Returning-Home Analysis in Tokyo Metropolitan Area at the time of the Great East Japan Earthquake using Twitter Data

Yusuke Hara
New Industry Creation Hatchery Center, Tohoku University
6-3-09 Aoba Aramaki Aoba-ku Sendai, Miyagi 980-8579, Japan
hara@plan.civil.tohoku.ac.jp

Abstract

This paper clarifies the occurrence factors of commuters unable to return home and the returning-home decision-making at the time of the Great East Japan Earthquake by using Twitter data. First, to extract the behavior data from the tweet data, we identify each user's returning-home behavior using support vector machines. Second, we create non-verbal explanatory factors using geotag data and verbal explanatory factors using tweet data. Then, we model users' returning-home decision-making by using a discrete choice model and clarify the factors quantitatively. Finally, by sensitivity analysis, we show the effects of the existence of emergency evacuation facilities and line of communication.

1 Introduction

The 2011 earthquake off the Pacific coast of Tohoku, often referred to in Japan as the Great East Japan Earthquake, was a magnitude 9.0 under sea megathrust earthquake that occurred at 14:46JST (05:46 UTC) on March 11, 2011. The focal region of this earthquake was widespread, spanning approximately 500 km north to south from off the Ibaraki shore to the Iwate shore and approximately 200km east to west. The number of deaths and missing persons attributed to this disaster totaled more than 19,000, and the complex, large-scale disasters of the earthquake, tsunami, and nuclear power plant accident had a major impact on people's lives. The Tokyo metropolitan area also was hit by a strong earthquake and various traffic problems occurred. For example, many railway and subway services were suspended for maintenance. Therefore, almost every railway and subway user was unable to return home easily, and they were called "victims unable to return home." According to (Measures Council for Victims Unable to Return Home by Earthquake that directly hits Tokyo Area, 2012), the number of people who were not able to go home during that day by paralysis of these transport networks is estimated about 5.15 million people and it is 30% of a going-out people of the day.

Assessing the problem of "victims unable to return home" in Tokyo metropolitan area is extremely important for anti-disaster measures. Although the questionnaire is performed ex post, it is not yet shown clearly what made going-home decision-making after the earthquake disaster. Moreover, since it was going-home behavior in big confusion, the problem that detailed time and position information are unknown exist.

Some previously studies have examined human behaviors via analysis of behavior log data at the time of a large-scale disaster. Because no rapid and accurate method existed to track population movements after the 2010 earthquake in Haiti, (Bengtsson et al., 2011) used position data from subscriber identity module (SIM) cards from the largest mobile phone company in Haiti to estimate the magnitude and trends of population movements after the 2010 Haiti earthquake and the subsequent cholera outbreak. Their results indicated that estimates of population movements during disasters and outbreaks can be acquired rapidly and with potentially high validity in areas of high mobile phone usage. (Lu et al., 2012) also used the same data in Haiti to determine that 19 days after the earthquake, population movements had caused the population of the capital Port-au-Prince to decrease by approximately 23% and that the destinations of people who left the capital during the first three weeks after the earthquake were highly correlated with their mobility patterns during normal times and specifically with the locations of people with whom they had significant social bonds. Lu

Proceedings of the Workshop on Language Processing and Crisis Information 2013, pages 44–50,
Nagoya, Japan, 14 October 2013. ©2013 Asian Federation of Natural Language Processing

et al. concluded that population movements during disasters may be significantly more predictable than previously thought. Overall, these previous studies clarified human movement over long periods of time. They showed that people in areas affected by an earthquake take refuge temporarily and that the population in the affected area is recovered over several months. Behavior log data should be able to clarify not only such long-term human behavior but also the human behaviors at the time of a disaster.

In this research, we analyze tweet data of Twitter as the behavior log data at the time of the Great East Japan Earthquake. Although tweet data does not contain actual behavior necessarily, there is possibility of containing thinking process and behavioral factors. We clarify the factors of going-home behavior in case of the Great East Japan Earthquake using Twitter data.

