Journal of the Eastern Asia Society for Transportation Studies, Vol. 6, pp. 2574 - 2588, 2005

ORIGIN-DESTINATION ESTIMATION USING CELLULAR PHONE BS INFORMATION

Byeong-Seok YOO
Senior Consultant
SCM Business Group
Samsung SDS
159-9 Gumi, Bundang, Seongnam, Gyeonggi
463-810 Korea
Fax: +82-2-6484-0055
E-mail: ybs74@hotmail.com

Seung-Pil KANG
Associate Professor
Graduate School of Engineering
Seoul National University
San 56-1 Shillim-Dong, Gwanak-Gu, Seoul
151-742 Korea
Fax: +82-2- 888-7144
E-mail: kangsp@snu.ac.kr

Kyungsoo CHON
Professor
Graduate School of Engineering
Seoul National University
San 56-1 Shillim-Dong, Gwanak-Gu, Seoul
151-742 Korea
Fax: +82-2-889-0032
E-mail: chonks@snu.ac.kr

Si-Gon KIM
Associate Professor
The Professional Graduate School of
Railroad Science and Technology
Seoul National University of Technology
172 Gongrung2-Dong, Nowon-Gu, Seoul
139-743 Korea
Fax: +82-2-975-6876
E-mail: sigonkim@snut.ac.kr

Abstract: Origin-Destination (O/D) is important information in route choices and trip assignments. In this study, methodology of estimating O/D based on cellular base station information is proposed. GPS coordinates and cellular base station coordinates have acquired on taxies in Cheongju. Three weeks' data were collected and used. Scatter diagram and sample correlation coefficients show that there are few significant differences of O/D pattern among weeks. However, there is a difference in O/D pattern between weekday and weekend. Besides, there is a difference between morning peak and evening peak. Two methodologies are proposed to convert cellular base station-based O/D into an administrative district-based O/D - using the distribution pattern ratio of GPS coordinates and using the coverage area ratio of the base stations. The statistical analyses through scatter diagram, MAPE and RMSE show that there is few significant difference of pattern between the estimated BS-based O/D and GPS-based O/D.

Key Words: Origin-Destination, Cellular Phone, Base Station, GPS

1. INTRODUCTION

Origin-Destination is information of the origin and the destination of trip. This is considered as one of the important information in transportation planning and management, especially in route choices, trip assignments and traffic simulation.

A household interview survey is deemed to be the traditional and the most widely used method in making sample O/D and its conversion to the total O/D. The accuracy of this method depends on sample size of survey and sincere answers of the respondent, moreover it require tremendous budget and time.

Some researchers have been studying on estimating O/D from the relationship between links' volumes and trip assignment model. They have concentrated on the estimation of dynamic O/D.

Nowadays, owing to the rapid spread of cellular phones, location information of the cellular phone is considered as a good alternative in O/D estimation. Location Information of cellular phone is gathered by wireless communication and can be collected in large samples. Because of using the pre-established system, it needs little budget to collect samples and to manage the system.

The location data from cellular phone can be XY coordinates or a base station(BS) coordinates or ID according to the system. In this study, only the base station coordinates were gathered due to the systematic constraints.

In this study, the methodology of generating cellular BS-based O/D and the methodology of converting this O/D into an administrative district(AD)-based O/D are proposed and the accuracy and the usefulness will be validated by various statistical analyses.

2. LITERATURE REVIEW

Origin-Destination is essential information in transportation planning, traffic operation and traffic simulation and so on. But because the accurate estimation is very difficult, it has been studying by various methodologies.

O/D is estimated and forecasted by administrative districts because of the homogeneity of the zones, the ease collection of socio-economic data, and the heterogeneity with other outer zones. The researches on O/D estimation can be divided into three categories according to the methodologies and techniques.

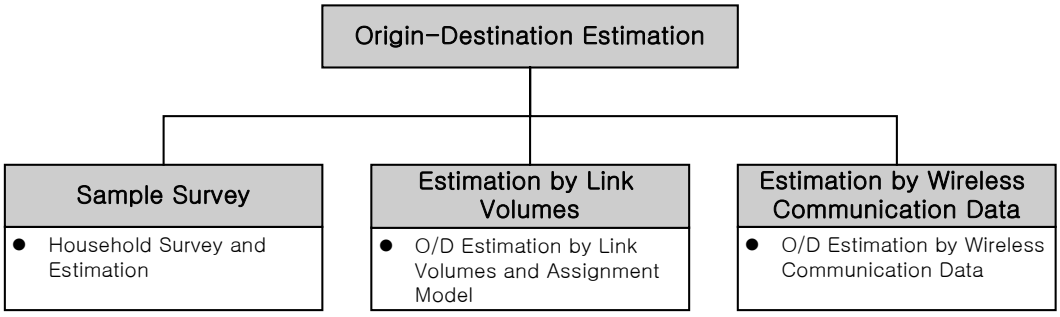

Figure 1. O/D Estimation Techniques

These techniques have their own strong points and weak points. Sample survey conduct household survey by mail or household interview and make sample O/D and convert it to the total O/D by the ratio of the sample and population. But only the small sample can be collected by this method and furthermore it is totally affected by the survey result even though it may be wrong sample. And the error is still much increased in conversion process of the sample O/D to total O/D. It even more needs much budget and time in estimating O/D.

Inductive loop detector technology has been in use for the detection of vehicles since the early 1960's. Researchers have been interested in estimating O/D from the relationship between links' volumes and trip assignment model. There were many researches in this methodology by Spiess(1987), Wu and Chang(1996), Ashok(1996), Hazelton(2000), Sherali et al(2003), and Lo and Chan(2003) etc. They have also concentrated on the estimation of dynamic O/D.

Their algorithms usually depend on the initial input O/D and they are constrained by the detector data where it is gathered and how often it is gathered. Furthermore, detector should be established to the large network and maintained regularly which needs much expense.

