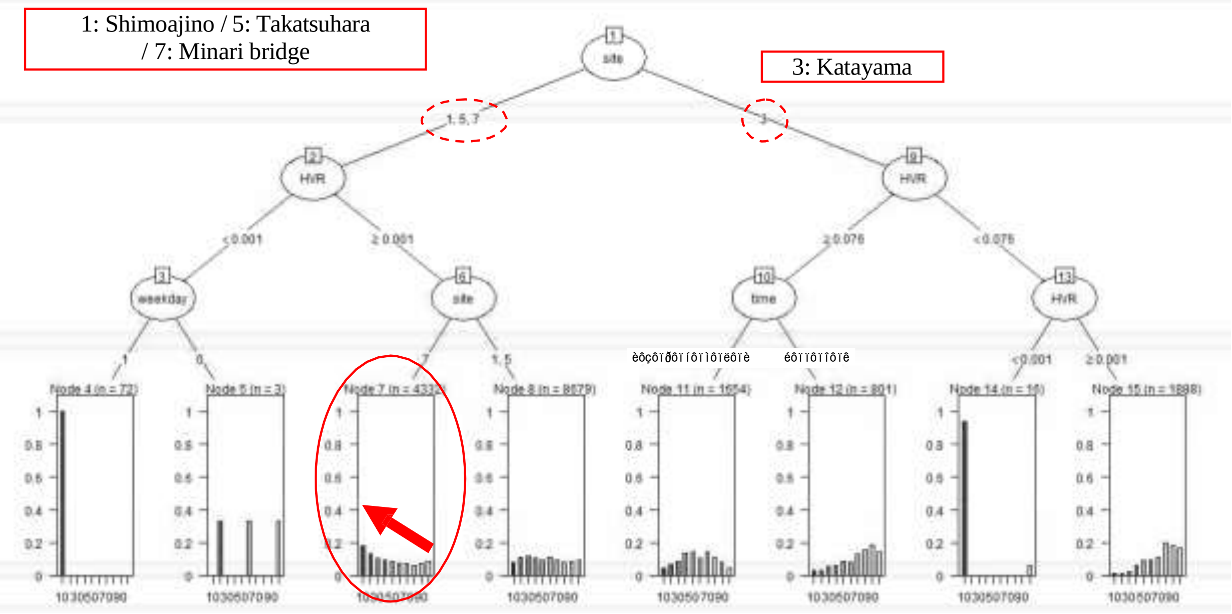

Figure 4. The decision tree for all inefficiency

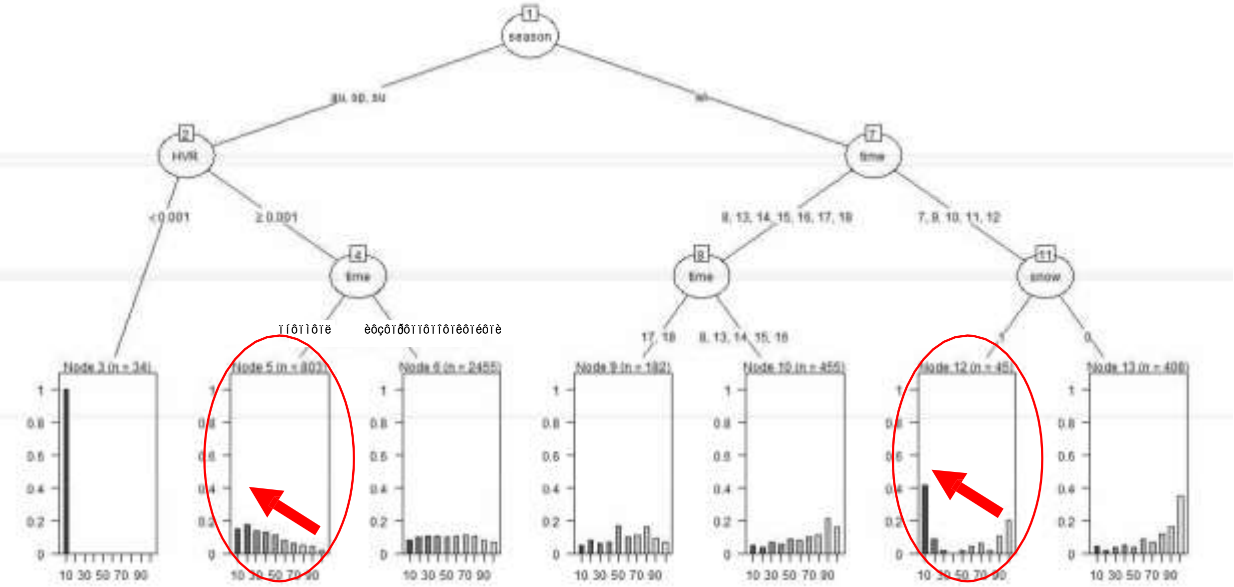

Figure 5. The decision tree for the inefficiency at Shimoajino

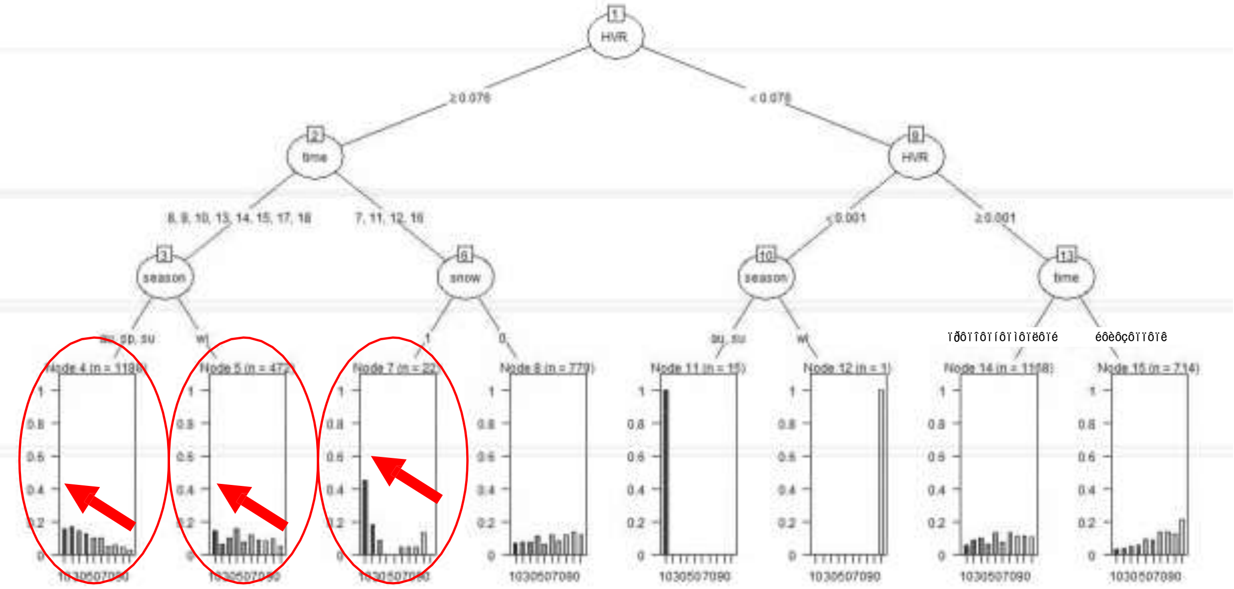

Figure 6. The decision tree for the inefficiency at Katayama