2 From Tweet Data To Behavioral Data

2.1 Framework

First, we provide a framework of this research to analyze users' going-home behavior using tweet data and geotag data. Figure 1 shows our framework: (1) behavior inference by tweet data,(2) feature engineering by geotag and tweet data, (3) estimation of behavioral model.

In (1) behavior inference by tweet data part, we inferred users' going-home behavior result using Support vector machine (SVM) and Bag-Of-Words (BOW) representation. In (2) feature engineering by geotag and tweet data part, we made explanatory factors of users' behavior from tweet data and geotag data. In (3) estimation of behavioral model part, we estimated users' behavior model (discrete choice model).

2.2 Data

In this section, we provide an outline of our data. This data is about 180-million tweet by Japanese in Twitter from March 11, 2011 to March 18, 2011. There are about 280 thousands tweet with geotag in this data. We sampled tweets whose timestamp is from 14:00, March 11 to 10:00, March 12 and whose GPS location is within Tokyo metropolitan area. The number of these tweet is 24,737 and the number of unique users (account) is 5,281. To observe users' trip on the day, we extracted users that had over 2 geotag tweet and the number of users is 3,307. We assume that these

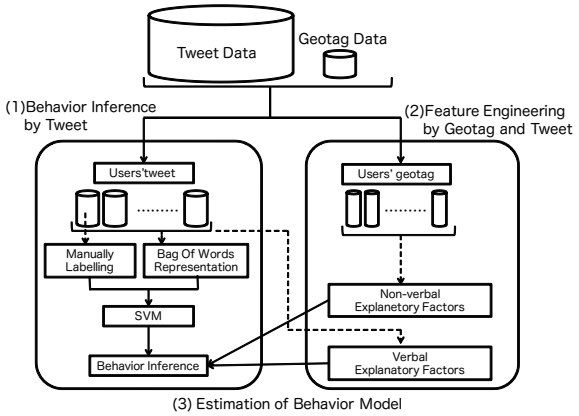

Figure 1: Framework in this research

users can tweet about the Great East Japan Earthquake and their going-home behavior. Therefore, we analyzed all tweet of these users from 14:00, March 11 to 10:00 (3,307 users, 132, 989 tweets).

We tagged 300 users' going-home behavior result manually to make supervised data. Our label set is composed of 1) going home by foot, 2) by train, 3) staying their offices or hotels until tomorrow morning, 4) other choice (taxi, bus, etc.), 5) unclear.

2.3 Morphological analysis

Next, we give morphological analysis by MeCab and obtained BOW representation by each user's tweet. To find the relationship between going-home behavior and each user's tweet, we use information gain. Information gain is index which shows decreasing degree of each class's entropy by existing word w. If word w is contained each user's tweet, Random variableX_w equals 1 and otherwise $X_w = 0$. Random variables which indicates each class is c and entropy $H(c)$ is written as

$$H(C) = -\sum_c P(c) \log P(c). \quad (1)$$

And conditional entropy is written as

$$H(c|X_w = 1) = -\sum_c P(c|X_w = 1) \log P(c|X_w = 1)$$
$$H(c|X_w = 0) = -\sum_c P(c|X_w = 0) \log P(c|X_w = 0).$$

Information gain $IG(w)$ of word w is defined as average decreasing entropy and written as

$$IG(w) = H(c) - (P(X_w = 1)H(c|X_w = 1) + P(X_w = 0)H(c|X_w = 0)) \quad (2)$$

Table 1: Illustrative examples of words whose information gain is high

1)by foot	駅 (station) 歩い (walk) 足 (foot) 休憩 (rest) 自転車 (bicycle) 電車 (train) ヤバイ (danger) 止まっ (stop) 半分 (half) 到着 (arrived) 歩ける (can walk) テレビ (TV) トイレ (toilet) 環七 (Kan-nana Street) km 川崎 (Kawasaki) 疲れ (tired) 遠い (far) 道 (road)
2)by train	大江戸 (O-edo subway line) 入場 (entry) 田園都市線 (Denen-toshi line) 奇跡 (miracle) なんとか (luckily) 順調 (smoothly) 京王 (Keio line) 乗れ (can take a train)
3)stay	泊め (sleep) 朝 (morning) 総武線 (Sobu line) 混雑 (congested) 検索 (search) JR (JR line) 乗車 (take a train) 満員 (full capacity) 明け (daylight) 暇 (a spare time) 始発 (first train in the morning) 悩む (worry)
4)other	Twitpic
5)unclear	jishin, skype

We calculated all words information gain $IG(w)$ by 5 class (walk, train, stay, other, unclear). Table 1 shows illustrative examples. For example, words whose conditional probability of walking is high are "half", "far", "km", "Kawasaki" and "Kannana Street". They show user's location. And "toilet", "tired" and "danger" indicates psychological factors during going-home by foot.