From the 2000's, the supply of cellular phones has been spreading out all over the world. In the case of GPS, the US government on 1 May 2000 removed SA(Selective Availability) which is an intentional error term. This measure improved the accuracy of GPS. Researchers started to have an interest in estimating O/D from the mobile data source such as cellular phone and GPS. These mobile data source don't need to establish detectors or other facilities. Besides, they can collect large samples. However, only conceptual idea and a pilot study were conducted in estimating O/D from mobile data source, and there were few researches that use real data.

In this study, methodology of estimating O/D from cellular BS information is proposed and the estimating O/D will be validated by various statistical analyses.

3. O/D ESTIMATION PROCEDURE BASED ON MOBILE INFORMATION

O/D tables are categorized by the transportation modes and travel purposes and time. Therefore, proper technique has to be chosen to estimate O/D and be applied according to the goal.

There are several location identification techniques such as E-CGI(Enhanced Cell Global Identity), A-GPS(Assisted Global Positioning System), TDOA, AOA and so forth. In these techniques only E-CGI and A-GPS methods can be used at detecting the location of cellular phones in Korea. E-CGI method can identify the base station that is serving a cellular phone. And the coverage of BS is known as over several hundreds meter to a few kilometers. On the other hand, A-GPS method can confirm a phone's location within about 50 meters.

In this study, the methodology of estimating O/D based on base station of cellular phone is presented like Figure 2.

Base stations' coordinates of cellular phones are gathered at every time interval, such as every a few seconds or every a few minutes from the sample that are determined prior. These data can be categorized into moving or standing by the change of the coordinates. If a phone is standing on a certain base station over a pre-defined threshold time, it can be inferred that the phone-owner finished his/her trip, and this makes an O/D.

Transportation mode can be identified by rule-based algorithm. For example, pedestrian walks below 5 kph speed. Bus or train would stops at stops or stations. Usually passenger car has a faster speed than bus. These characteristics can make a rule-based algorithm. And this makes an O/D by modes. However, it seems to be difficult in urban cities with an exclusive bus lane. It may be possible by wide consideration of mode characteristics.

Cellular BSs may be divided into residential area, commercial area, school zone, shopping area, and so on according to the zone characteristics. Using these BS-based zone characteristics, trip purposes such as commuting, shopping etc may be inferred and this makes an O/D by purposes.

Finally, hourly O/D can be calculated by aggregating along time. In this way, cellular phone BS-based O/D could be estimated by time, purposes, and modes. The next step is to convert this cellular BS-based O/D to AD-based O/D. This can be accomplished by the relationship between base stations and administrative districts. Finally using the sample ratio, the sample O/D can be totalized.

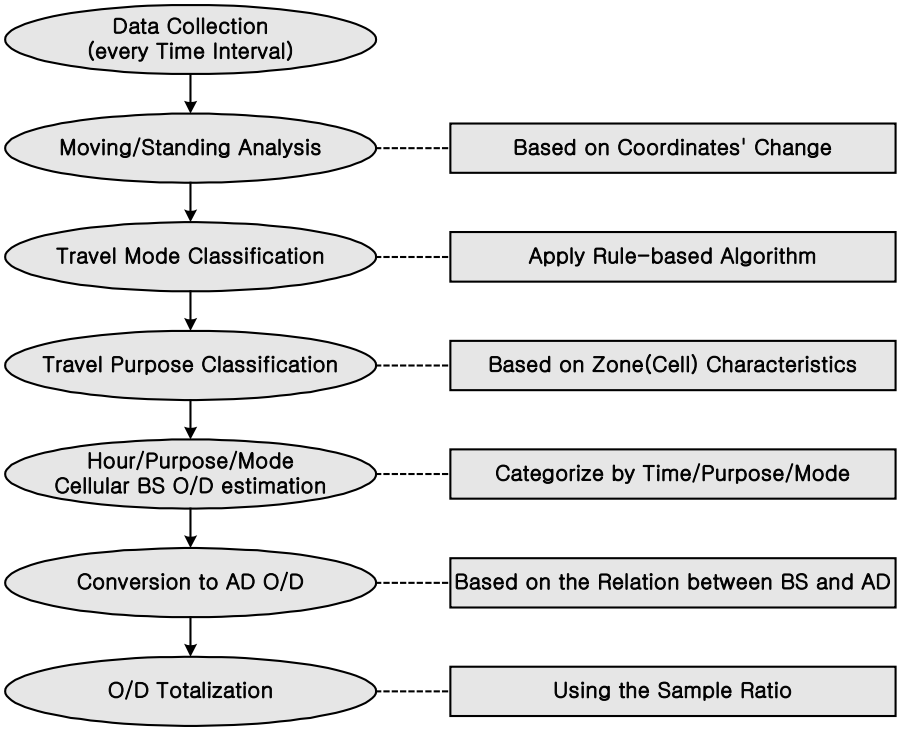

Figure 2. O/D Estimation Procedure based on Mobile Information

4. REAL DATA COLLECTION USING TAXI

A general procedure is proposed in the prior section, but cellular phone service provider is reluctant to expose the users' exact location information because of private privacy and system constraints. Only base stations' coordinates of cellular phones can be gathered in a limited extent. However, true coordinates needs to be collected for validation. For these reasons, real coordinates and BS coordinates of cellular phones should be gathered in a pair.

The information of GPS coordinates and BS coordinates of cellular phones have acquired by establishing GPS equipments and cellular phones on taxies in Cheongju by the help of Openpass Inc. The data acquisition system is like as Figure 3.

The data is composed of GPS information and cellular BS information. GPS information is composed of vehicle ID, GPS coordinates, time, and passenger existence code. Cellular BS information consists of vehicle ID, cellular base station coordinates, and time. Because the data is gathered by means of taxi, the paired data is transmitted to center whenever passenger boards the taxi and after then the data is transmitted at every 5 minutes until the passenger alights taxi. Hence, the passenger's O/D can be gathered directly. Because GPS coordinates and cellular BS is known, GPS-based O/D(AD-based O/D) and Cellular BS-based O/D can be collected.