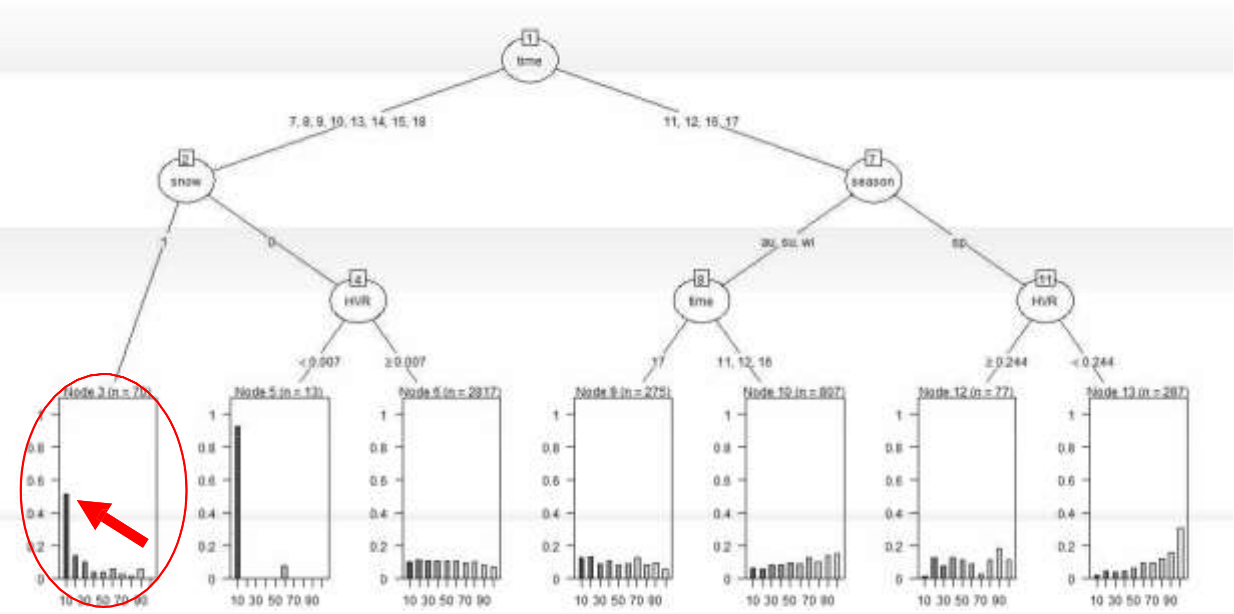

Figure 7. The decision tree for the inefficiency at Takatsuhara

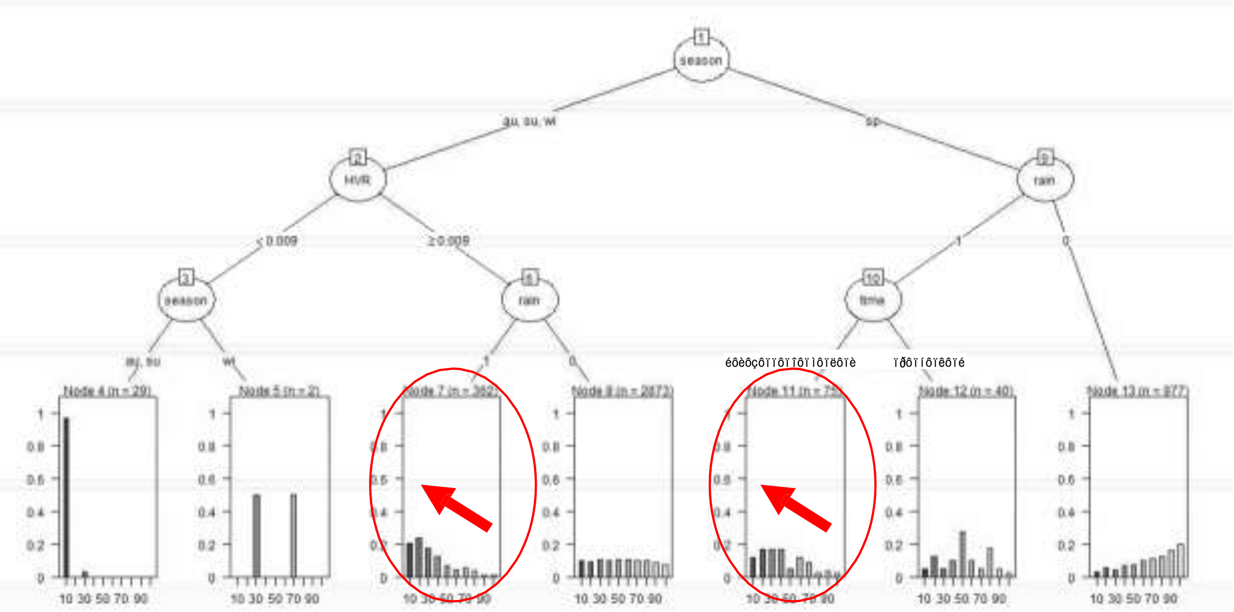

Figure 8. The decision tree for the inefficiency at Minari bridge

REFERENCES

AASHTO. (2011) A Policy on Geometric Design of Highways and Streets. Washington, D.C.

Aigner, D., C. A. K. Lovell, and P. Schmidt. (1977) Formulation and Estimation of Stochastic Frontier Production Function Models. *Journal of Econometrics*, Vol. 6, 21–37.

Ali, K.M.E., Samad, Q.A. (2013) Resource use efficiency in farming: an application of stochastic frontier production function. *J. Agric. Econ. Dev.* 2 (5), 194 –202.