In the case of train, "miracle", "luckily" is contained and "O-edo line" and "Denen-toshi line" are the train and subway lines which is operated in March 11. In the case of stay, "morning", "daylight" and "sleep" indicates that users slept at hotel or their offices and "first train in the morning", "worry" and "search" shows their going-home timing. Other choices users, who choose bicycle, taxi etc, and unclear users don't show the understandable tendency. However, they submitted pictures for Twitpic, which is photo share site, and tweeted with #jishin hashtag.

As seen above, the words whose information gain is high is useful to infer their going-home behavior. Therefore, we made classifier by using these words as features.

2.4 SVM and behavior inference

In this section, we infer each user's behavioral result by SVM. we use 300 labeled data as supervised data and we treat top 500 words of information gain as features of SVM. In learning, we did 9-fold cross validation and average accuracy rate is 73.3%.

Figure 2 shows the inferred result. The number of users by foot is $1,913$, the number of users by train is 359, the number of users staying is 385, the number of users by other choice is 15 and the number of users whose choice is unclear is 635. This result indicates that the ratio of all going-home users except unclear users is $84,9\%$.

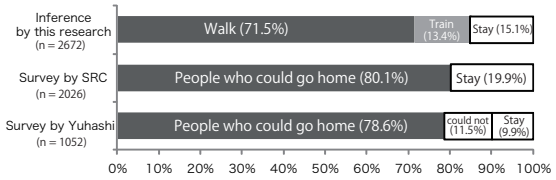

Figure 2: Inferred result and comparison of other survey

To discuss the accuracy of this inference result, we compare our result with other survey results. Figure 2 shows the survey result by (Survey Research Center, 2011) and the survey result by (Yuhashi, 2012). The result of Survey Research Center says 80.1% of all could get home and the result of Yuhashi says 78.6% of all could get home.

3 Behavioral Analysis

3.1 Non-verval factors

Based on the prediction of going-home decision-making classified by user, nonverbal / verbal explanation factor is created from tweet data or geotag data, and the factor of each individual's going-home decision-making is analyzed.

First, the explanation factor about travel behavior is created using the geotag data classified by user. In this research, for simplicity, we assume that a position before the earthquake is the location of office (origin) and a position of 12:00, March 12, 2011 is the location of home (destination). Next, road network distance, the on foot time required, the station nearest office, the station nearest home, the railroad time required, railroad expense, and the number of times of a railroad change are created using these GPS data. These are the features created using the network at the time of usual.

In order to express a spatial spread of people's going-home behavior, Figure 3, 4 shows the spatial dstribution of users' location of before the earthquake and the next day of the earthquake by plotting each user's geotag. As an overall trend, office distribution and house distribution are spatially different, and home distribution is spread in