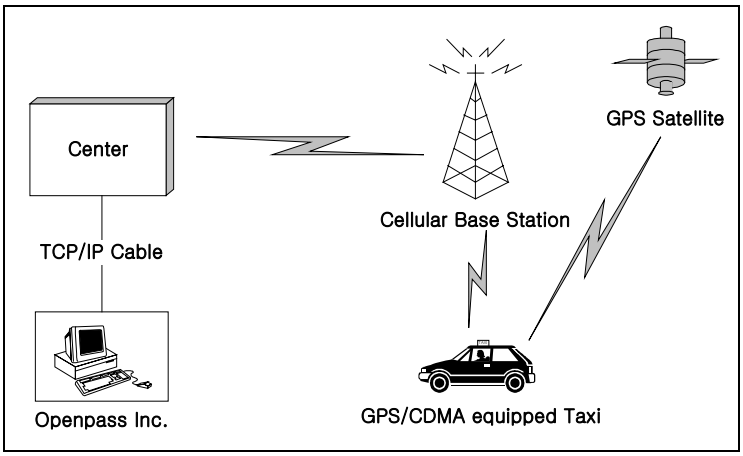

Figure 3. Data Acquisition System

The real data were gathered for 31 days from January 13, 2004 to February 12, 2004 with the average 10 probe vehicles every day. Total 8,964 O/D trips were identified. After eliminating the data of national holidays, total 6,398 O/D trips of 21 days were used in the study.

Only Taxi O/D is collected in this study. However, the characteristics of O/D patterns can be identified and the methodology to convert BS-based O/D into AD-based O/D can be studied.

5. TAXI O/D CHARACTERISTICS ANALYSES

5.1 Total O/D Trips

Before the analysis of cellular BS-based O/D, the GPS-based O/D is analyzed to have an insight of the characteristics of taxi O/D trips.

Collected real data is composed of 6,398 O/D trips for 21 days like as follows.

Table 1. Collected O/D Trips

Date	Vehicles	Trips	Date	Vehicles	Trips	Date	Vehicles	Trips
01/14 (Wed)	16	355	01/28 (Wed)	11	360	02/04 (Wed)	11	235
01/15 (Thu)	19	386	01/29 (Thu)	9	228	02/05 (Thu)	13	282
01/16 (Fri)	16	305	01/30 (Fri)	11	169	02/06 (Fri)	10	322
01/17 (Sat)	14	428	01/31 (Sat)	13	169	02/07 (Sat)	12	301
01/18 (Sun)	14	306	02/01 (Sun)	14	232	02/08 (Sun)	11	236
01/19 (Mon)	15	464	02/02 (Mon)	11	275	02/09 (Mon)	9	250
01/20 (Tue)	13	473	02/03 (Tue)	12	297	02/10 (Tue)	13	325
Sum	107	2,717	Sum	81	1,730	Sum	79	1,951

If total O/D trips are divided into hours and days, it seems to be Figure 4. There are differences in O/D frequencies along time, namely there were fewer trips before morning peak and there were more trips in evening peak than morning peak. This result can be come from the number of running vehicles and real demands.

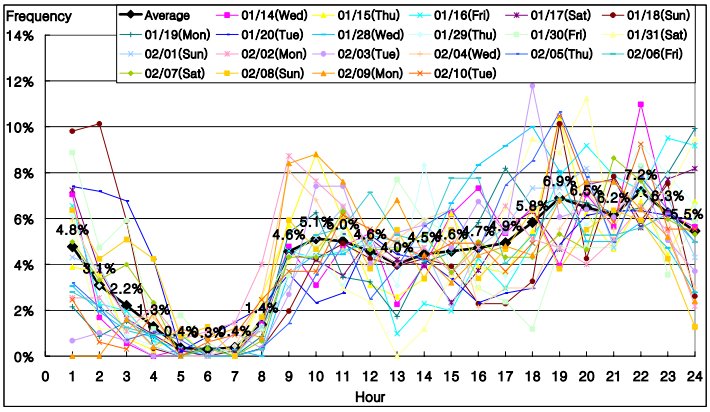

Figure 4. Hourly O/D Frequency

Figure 5 shows O/D variations along days. There are many taxi trips on Monday, Tuesday, and Wednesday, and there are fewer taxi trips on Thursday, Friday, Saturday, and Sunday. Especially, the trips are the least on Sunday that means people use automobiles rather than taxi even though the demand of Sunday may be more than that of weekday.

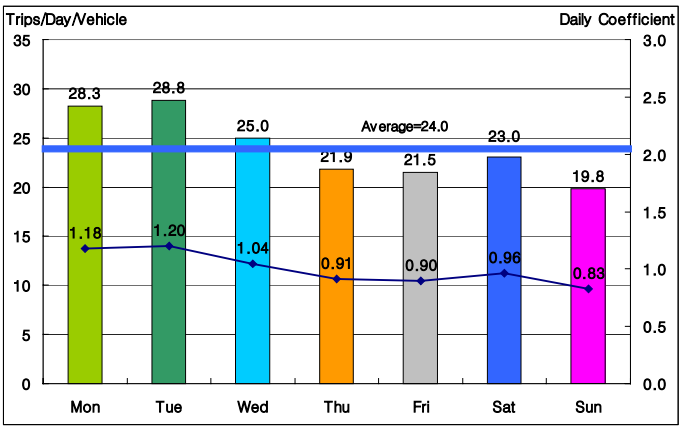

Figure 5. Daily O/D Trips and Daily Coefficient

5.2 Weekly O/D Trips Comparison

Weekly O/D Trips are compared to identify the differences between weeks. There were more trips in the first week that caused from not only the number of probe vehicles which is more than the others but also the seasonal characteristics that the week was before national holidays – people buy gifts for families before holidays, therefore there are more trips usually before national holidays. However, in the case of average O/D trips per vehicles, there were little differences between 1 week and 3 week.

