

DEVELOPMENT OF 3D MONITORING DATASET FOR SHOP AND OFFICE TENANTS VARIATIONS IN BROAD URBAN AREA BY SPATIO-TEMPORAL INTEGRATING DIGITAL MAPS AND YELLOW PAGE DATA

Yuki Akiyama^{a,*}, Takeshi Shibuki^b, Satoshi Ueyama^a, Ryosuke Shibasaki^c

^aGraduate School of Frontier Science, The University of Tokyo, Japan – aki@iis.u-tokyo.ac.jp, uym@csis.u-tokyo.ac.jp

^bDepartment of Civil Engineering, The University of Tokyo, Japan – tshibuki@iis.u-tokyo.ac.jp

^cCenter for Spatial Information Science, The University of Tokyo, Japan – shiba@csis.u-tokyo.ac.jp

KEY WORDS: Spatial Database, Spatio-temporal Data Mode, Image Understanding, Spatial-temporal Analysis, Urban Planning, Digital Mapping

ABSTRACT:

Various methods of analysis and dataset have been used for urban studies. However, used urban data often has low spatial resolution, especially regional statistics, though statistical data can cover large areas with homogeneous quality. For detailed analysis, some studies rely on field surveys which have fine spatial resolution, but fail to cover entire urban areas or large regions including suburbs and rural areas. Many prior studies also indicated the need of detailed time-series urban dataset all over Japan. In this study detailed time-series urban dataset in broad area have been developed. This dataset have been developed to integrate multi-year digital house maps and yellow page data using GIS system and developed programs in this study. Developed dataset cover South Kanto region for the period from 1995 to 2005 every five years. South Kanto region is the geopolitically important area of Japan including Tokyo metropolitan area. Using this dataset, not only name, business category, and detailed location including three-dimensional information but also time-series variation of almost all shops and offices can be monitored. Time-series variation means continuation, change, emergence, demise and immigration from other room or floor of tenants between two times.

1. INTRODUCTION

1.1 Background

So far many kinds of dataset were developed for understanding “urban space”. However, development of dataset with high spatial resolution and reliability needs a large amount of labour. Therefore, they are able to cover limited area. On the other hand, spatial resolution and reliability of dataset it can cover large area is not so fine. Local statistical information is one of the examples of them.

Since the collapse of the asset-inflated economy in Japan (after 1991), wide range of urban problems have happened, for example, urban decay at many local cities. Time-series urban dataset which can closely monitor changes of urban space has been demanded to analyze and solve such problems. Previous studies also pointed out this necessity (Yasuyuki, M., 1994).

In contrast, many skyscrapers are built in major large cities and as a result large cities increase its density of houses, shops and offices. Therefore, it seems that data which can monitor urban spaces in three dimensions is needed. Moreover, property investments expand in energetic urban areas in recent years. Due to the expansion of property investments, pace of urban change is increasing now as never before.

To make a more detailed urban analysis, a method is required that can monitor spatial information including three dimensional distribution as micro as possible. A development of dataset is demanded that can monitor time-series variations

of almost all shop and office tenants in broad area or all of urban areas, prefectures and national land.

To realize such dataset, a development of methodology is needed that integrates appropriate existing multi-year spatial information at low cost and short time.

1.2 Previous studies

Most studies have also tried to develop spatio-temporal urban dataset using existing information.

Itai developed 3D digital map around the Kawaguchi station (Saitama prefecture, Japan) that visualizes foundry industrial decline and urbanization (Youichi, I., 1994). This dataset was developed using distributional information of foundries in 1974 and 1994. 3D time-series variation of foundries can be monitored using this dataset. However, base maps used in this study made manually. Therefore this method is not considered proper to develop dataset that can cover broad area.

Ito developed a methodology to integrate digital house maps between different two years (Kaori, I., 2001). The methodology used in her study is helpful to us in conducting our research. However, this methodology was limited to integrate digital house maps. Moreover, dataset developed in her study is covered only the special words of Tokyo. Due to this limitation, it is difficult to apply this methodology in widely area than it.

In contrast, some studies have collected time-series information using the questionnaire survey to local governments without relying on existing spatial information (Junichiro, A., 2002).