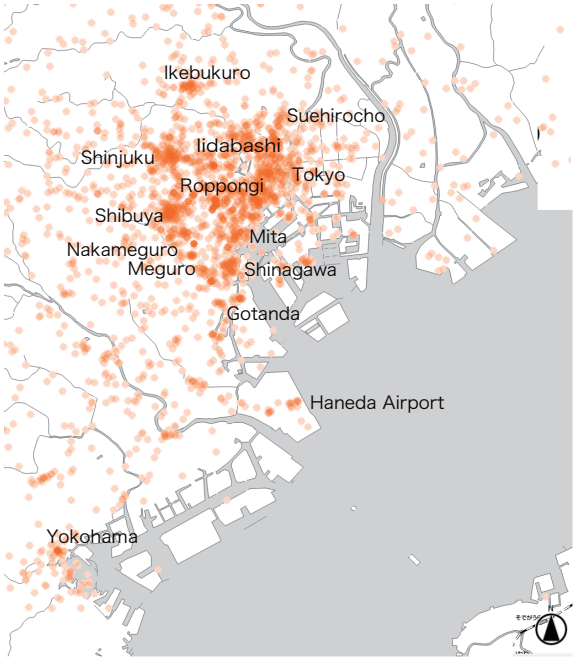

Figure 3: Users' location distribution before the earthquake

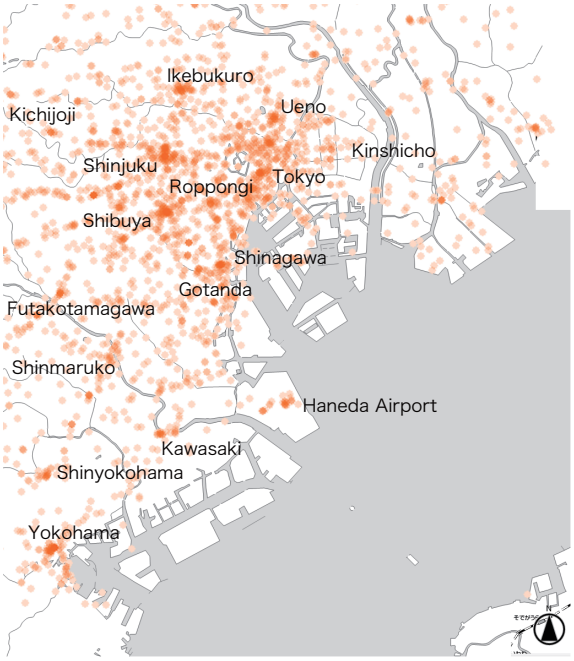

Figure 4: Users' location distribution in the morning on March 12

the direction of the suburban area.

Next, the cross tabulation result of going-home decision-making by the road network distance between offices and houses is shown in Figure 5. This result indicates that the rate of on foot decreases relatively as distance with a house becomes long, but 50% or more of people got home on foot if their distance is 20 km over.

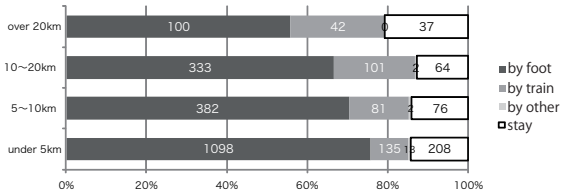

Figure 5: The relationship between going-home behavior and the distance

3.2 Verbal factors

Finally, a verbal explanation factor is generated. Since it is surmised that a family's existence and with or without information has affected going-home decision-making, the factor which affects going-home decision-making behavior is extracted from each user's tweet.

First, we analyze the effect of a safety check with a family. In this research, the family was defined as a spouse and children living together. And 353 of 3,307 persons had spoken existence of a family living together. We extracted safety check tweet such as "I got e-mail from my wife! I felt easy,", "The telephone led to the wife and the daughter at last! " and "My telephone is not connected to my son's nursery school. "

Figure 6, 7 shows the time zone rate of the safety checked tweet and the safety unidentified tweet according to going-home decision making.

Safety checked tweets are concentrated before 18:00 (42% of on foot, 45% of by train, and 65% of stay). Safety unidentified tweets are also concentrated before 18:00. We assume that the safety unidentified tweets are strongly reflecting each individual's psychological state because they can perform every time zone until safety checked. If we assume that the tweet in a earlier time zone is more important for each user, an on foot going-home person will regard his/her family's safety unidentified situation as more questionable than a railroad going-home person, and he may make decision of going-home by foot.