Table 2. Weekly O/D Characteristics

Week	Period	O/D Trips	Average Vehicles per day	Average O/D Trips per vehicles
1 Week	1.14~1.20	2,717	15.3	25.4
2 Week	1.28~2.03	1,730	11.6	21.4
3 Week	2.04~2.10	1,951	11.3	24.7

To identify the difference between O/D patterns, O/D frequency ratio is compared.

$$O/D\ Frequency\ Ratio(\%) = \frac{T_{ij}}{\sum T_{ij}} \times 100 \qquad (1)$$

where, T_{ij} = O/D trips from i to j

Scatter diagrams with this O/D frequency ratio are depicted among three weeks and sample correlation coefficients are calculated to identify the linear correlation between weeks.

There are positive correlations above 0.83 between weekly O/D pairs, this means that O/D pattern is not different between weeks. Furthermore, this means that sample O/D can be gathered at any week without distorting the O/D pattern.

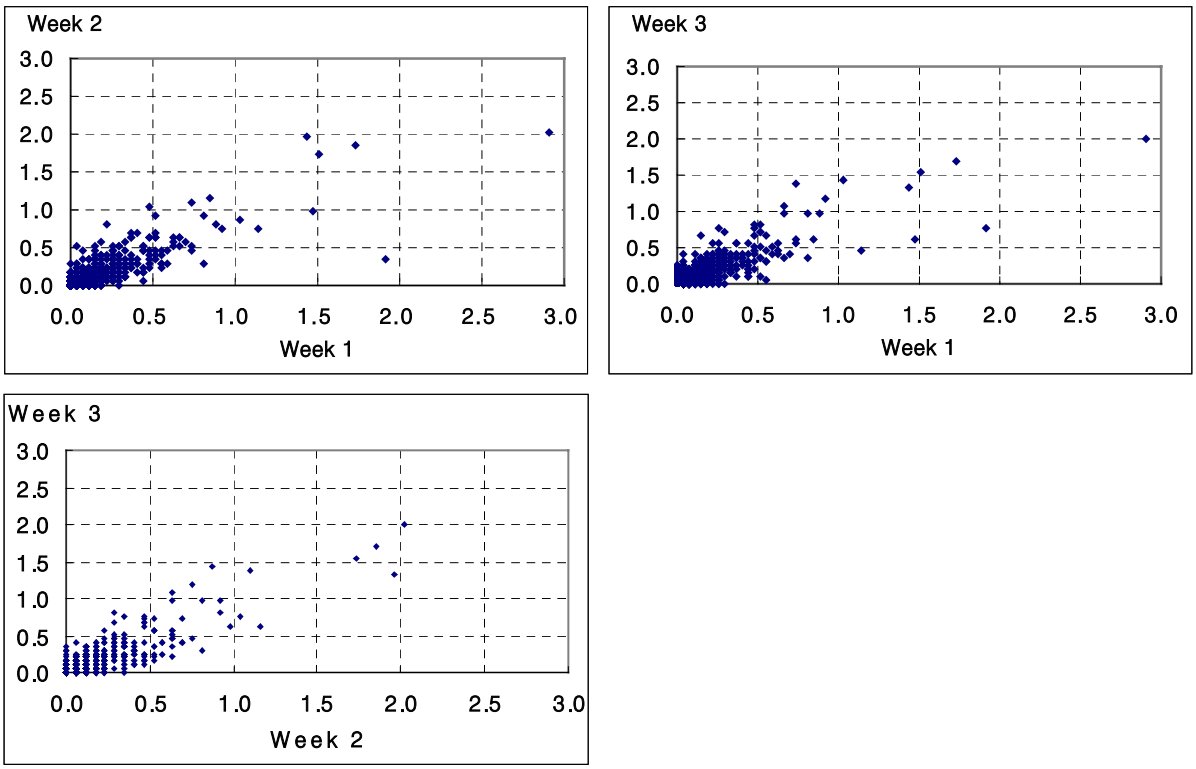

Figure 6. Scatter Diagram between Weeks

Table 3. Sample Correlation Coefficient between Weeks

Weeks	Sample Correlation Coefficient
Week 1 – Week 2	0.8316
Week 1 – Week 3	0.8311
Week 2 – Week 3	0.8448

5.3 Daily O/D Trips Comparison

Daily O/D Trips are compared to identify the differences between weekday and weekend. Usually the O/D patterns between weekday and weekend are considered as to be different because of commuting and leisure trips. Total 3 weeks data are classified into weekday, Saturday and Sunday like Table 3.

Table 4. Daily O/D Characteristics

Day	Period	O/D Trips	Average Vehicles per day	Average Daily O/D Trips per vehicle
Weekday	15 days	4,726	12.6	25.0
Saturday	3 days	898	13.0	23.0
Sunday	3 days	774	13.0	19.8

Weekday has more O/D trips per vehicle rather than that of Sunday. This means that there is low taxi demand on Sunday. This result coincides with general idea about taxi O/D because people use automobiles rather than taxi on Sunday.

According to the scatter diagram and sample correlation coefficients, weekday and Saturday have strong correlation but weekday and Sunday has low correlation and even more there is low correlation between Saturday and Sunday. This means that there are differences between O/D patterns among days of week.