Baten, M.A., Kamil, A.A., Haque, M.A. (2009) Modeling technical inefficiencies effects in a stochastic frontier production function for panel data. *Afr. J. Agric. Res.* 4 (12), 1374–1382.

Cullinane, K., Song, D.-W., Gray, R. (2002) A stochastic frontier model of the efficiency of major container terminals in Asia: assessing the influence of administrative and ownership structures. *Transportation Research. Part A: Policy Pract.* 36 (8), 743–762.

Eleonora D'Andrea and Francesco Marcelloni (2016) Detection of Traffic Congestion and Incidents from GPS Trace Analysis. *Expert Systems with Applications*.

Figueroa Medina, A. M., and A. P. Tarko. (2004) Reconciling Speed Limits with Design Speeds. *Publication FHWA-IN-JTRP-2004/26*. Joint Transportation Research Program, Indiana Department of Transportation and Purdue University, West Lafayette, Ind.

Figueroa Medina, A. M., and A. P. Tarko. (2005) Speed Factors on Two-Lane Rural Highways in Free-Flow Conditions. *Transportation Research Record: Journal of the Transportation Research Board,* No. 1912, 39–46.

Gaetano Fusco. Chiara Colombaroni and Natalia Isaenko. (2016) Short-Term Speed Predictions Exploiting Big Data on Large Urban Road Networks. *Transportation Research Part C: Emerging Technologies,* 73, 183–201.

Greene, (2008) W. H. *Econometric Analysis*, Prentice Hall, Upper Saddle River, N.J.,.

Hattori, T., (2002) Relative performance of U.S. and Japanese electricity distribution: an application of stochastic frontier analysis. *J. Productivity Anal*. 18 (3), 269–284.

Holmgren, J., (2013) The efficiency of public transport operations– an evaluation using stochastic frontier analysis. *Res. Transp. Econ.* 39 (1), 50–57.

Ilyes Jenhani, Nahla Ben Amor and Zied Elouedi, (2008) Decision Trees as Possibilistic Classifiers. *International Journal of Approximate Reasoning*, 48.3, 784–807.

Lobo, A. Rodrigues, C. and Couto, A., (2014) Estimating Percentile Speeds from Maximum Operating Speed Frontier. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2404, 1–8.

Meeusen, W., and J. van den Broeck., (1977) Efficiency Estimation from Cobb- Douglas Production Functions with Composed Error. *International Economic Review,* Vol. 18, 435–444.

Neffati, A. Fredj, I.B. Schalck, C., (2011) Earnings management and banking performance: a stochastic-frontier analysis on U.S. bank mergers. *Interdiscip. J. Res. Bus*. 1 (6), 58–65.

Pendyala, R.M., Yamamoto, T., Kitamura, R., (2002) On the formulation of time-space prisms to model constraints on personal activity-travel engagement. *Transportation* 29 (1), 73–94.

Seungwoo Jeon and Bonghee Hong., (2016) Monte Carlo Simulation-Based Traffic Speed Forecasting Using Historical Big Data. *Future Generation Computer Systems*, 65, 182–95.

Shi An and others, (2016) Mining Urban Recurrent Congestion Evolution Patterns from GPS-Equipped Vehicle Mobility Data. *Information Sciences*, 373, 515–26.

Tanishita, M. and Bert van Wee., (2016) Impact of Vehicle Speeds and Changes in Mean Speeds on per Vehicle-Kilometer Traffic Accident Rates in Japan. *IATSS Research*.

Tarris, J. P., C. M. Poe, J. M. Mason, Jr., and K. G. Goulias., (1996) Predicting Operating Speeds on Low-Speed Urban Streets: Regression and Panel Analysis Approaches. *Transportation Research Record,* No.3, 46–54.

Vishwakarma, A. Kulshrestha, M., (2010) Stochastic production frontier analysis of water supply utility of urban cities in the state of Madhya Pradesh India. *Int. J. Environ. Sci.* 1 (3), 357–367.

Wang, H.-J., (2003) A stochastic frontier analysis of financing constraints on investment: the case of financial liberalization in Taiwan. *J. Bus. Econ. Stat*. 21 (3), 406–419.

Weimin Zheng, Xiaoting Huang and Yuan Li., (2017) Understanding the Tourist Mobility Using GPS: Where Is the next Place?. *Tourism Management*, 59, 267–80.

Xiangjie Kong and others, (2016) Urban Traffic Congestion Estimation and Prediction Based on Floating Car Trajectory Data. *Future Generation Computer Systems*, 61, 97–107.