* Yuki Akiyama (M.Env.). Cw-503, Institute of Industrial Science, 4-6-1, Komaba, Meguro-ku, Tokyo, 153-8505, Japan, Phone: +81-3-5452-6417, Fax: +81-3-5452-6414, e-mail: aki@iis.u-tokyo.ac.jp

The questionnaire survey can collect faithful information. However, collectable data can cover limited areas or buildings due to a property of questionnaire survey.

Urban space consists of elements that have many kinds of functions, for example, commercial facilities, living accommodation and industrial installations. Not only shops and offices locations and names but also their functions, namely “business categories” are needed to monitor detailed variation in urban space.

1.3 Purpose of this study

The dataset which we tried to develop in this study is required for performance as follows. It has spatial information as micro as possible, it can cover broad area, namely throughout urban area and even national land, it has 3D location information of shops and offices, it has business categories of shops and offices, and it can monitor not only current information but also some past points of time.

The purpose of this study is the development of methodology to create the dataset for urban analysis that satisfies foregoing requirements. This study has also revealed a processing accuracy of this system and reliabilities of source data.

The dataset developed in this study covers South Kanto region (Tokyo, Saitama, Chiba and Kanagawa prefecture) in 1995, 2000 and 2005. South Kanto region is a geographically-diverse area, for example, this area contains Tokyo metropolitan area, suburbs and rural regions.

Finally, we introduce some studies and results using the dataset developed in this study.

2. DEVELOPMENT OF THE DATASET

This section explains a development flow and the methodology to develop the dataset. First we introduce source data and explain properties of them. Secondly, we explain how to develop the dataset.

The dataset covers whole area of South Kanto region. However, processable area of spatio-temporal integration system developed in this study is not limited in South Kanto region. This system can deal with data in whole of Japan.

2.1 Source data

We notice variations of “tenants” to monitor diverse phenomenon in urban space. The “tenant” in this study means all kinds of shops and offices without personal residences. Tenant information is collected from digital house maps and yellow page data of Japan. Both data contain micro scale data, namely tenant scale data, the whole of Japan. Digital house maps cover 1817 cities, towns and villages out of 1827 in April 1, 2007. It is published by ZENRIN CO., LTD. On the other hand, yellow page data is “Town Page database” published by NTT Business Information Service, Inc. The Town Page database includes information of about ten-million tenants. Town Page dataset is the largest yellow page data of Japan. Table 1 shows information held by both data. Both data have tenant names and location information. Digital house maps have more detailed location information than yellow page data. In contrast, yellow page data have business categories.

2.2 Development flow of the dataset

The system developed in this study integrates digital house map with yellow page data. After that processing, integrates multi-year digital house maps that have information of yellow page data. Figure1 shows an integration flow of digital house maps and yellow page data in 1995, 2000 and 2005. Using spatio-temporal house maps developed in this study, not only almost all tenant information in 1995, 2000 and 2005, but also time-series variations of tenants can be monitored.

“Time-series variations” mean continuations, changes, emergences, demises and immigration from other room or floor of all tenants. Figure2 shows types of time-series variations in this study.

	Digital house map	Yellow page data
Tenant names	○	○
Business categories	×	○
Building information	○	△1)
- Building names	○	△2)
- Tenant-occupied floor	○	△2)
- Tenant-occupied room	○	△2)
Number of stores	○	×
Shapes of building	○	×
Addresses	○	○
longitude and latitude	○	△3)
frequency of data update	Urban area: about every year Other: about every 1-3 years	About every 2 months

1) In case that one tenant occupies one building, here is no data.

2) They are extracted from building information of yellow page data (A method is explained later.)

3) They are created from address using address geocoding system.