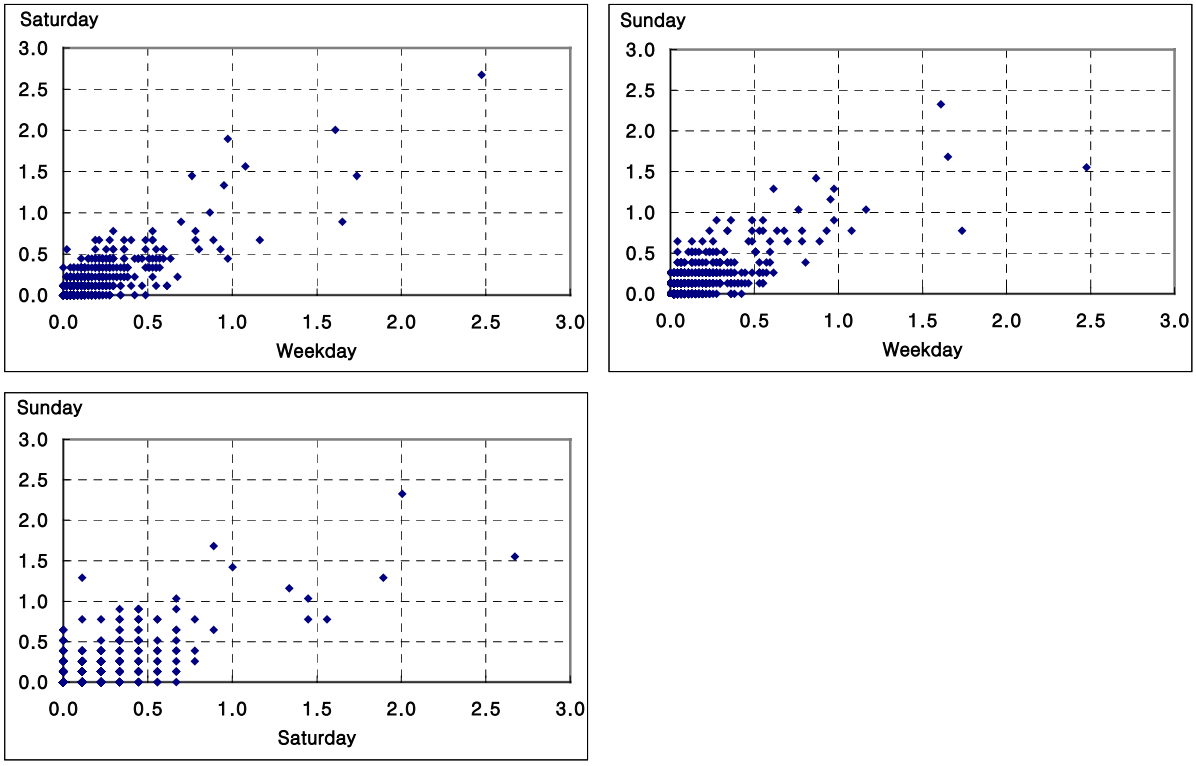

Figure 7. Scatter Diagram between Days

Table 5. Sample Correlation Coefficient between Days

Days	Sample Correlation Coefficient
Weekday – Saturday	0.8186
Weekday – Sunday	0.7743
Saturday - Sunday	0.6979

5.4 Hourly O/D Trips Comparison

In transportation planning and demand management, various O/D need to be estimated. These O/Ds consist of average daily O/D and peak hour O/D and so on. In this reason, average daily O/D and peak hour O/D are compared. The average hourly O/D trips per vehicle are lower in a morning peak than in the evening peak.

Table 6. Hourly O/D Characteristics

Hour	Period	O/D Trips	Average Daily Vehicles per hour	Average Hourly O/D Trips per vehicle
Day	21 days – 24 hour	6,398	4.0	3.2
Morning Peak	21 days – 2 hour	380	3.1	2.9
Evening Peak	21 days – 3 hour	1,229	5.7	3.4

Morning Peak: 07:00 – 09:00, Evening Peak: 17:00 – 20:00

According to the scatter diagram and sample correlation coefficients, hourly O/D trips show different O/D patterns. Especially, there is little correlation between morning peak and evening peak. But they don't show negative correlation that generally shows in a commuter trip. This means that taxi O/D pattern is different between morning peak and evening peak, but they don't show clear negative correlation between morning peak and evening peak.