Next, the relationship between the information of train operation again and going-home decision-making is analyzed. The train line on the day was resumed one by one after 20:40. It is dependent on the acquisition existence of railroad resumption information whether he stays in his office or he goes home using the resumed railroad. Figure 8 shows the relationship between the rate of rail-

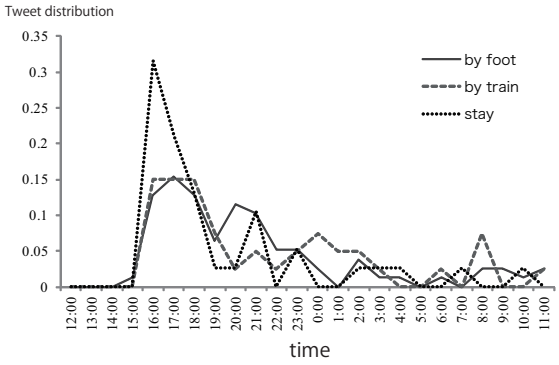

Figure 6: The distribution of safety checked tweets

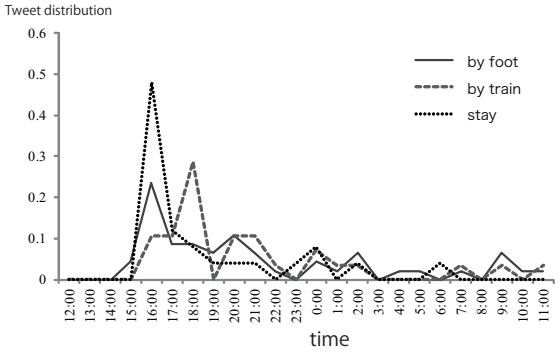

Figure 7: The distribution of safety unidentified tweets

road resumption tweet and going-home decision-making and it indicates that a railroad chooser tend to speak of railroad resumption information.

Finally, we analyze the relationship between individual psychological factor and going-home decision-making. On March 11, many utterances about their mental situation were seen. Figure 9 shows the utterance rate of uneasy and going-home decision-making result. Interestingly, individuals whose utterance rate of uneasy is under 5% tend to stay at office or hotel but people whose utterance rate of uneasy is over 5% tend to go home by foot. This results shows the person who felt fear tend to walk to home.

4 Behavioral Model

4.1 Discrete choice model

We built discrete choice model based on the explanatory variable generated in **3**. Discrete choice model is a statistical model used in fields, such as econometrics, travel behavior analysis, and marketing, and is also called Random utility model ((Ben-Akiva and Lerman, 1985); (Train, 2003)). In this research, Multinomial Logit Model (MNL) is used and it is the most fundamental model in a discrete choice model.

Discrete choice models describe decision mak-

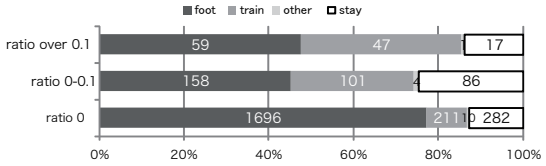

Figure 8: The relationship between the rate of railroad resumption tweet and going-home decision-making

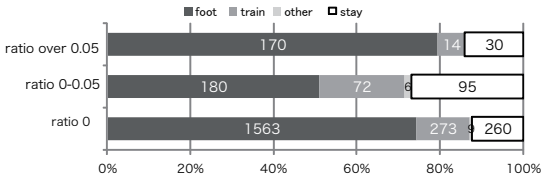

Figure 9: The relationship between uneasy tweet and going-home decision-making

ers' choices among alternatives. A decision maker, labeled n, faces a choice among J alternatives. The decision maker would obtain a certain level of utility from each alternative. The utility that decision maker n obtains from alternative j is $U_{nj},\ j = 1, \ldots, J$. This utility is known to the decision maker but not, as we see in the following, by the researcher. The decision maker chooses the alternative that provides the greatest utility. The behavioral model is therefore: choose alternative i if and only if $U_{ni} > U_{nj},\ \forall j \neq i$.

Consider now the researcher. The researcher does not observe the decision maker's utility. The researcher observes some attributes of the alternatives as faced by the decision maker, labeled $x_{nj}\ \forall j$, and some attributes of the decision maker, labeled s_n, and can specify a function that relates these observed factors to the decision maker's utility. The function is denoted $V_{nj} = V(x_{nj}, s_n)\ \forall j$ and is often called representative utility. Usually, V depends on parameters that are unknown to the researcher and therefore estimated statistically.