Address geocoding system (<http://pc035.tkl.iis.u-tokyo.ac.jp/~sagara/geocode/>)

Table 1. Information held by house map and yellow page data

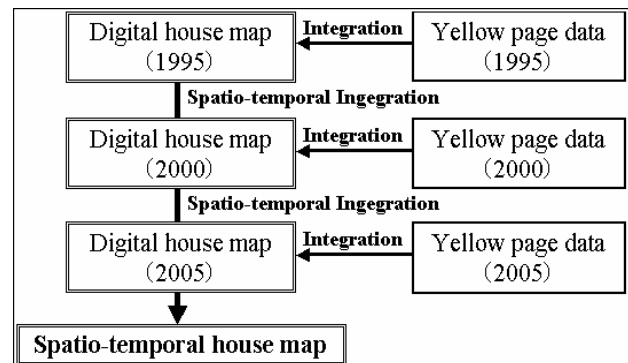


Figure 1. Integration flow of source data

2.3 Challenges of dataset development

A method to integrate digital house maps with yellow page data is required to produce this dataset. Moreover, a method to specify time-series variation of tenants between different two years is also required. Automation of these methods is needed because amount of data is large.

To realize these methods, a methodology to integrate source data based on location information is needed. In addition, a methodology to evaluate similarity of integrated tenant names is also needed.

Thus, a development of spatial integration methodology based on longitude, latitude, address and building information and evaluation methodology of tenant names are main tasks.

2.4 Integration of location information

Depending on location information which tenant data have, one tenant data are integrated other tenant data that has the most similar location. Figure 3 shows an integration algorithm of location information.

In the case of yellow page data, building information of one tenant is one text. In other words, it is not divided into building name, floor, and room number. Therefore, we have developed a division pattern library for building information using building information of yellow page data at the special words of Tokyo in 2005.

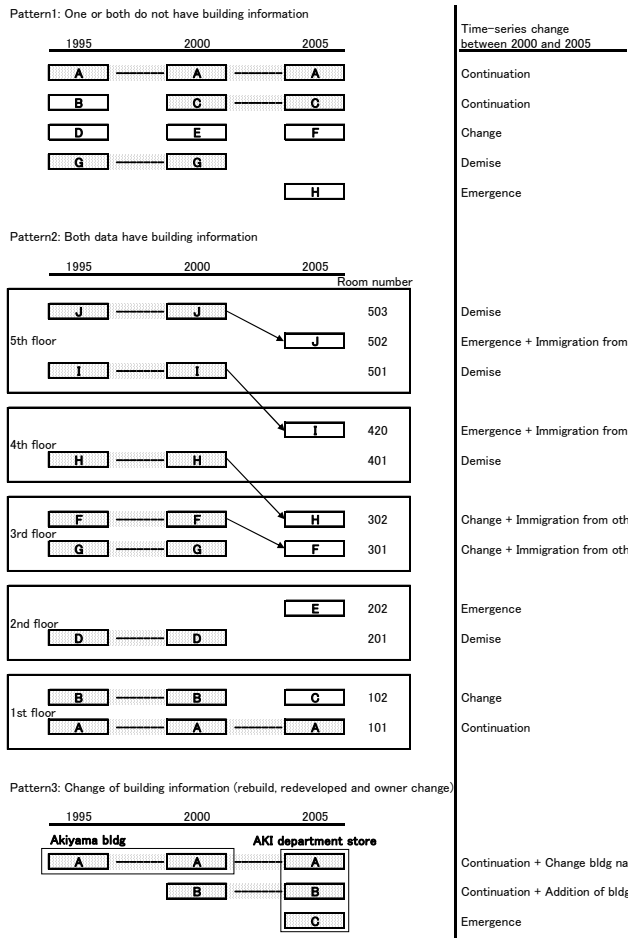


Figure 2. Types of time-series variation

Using this library, building information of tenant data is divided into building name, floor, and room number. Figure 4 shows an example of building information division.

The location integration methodology developed in this study is able to deal with not only longitude and latitude, but also nonquantifiable location information, namely, address and building information. It is a great difference between location integration in this study and integration of typical GIS software.