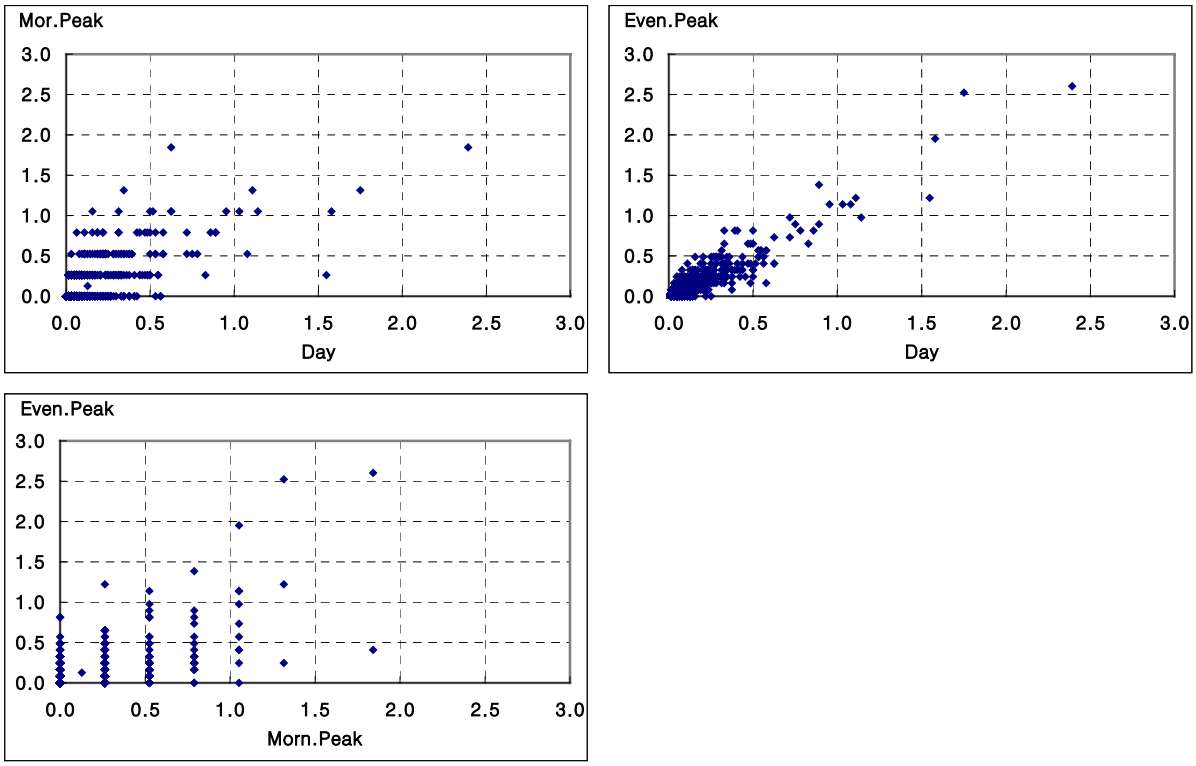

Figure 8. Scatter Diagram between Hours

Table 7. Sample Correlation Coefficient between Hours

Hours	Sample Correlation Coefficient
Day – Morning Peak	0.6772
Day – Evening Peak	0.9240
Morning Peak – Evening Peak	0.5893

5.5 Results Summary

According to the above analyses, taxi O/D shows same pattern between weeks, but there are differences in O/D patterns between weekday and weekend. Also, there is a difference in O/D patterns between morning peak and evening peak, but they don't show negative correlation.

These results coincide with general O/D characteristics, and can be inferred that researchers had to pay attention to what is proper O/D table according to the purposes. For example, if a goal is to check the morning congestion, a morning O/D table should be estimated not a daily O/D table.

6. O/D ESTIMATION BASED ON CELLULAR BS INFORMATION

6.1 O/D Estimation based on Cellular BS Information

39 cellular phone BSs are identified by the BSs' coordinates and there are 28 ADs in Cheongju like as follows in Figure 9.

Almost every administrative district has more than one cellular base stations except for 3 districts which probably don't need to have cellular phone base stations because there is one near that district.

As mentioned earlier, the data acquisition system can calculate 39×39 cellular phone BS-based O/D tables directly with simple manipulation. Also, 28×28 AD-based O/D tables can be made simply. The next step is to convert 39×39 BS-based O/D table to 28×28 AD-based O/D table.

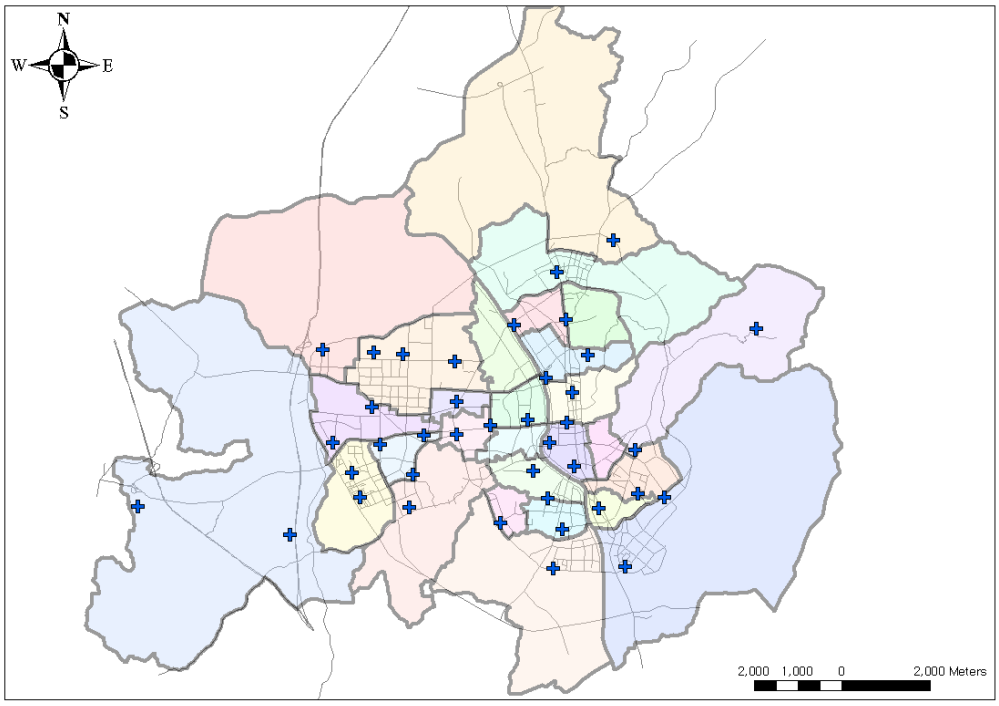

Figure 9. Cellular Phone Base Station Distribution

6.2 Conversion Techniques

Cellular BS-based O/D needs to be converted into AD-based O/D according to the relationship between BS and AD. This conversion process needs to be done because future O/D may be predicted according to the relationship between current O/D table and AD-based socio-economic characteristics.

After converting the BS-based O/D to AD-based O/D, scatter diagram and sample correlation coefficient will be analyzed. Also, MAPE(Mean Absolute Percent Error) and RMSE(Root Mean Square Error) will be calculated to verify the confidence of the conversion result.

$$MAPE = \frac{1}{N}\sum \frac{\left|T_{ij}^{*} - T_{ij}\right|}{T_{ij}} \times 100 \tag{2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum \left(T_{ij}^{*} - T_{ij}\right)^{2}} \tag{3}$$

where, T_{ij} = Observed trips from i to j (GPS AD-based O/D)

T_{ij}^{*} = Estimated trips from i to j (Converted O/D from BS-based O/D)

N = Total number of O/D table ($28 \times 28 = 784$)

6.2.1 GPS Distribution Ratio Techniques

In this study, both GPS coordinates and cellular phones' BSs coordinates are collected. Therefore, one can identify how many GPS coordinates are distributed in ADs in one cellular phone BS.