Since there are aspects of utility that the researcher does not or cannot observe, $V_{nj} = U_{nj}$. Utility is decomposed as $U_{nj} = V_{nj} + \varepsilon_{nj}$, where ε_{nj} captures the factors that affect utility but are not included in V_{nj}. This decomposition is fully general.

The researcher does not know $\varepsilon_{nj}\ \forall j$ and therefore treats these terms as random. The joint density of the random vector $\varepsilon_n = (\varepsilon_{n1}, \ldots, \varepsilon_{nJ})$ is denoted $f(\varepsilon_{nj})$. With this density, the researcher can make probabilistic statements about the decision maker's choice. The probability that decision

maker n chooses alternative i is

$$\begin{aligned} P_{ni} &= \Pr(U_{ni} > U_{nj}\ \forall j \neq i) \\ &= \Pr(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj}\ \forall j \neq i) \\ &= \Pr(V_{ni} - V_{nj} > \varepsilon_{nj} - \varepsilon_{ni}\ \forall j \neq i) \end{aligned} \quad (3)$$

This probability is a cumulative distribution, namely, the probability that each random term $\varepsilon_{nj} - \varepsilon_{ni}$ is below the observed quantity $V_{ni} - V_{nj}$. MNL model is derived under the assumption that the unobserved portion of utility is distributed iid extreme value.

$$f(\varepsilon_{nj}) = e^{-\varepsilon_{nj}} e^{-e^{-\varepsilon_{nj}}} \quad (4)$$

$$F(\varepsilon_{nj}) = e^{-e^{-\varepsilon_{nj}}} \quad (5)$$

And decision maker n chooses alternative i is derived as

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}. \quad (6)$$

This is choice probability of MNL model.

4.2 The setting of utility function

In discrete choice model, observed utility term V_{ni} is generally defined as $V_{ni} = \boldsymbol{\beta}' \boldsymbol{x}_{ni}$. $\boldsymbol{\beta}$ is coefficient vector and $\boldsymbol{x}_{ni}$ is explanatory vector of decision maker n's alternative i.

In this research, data set is 2672 samples identified by SVM except persons unclear and choice set is on foot, train, other and stay. Explanatory variables of on foot are required time by foot, the ratio of uneasy tweets and alternative specific constant. Explanatory variables of train are required time by train, log of the distance between office and home, the ratio of train resumption tweets, the dummy variables of family safety checked tweets and alternative specific constant. Explanatory variables of stay are the ratio of uneasy tweets, the ratio of waiting position tweets, the dummy variables of family safety checked tweets and alternative specific constant. We normalized the utility of other to 0.

Next, we outlines the estimation method of the coefficient parameter of a utility function. MNL model's likelihood function is written as

$$LL(\beta) = \sum_{n=1}^{N} \sum_{i} \delta_{ni} \ln P_{ni} \quad (7)$$

where δ_{ni} is Kronecker delta if decision maker n choice i, $\delta_{ni} = 1$ and otherwise $\delta_{ni} = 0$. This likelihood function is globally concave (McFadden, 1974). Therefore, parameters can be estimated uniquely with a maximum likelihood estimation.

Table 2: The estimation result of MNL model

variables	estimator	t-value
required time (min/10) [foot, train]	-0.012	-2.20
log(distance(km)) [train]	0.36	5.50
the ratio of train resumption [train]	4.17	5.72
the ratio of train uneasy [foot]	6.05	2.71
the ratio of train uneasy [stay]	4.52	1.82
the ratio of waiting position [stay]	2.98	4.52
family safety checked [train, stay]	1.14	3.54
alternative specific constant [foot]	4.88	18.50
alternative specific constant [train]	2.46	8.48
alternative specific constant [stay]	3.08	11.61
observations	2672	
initial log likelihood	-3704.179	
final log likelihood	-2107.771	
likelihood ratio index($\overline{\rho}^2$)	0.428	

4.3 the results and simulation

Under the above setting, the estimation result is shown in Table 2. A likelihood ratio index is 0.428 and its goodness of fit is good enough. Moreover, the result that the coefficient parameter of the required time is negative and the choice probability of train increases as the distance between office and home is far is suitable for basic analysis and intuition,

Moreover, we estimated parameters of the rate of the uneasy tweet separately by on foot and stay. It turns out that the uneasy tweet rate has had bigger influence to on foot choice. For example, from the ratio of parameters, the increase of 5 point uneasy tweet ratio is equivalent to the increase of 64 minutes required time by foot. From a perspective of family safety check, decision maker who could check family's safety tend to choice stay. Therefore, family's safety check is the important factors for the avoidance of confusion at the great disaster.