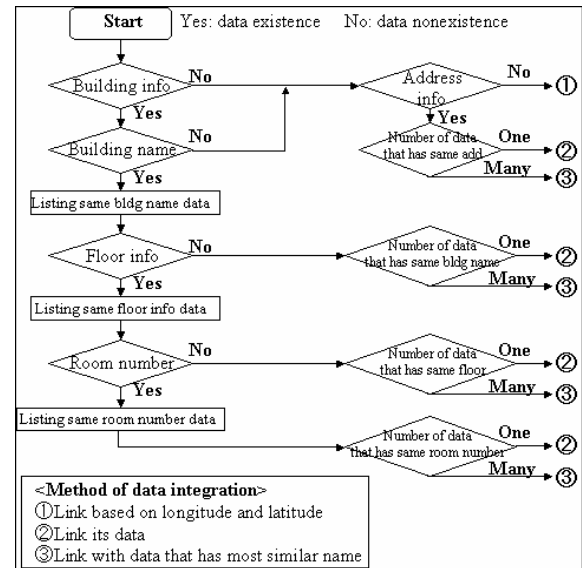


Figure 3. Integration algorithm of location information

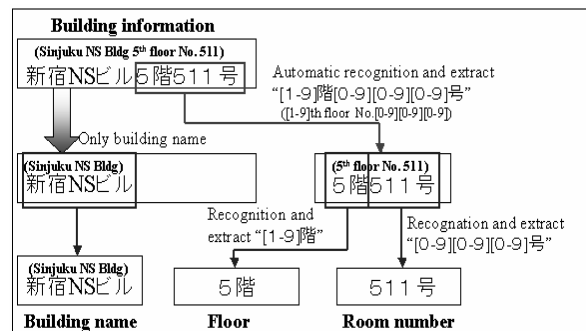


Figure 4. Example of building information division

2.5 Identification judgment of tenant names

An identification of tenants which are integrated based on location information is judged based on their tenant names. A quantification methodology of name similarity, namely, similarity between two words is needed to realize this identification. Tenant names are spelled in different ways even they are same tenants. Due to this fluctuation of description, a quantification of name similarity is demanded.

2.5.1 Quantification of text similarity: In this study, the "n-gram" has quantified name similarity. The n-gram is one of the methods of natural language processing. The n-gram is able to quantify a similarity between two words or texts. Much attention has been paid to the n-gram in the fields of literature and linguistic in recent years (Miyuki, K., 2000). The "bi-gram" calculates name similarity in this study. The bi-gram extracts neighbouring two characters and compares them. Figure 5 shows an example of the be-gram processing. Equation 1 defines similarity between word i and word j.

$$S_{ij}^{(n)} = \frac{n_{ij}^{(n)} + n_{ji}^{(n)}}{m_i^{(n)} + m_j^{(n)}} \quad (1)$$

where $S_{ij}^{(n)}$ = name similarity between word i and word j

$m_i^{(n)}$ = the number of n-letter words blocks extracted from text i
 $m_j^{(n)}$ = the number of n-letter words blocks extracted from text j
 $n_{ij}^{(n)}, n_{ji}^{(n)}$ = the number of $m_i^{(n)}$ the same as $m_j^{(n)}$

A proper threshold of $S_{ij}^{(n)}$ is 0.4. This means if $S_{ij}^{(n)}$ is more than 0.4, these words are same.

2.5.2 Removal of frequently-appearing words: Tenant names contain frequently-appearing words, geographic names, and station names. Owing to these words, identification judgment of tenant names using only the n-gram was not able to achieve accurate results. Sagara has also pointed out the difficulty to extract pure tenant names (Takeshi, S., 2006). We have developed libraries of frequently-appearing words, geographic names, and station names against this problem. The library of frequently-appearing words was developed using all tenant names at the special words of Tokyo in 2000 and 2005. The library of geographic names was developed using the “data file of Japanese administrative area handbook” published by Nihon Kajo-Syuppan corporation. The library of station names was developed using railroad timetables of Japan. Figure 6 shows an example of removal processing.

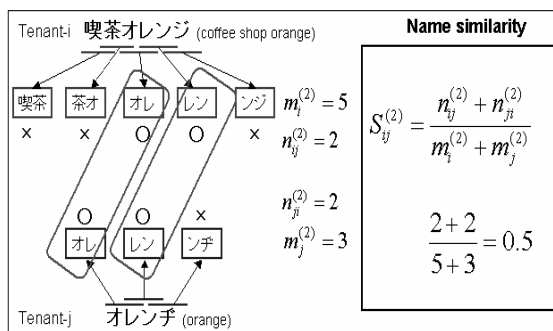


Figure 5. Example of the n-gram processing (bi-gram)