Figure 10 shows the relationship between GPS distribution and Cellular BS. 'X' cellular BS has many GPS coordinates that are distributed in 'A', 'B', 'C', and 'D' administrative districts. Hence, using the percentage of the GPS distribution, 39×39 BS-based O/D table can be converted into 28×28 AD-based O/D table.

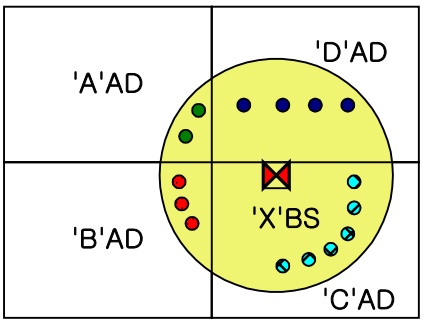

Figure 10. Cellular BS and GPS Distribution Ratio

According to this GPS distribution technique, 28×28 AD-based O/D table is calculated and the result can be seen in Figure 11 and Table 8. The sample correlation coefficient is 0.9320 and the scatter diagram shows high positive correlation between observed O/D and estimated O/D. In this case, MAPE is 53.99 % that seems to be relatively low value.

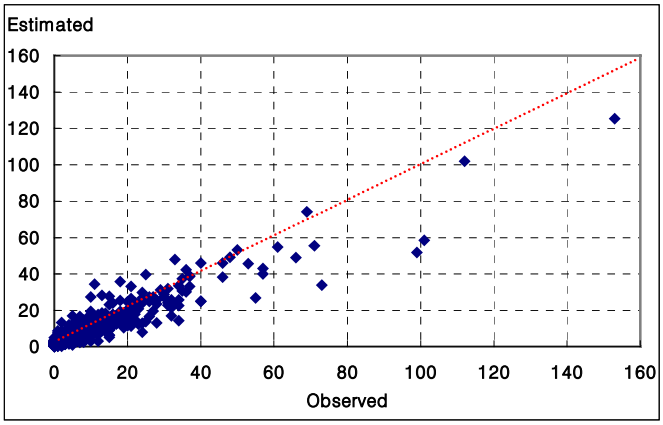

Figure 11. Scatter Diagram between Observed and Estimated by GPS Distribution Ratio

Table 8. Statistical Result Using GPS Distribution Ratio

	GPS Distribution Ratio Techniques
Sample Correlation Coefficient	0.9320
MAPE	53.99
RMSE	4.88

To use this technique, advance survey about GPS distribution should be conducted. Also, it cannot be applied to another region. Therefore, this technique has limitation to be applied in real world.

6.2.2 Cellular BS Coverage Area Ratio Techniques

Every cellular base station has its service coverage, and this coverage covers administrative districts to some extent. This relationship can be depicted like Figure 12. 'X' cellular BS covers the 'A' -AD to x % and cover the 'B'-AD to y% etc. Using these area ratio of cellular phone BSs, 39×39 BS-based O/D table can be converted into 28×28 AD-based O/D table. The coverage of each cellular base station is like as Figure 13.

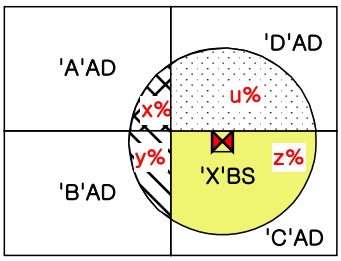

Figure 12. Cellular BS Coverage Ratio

According to this BS coverage area ratio technique, 28×28 AD-based O/D table is calculated and the result can be seen in Figure 14 and Table 9. The sample correlation coefficient is 0.8338 and the scatter diagram shows high positive correlation between observed O/D and estimated O/D. In this case, MAPE is 84.13 % that seems to be relatively low value. Therefore, this BS coverage area ratio technique can be a good conversion technique in converting the sample BS-based O/D table to AD-based O/D table.

Also this technique has a strong advantage that can be applied at any region without any additional survey, as long as the coverage area is provided from mobile communication company.