A sensitivity analysis is conducted based on this result. One is the analysis of the effect of the existence of a stay place on going-home behavior and another is the analysis of effect of family's safety check in the early time zone. Figure 10 shows the results.

First, we consider the case where all people have the waiting place. If the ratio of waiting position tweets of users who choose by foot, train and other is same as the average ratio by stay choosers, the number of choice staying will increase by 1.18

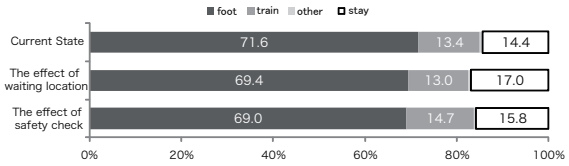

Figure 10: The result of sensitivity analysis

times and the share of stay is 17.0%.

On the other hand, the share of going-home behavior such as on foot and train decreases by 3%. Although 3% of reduction seems to be very small influence apparently, generally the traffic congestion and confusion in a transport system occur by exceeding only 10 % of supplied capacity. From this point, 3% of reduction effects is not few.

Next, we analyzed the influence of the safety check within a family. It is checked from the tweets that there are 353 decision makers who have family living together. When all of these 353 persons was able to check family's safety by 17:00, as shown in Figure 10, the number of agents who choice train or stay increase by 1.1 times, and the number of people who go home by foot decrease by 0.95 time. Needless to say, the safety check within a family at the time of a disaster is the important information. Since lines of communication other than a mobile phone carried out the big contribution by this earthquake disaster, these communication tools can prevent the confusion of transport network partially.

5 Conclusions

In this paper, we inferred the going-home behavior in Tokyo metropolitan area after the Great East Japan Earthquake using tweet data and geotag data of Twitter and clarified the decision-making factors. Although the inference method of going-home behavior and the behavioral model were the existing techniques, by combining two data sources and techniques, the going-home behavior for each individual and its factors were clarified only from Twitter data. And the virtual scenario simulation was carried out and we analyzed the effect of waiting space and communication tools.

In the ex post survey about the behavior in the earthquake disaster, the orders of samples is about thousands of people. In this research, the number of users whose tweets were with geotag is 3,307 people in Tokyo metropolitan area and it is also same order. However, if we can calculate the similarity of users who have geotag and not have geotag from the similarity of users' tweet, human behaviors in the great disaster can be clarified in hundreds thousands of people's order. We would like to consider these approach as future tasks.

Acknowledgments

We specially thank the Great East Japan Earthquake Big Data Workshop and Twitter Japan.

References

Ben-Akiva, M. and Lerman, S. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., von Schreeb, J. 2011. Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. *PLoS Medicine*, 8(8), e1001083.

Lu, X., Bengtsson, L. and Holme, P. 2012. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences of the United States of America*, 109(29), 11576–11581.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, Academic Press, New York, 105–142.

MeCab Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/.

Train, K. 2003. *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge.

Survey Research Center. 2011. Survey of the Great East Japan Earthquake disaster ("victims unable toreturn home"). http://www.surece.co.jp/src/press/backnumber/20110407.html.

Measures Council for Victims Unable to Return Home by Earthquake that directly hits Tokyo Area. 2012. Measures Council for Victims Unable to Return Home by Earthquake that directly hits Tokyo Area Final Report. http://www.bousai.metro.tokyo.jp/japanese/tmg/kitakukyougi.htm.

Yuhashi, H, 2012. Returning-Home Situation and Information Behavior in the Great East Japan Earthquake. Japan Society for Disaster Information Studies 14th workshop. A-4-2. 140–143.