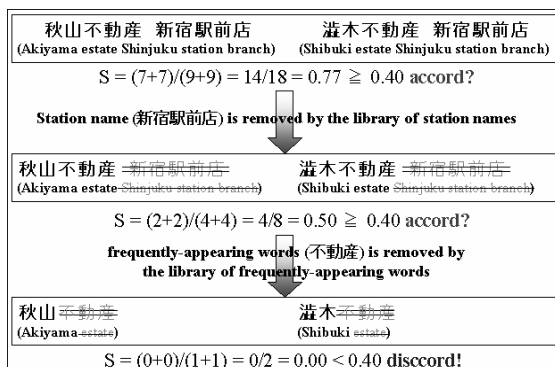


Figure 6. Removal effect of frequently-appearing words

3. ASSESSMENT OF THE RELIABILITY

We have assessed the reliability of methodologies developed in this study in this section. The processing accuracy and the reliability of source data are required to assess the reliability.

3.1 Processing accuracy of the system

In order to assess processing accuracy of the system, first, samples of each data were extracted from some areas of the special words of Tokyo. Secondly, integration results created by the system and results created by manual processing are compared. Sample areas were determined based on the tendency of tenant variations in Tokyo revealed by Ito (Kaori, I., 2001). Residential area is also added to sample areas. Figure 7 shows the tendency of tenant variations and sample areas. First digital house maps and yellow page data in 2000 and 2005 (the case of integration digital house maps with yellow page data is in 2005) were integrated by the system in each area. Secondly, 500 samples of integration results (100 samples from each area) were extracted with random sampling, and were compared with results created by manual processing. Classification ratios of time-series variations of each sample were adjusted to be in the same range of ratios of total result. Table 2 shows processing accuracy of spatio-temporal integration of digital house maps. Table 3 shows of yellow page data. Table 4 shows processing accuracy of spatial integration of digital house maps with yellow page data. Difference between accuracy to integrate each data with accuracy of spatio-temporal integration is little. In the case of spatio-temporal integration of digital house maps, 93.00 % samples accorded. In the case of spatio-temporal integration of yellow page data, 95.88 % samples accorded. In the case of integration of each data, 94.00 % samples accorded. In particular, Areas where variations of tenants are small show a high accuracy. In addition, Accuracies of continuation and integration success are excellent results in all areas.

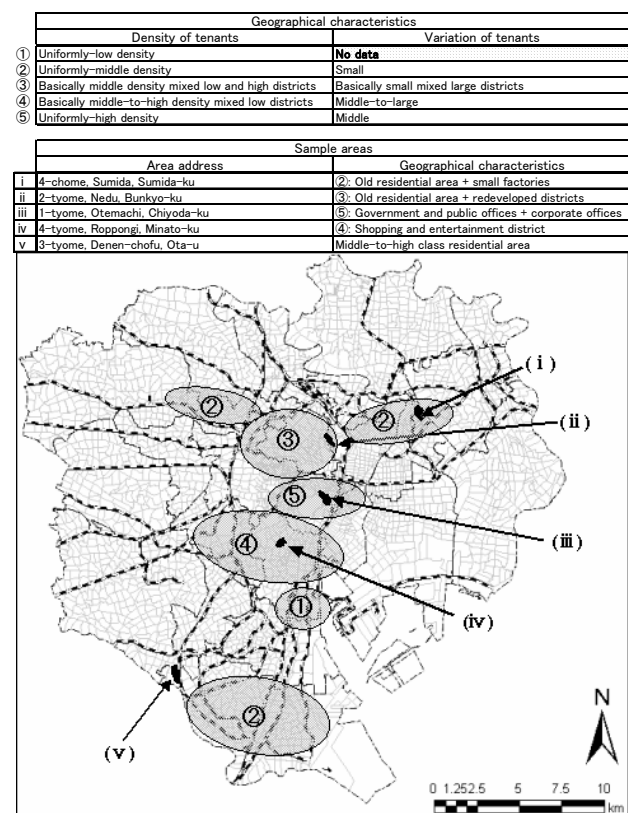


Figure 7. Tendency of tenant variations between 2000 and 2005 in the special words of Tokyo and sample areas