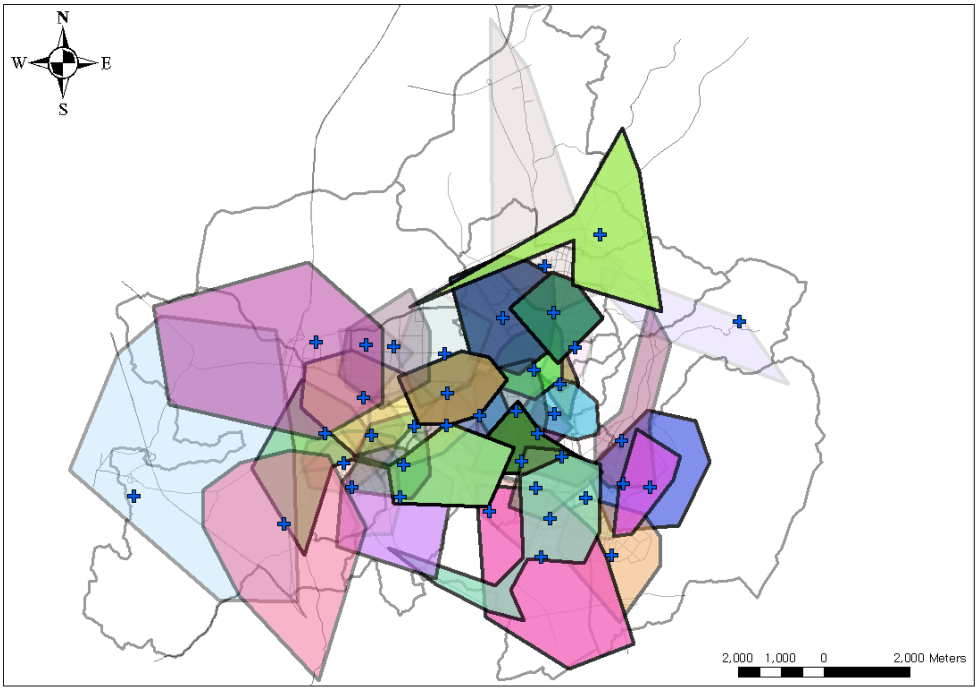

Figure 13. Cellular Phone Base Station Coverage

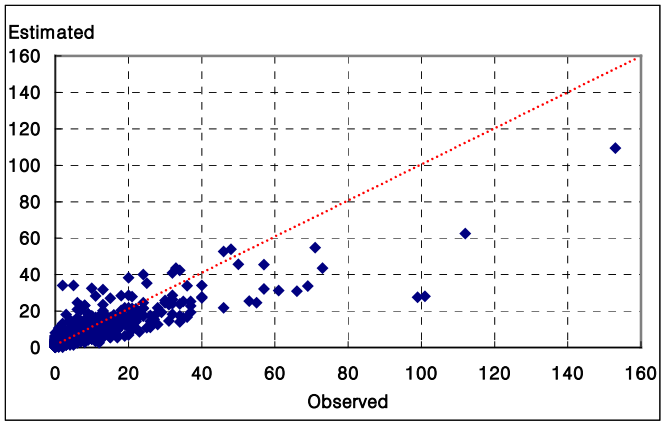

Figure 14. Scatter Diagram between Observed and Estimated by BS Coverage Area Ratio

Table 9. Statistical Result Using BS Coverage Area Ratio

	GPS Distribution Ratio Techniques
Sample Correlation Coefficient	0.8338
MAPE	84.13
RMSE	7.38

7. TOTALIZATION TECHNIQUE

After estimating sample O/D table, O/D table should be totalized. If the goal is estimating daily O/D table, totalization can be achieved using the number of running taxi and the number of sample taxi. First, one can conduct sampling and converting to make AD-based O/D table. And then using the number of probe vehicles, average daily vehicular O/D table can be calculated. Then, using the total number of running vehicles, daily O/D table can be totalized. This totalization process can be described like Figure 15.

$$Pop_{ij} = Sam_{ij} \times TC_{ij} \tag{4}$$

where, Pop_{ij} = Population O/D

Sam_{ij} = Sample O/D

TC_{ij} = Totalization Coefficient

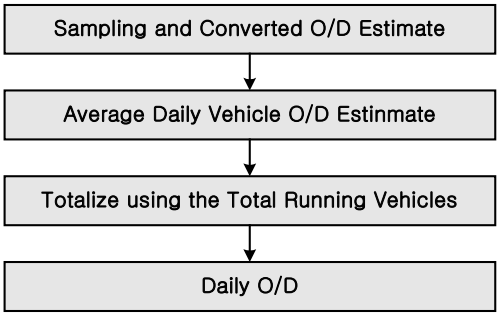

Figure 15. Totalization Process

The totalization coefficient is the number of total running taxies in this study, and the number of total running taxies was 2,710. Total Taxies' O/D table can be estimated by using this coefficient and the converted AD-based O/D from BS-based O/D with BS coverage area ratio technique as proposed earlier.

Table 10. Taxi Service Status (January 2004 ~ February 2004)

	No. of Registered Taxies	Suspension Ratio	No. of Daily Running Taxies
Company Taxi	1,481	1/6	1,230
Owner-driver Taxi	2,220	1/3	1,480
Sum	3,701		2,710

8. CONCLUSION

Researchers started to focus on the mobile data such as GPS and cellular phone in making transportation information. In this study, methodology of estimating O/D table from cellular base station information is proposed.

Analyses on GPS-based O/D pattern show that there is a difference both between weekday O/D pattern and weekend O/D pattern and between morning peak O/D pattern and evening peak O/D pattern.

Two techniques are presented to convert cellular base station-based O/D table to administrative district-based O/D table – one by using GPS distribution ratio and the other by using cellular base station coverage area ratio. Two techniques show that they are good estimators by judging from scatter diagram, correlation coefficient, MAPE and RMSE. Between these two techniques, cellular base station coverage area ratio may be used at any region without any additional survey.

In this study, only taxi was used as a sampling mode, but in the future information of all modes would be collected. In such case, more algorithms should be studied such as travel mode classification algorithm, and trip purpose classification algorithm, and so on.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to Ministry of Construction & Transportation and Engineering Research Institute SNU by which they are supported, for giving them the opportunity to carry out this paper.

REFERENCES

Ashok, K.(1996), Estimation and Prediction of Time-Dependent Origin-Destination Flows, Ph.d Dissertation, MIT

Bang, S.S.(2001), A Study on Trip O/D Construction using Position Data from Cellular Mobile Station, Master Dissertation, Seoul National University

Hazelton, M.L.(2000), Estimation of origin–destination matrices from link flows on uncongested networks , **TR-B, Vol. 34**, 549-566.

Korea Transportation Institute.(2004) Travel Speed Estimation and O/D Estimation Technique Development based on LBS,

Lo, H.P. and Chan, C.P.(2003), Simultaneous estimation of an origin–destination matrix and link choice proportions using traffic counts, **TR-A, Vol 37**, 771-788.

Lo, H.P. et al.(1999), Decomposition algorithm for statistical estimation of OD matrix with random link choice proportions from traffic counts, **TR-B, Vol. 33**, 369-385.

Ormsby, T. et al.(2001), **Getting to know ArcGis Desktop**, ESRI Press.

Sherali, H.D. et al.(2003), Estimation of origin-destination trip-tables based on a partial set of traffic line volumes, **TR-B, Vol 37**, 815-836

Spiess, H.(1987), A maximum likelihood model for estimating origin-destination matrices, **TR-B, Vol. 21**, 395-412.

Wu, J. and Chang, G.L.(1996), Estimation of time-varying origin-destination distributions with dynamic screenline flows**, TR-B, Vol.30**, 277-290.