	4 Sumida (458)			2 Nedu (576)			4 Roppongi (1109)			1 Otemachi (773)			3 Denen-chofu (198)			Total
	Count	Accord	Accuracy	Count	Accord	Accuracy	Count	Accord	Accuracy	Count	Accord	Accuracy	Count	Accord	Accuracy	
Continuation	84	82	97.89	56	54	96.43	33	33	100.00	48	44	91.67	54	52	96.30	98.08
Change	9	8	88.89	16	15	93.75	23	19	82.61	23	17	73.91	11	11	100.00	86.37
Change + (floor)	0	0		0	0		0	0		0	0		0	0		
Change + (room)	0	0		0	0		0	0		0	0		0	0		
Emergence	8	8	100.00	15	13	86.67	16	13	81.25	19	18	94.74	13	12	92.31	90.14
Emergence + (floor)	0	0		0	0		0	0		0	0		0	0		
Emergence + (room)	0	0		0	0		1	1	100.00	0	0		0	0		100.00
Demise	19	19	100.00	13	13	100.00	27	23	85.19	10	8	80.00	22	22	100.00	93.41
Sum	100	97	97.00	100	95	95.00	100	89	89.00	100	87	87.00	100	97	97.00	93.00

* "(floor)" means immigration from other floor.

* "(room)" means immigration from other room.

* Numbers in parentheses of area names (e.g. 458 at "4 Sumida (458)") mean total number of tenants.

* "Count" means the number of sample tenants. "Accord" means the number of according tenants with result of manual processing.

* "Accuracy" = ("Accord") / ("Count")

* "Total" = (sum of "Accord") / (sum of "Count")

Table 2. Processing accuracy of spatio-temporal integration of digital house maps in 2000 and 2005

	4 Sumida (297)			2 Nedu (365)			4 Roppongi (643)			1 Otemachi (562)			3 Denen-chofu (85)			Total
	Count	Accord	Accuracy	Count	Accord	Accuracy	Count	Accord	Accuracy	Count	Accord	Accuracy	Count	Accord	Accuracy	
Continuation	80	79	98.75	62	62	100.00	45	45	100.00	41	38	92.68	46	46	100.00	98.54
Change	3	3	100.00	9	9	100.00	20	19	95.00	23	16	69.57	12	11	91.67	86.57
Change + (floor)	0	0		0	0		0	0		0	0		0	0		
Change + (room)	0	0		0	0		0	0		0	0		0	0		
Emergence	3	3	100.00	9	9	100.00	15	15	100.00	6	5	83.33	10	10	100.00	97.67
Emergence + (floor)	0	0		0	0		0	0		0	0		0	0		
Emergence + (room)	0	0		0	0		1	1	100.00	0	0		0	0		100.00
Demise	14	14	100.00	20	20	100.00	19	17	89.47	30	27	90.00	17	16	94.12	94.00
Sum	100	99	99.00	100	100	100.00	100	97	97.00	100	86	86.00	85	83	97.65	95.88

Table 3. Processing accuracy of spatio-temporal integration of yellow page data in 2000 and 2005

	4 Sumida (429)			2 Nedu (510)			4 Roppongi (863)			1 Otemachi (713)			3 Denen-chofu (163)			Total
	Count	Accord	Accuracy	Count	Accord	Accuracy	Count	Accord	Accuracy	Count	Accord	Accuracy	Count	Accord	Accuracy	
Integration success	33	33	100.00	39	38	97.44	32	30	93.75	36	33	91.67	28	27	96.43	95.83
Integration failure	13	10	76.92	15	15	100.00	17	15	88.24	14	12	85.71	9	8	88.89	88.24
Only yellow page	14	13	92.86	5	5	100.00	13	12	92.31	8	6	75.00	7	7	100.00	91.40
Only house maps	40	40	100.00	41	40	97.56	38	35	92.11	42	39	92.86	56	52	92.86	94.93
Sum	100	96	96.00	100	98	98.00	100	92	92.00	100	90	90.00	100	94	94.00	94.00

* "Integration failure" means that one name discords other name despite each location is same.

Table 4. Processing accuracy of spatial integration of digital house maps and yellow page data in 2005

3.2 Reliability assessment of source data by field surveys

Information of digital house maps is collected by the field survey (verification of name plate and name board). On the other hand, information of yellow page data is collected by self-return from business proprietor. Due to this difference, digital house maps are more exhaustive data than yellow page data. However, it seems that data entry error happened due to the manual processing. In contrast, it seems that the yellow page data possesses higher reliability than digital house maps. In order to reveal reliabilities of source data, we conducted field surveys, next, compared tenant locations and names of source data with results of field surveys.

Survey areas are districts where the tenant stability is relatively high and low based on the "tenant stability" created by yellow page data in 2000 and 2005. Figure 8 shows the tenant stability in the special words of Tokyo and survey areas. A selected district where tenant stability is relatively high is the eastern area of Akabane station, and where that is relatively low is around Shimo-takaido station.

Table 5 and table 6 show results of field surveys. The "survey ratio" means the ratio of tenants that could be surveyed to all tenants in a survey district. Tenants that could not be surveyed mean the case that field surveyor could not approach or find these tenants (e.g. locked lobby of apartment building). The "Reliable rate" is defined as equation 2.

$$R_i = \frac{F_i}{S_i - D_i} \quad (2)$$

where R_i = Reliable rate
 F_i = Number of findable tenants
 S_i = Number of tenants that could be surveyed
 D_i = Number of demised tenants

88.16 % of tenant data of digital house maps and 87.74 % of that of yellow page data accord real-space information in the western side area of Akabane station. 89.51 % of tenant data of digital house maps and 91.51 % of that of yellow page data accord real-space information around Shimo-takaido station. As a result of these surveys, it can be seen that about 90 % source data of the dataset accord real space. Source data can meet the demands of our study, namely, source data can monitor detailed real space in a broad area.

4. RESULTS

In this section, we introduce some examples of study and data using the spatio-temporal dataset all over South Kanto region developed in this study.

4.1 Macro scale monitoring of tenant variations

Osada developed time-series tenant dataset in the special words of Tokyo based on yellow page data in 2000 and 2005 using the system (Tatsuro, O., 2007). Next, he developed the 100-square-meter grids data that covers all over the special words of Tokyo. Finally, he calculated the "Relative changing index" which means the intensity of tenant variation of each grid. The relative changing index is defined as equation3.

$$F_a(i) = A_i - S_i R_{sum} \quad (2)$$

where $F_a(i)$ = Relative changing index in grid i
 A_i = changed and emergence tenants in grid i
 S_i = Total number of tenants in grid i
 R_{sum} = Ratio of changed and emergence tenants to total in the special words of Tokyo

Figure 9 shows 100 meter grid map of the relative changing index at the special words of Tokyo. In addition, figure 10 shows the same map zoomed in around Shibuya to Shinjuku. His study shows that the dataset developed in our study can quantitatively monitors tenant variations in broad area.

Survey areas		Continuation	Sum of tenants	Tenant stability (%)	Rank
Akabane	1-chome Akabane	523	1017	51.43	1021 / 3120
Shimo-takaido	1-chome Shimo-takaido	158	267	59.18	1907 / 3120
	3-chome Matsubara	112	204	54.90	1397 / 3120
	4-chome Akatutumi	152	268	56.72	1600 / 3120

*Tenant stability = (number of continuation tenants) / (sum of tenants)

*There are 3120 districts in the special words of Tokyo.

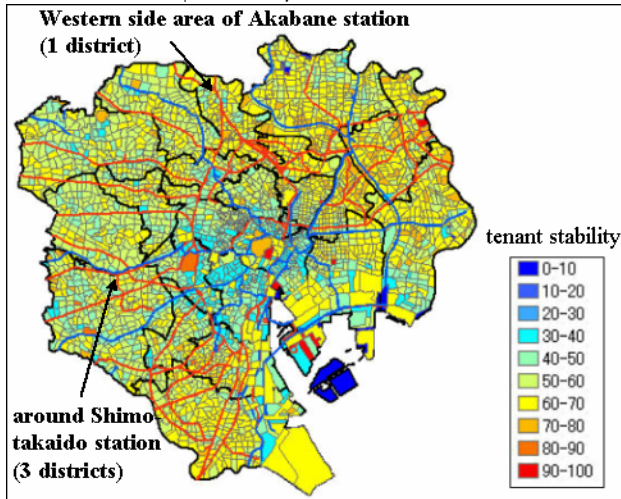


Figure 8. The tenant stability in the special words of Tokyo and field survey areas

	Digital house maps	Yellow page data
Time of data creation	~2005/06/28	2004/10/02~12/04
Time lag	About 2years and 2months	About 2years and 9months
Survey ratio (%)	97.39 (1269/1303)	96.43 (810/840)
Number of demised tenants	154	84
Reliable rate (%)	88.16 (983/1115)	87.74 (637/726)

Table 5. Results of the field survey in the eastern area of Akabane station

	Digital house maps	Yellow page data
Time of data creation	~2005/08/29	2004/10/02~12/04
Time lag	About 2years and 1month	About 2years and 10months
Survey ratio (%)	98.96 (571/577)	94.07 (349/371)
Number of demised tenants	56	31
Reliable rate (%)	89.51 (461/515)	91.51 (291/318)

Table 6. Results of the field survey around Shimo-takaido station

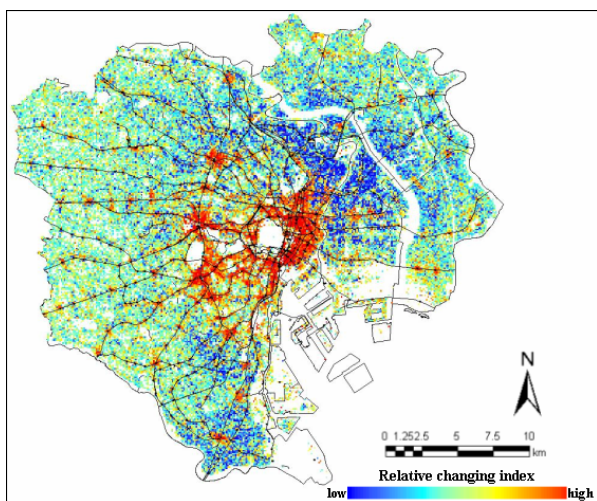


Figure 9. 100 meter grid map of the relative changing index at the special words of Tokyo (2000 to 2005)

4.2 3D visualization of tenant variations

Figure 11 shows 3D maps of tenant variations between 2000 and 2005 using spatio-temporal dataset by time-series processing of digital house maps and yellow page data. Owing to information of business categories that yellow page data has, 3D distribution of specific business category and vacant rooms or floors can be revealed. Figure 12 shows the 3D distribution map of restaurants, clothing shops and vacant rooms or floors in the western area of Shibuya station in 2005.

5. CONCLUSION

In the present work the spatio-temporal urban dataset which can monitor micro scale data of tenants has been developed. The Spatial integration methodology based on not only longitude and latitude but also nonquantifiable information, namely addresses and building information and the identification methodology of texts which considers ambiguity of texts have been proposed. Using these methodologies, the dataset which we aimed to develop has been realized. Information of business categories and time-series variations of almost all tenants can be monitored by the dataset. Source data are commercially available data. Moreover, the processing of data development has been automated. In conclusion, it is the practical method that variation in broad urban area can be continuously monitored at low cost using this methodology.

REFERENCES

- Junichiro, A., 2002. A Study on the Actual Condition of Big Shopping Stores' Closing and the Tendency of Its Uses After the Closing in the Central Areas of Local Cities ~From Researches to Municipalities in Hokuriku-Koshin-Etsu Area~. Architectural Institute of Japan. Planning. Environment. Eng., AIJ, No.557, pp.257-264.
- Kaori, I. and Hidekuni M., 2001. Method for Making Spatio-Temporal Data from Accumulated Information: Using the Identification by Resolving Geometric and Non-Geometric Ambiguity. Papers and Proceedings of the Geographic Information Systems Association, vol.10, pp.147-150.
- Miyuki, K., 2000. An Analysis of Japanese Classical Literature Using Character-based N-gram model. The journal of humanities, Chiba Univ, vol.29, pp.187-238.
- Takeshi, S., Toshiro, M., Syuichi, K., Hideki, O. and Masaru, K., 2006. A Method to Normalize Shop Names Using Their Address Descriptions for Accuracy Improvement of Shop Information Search from the Web. Proceedings of Data Engineering Workshop (DEWS) 2006, 2C-o1, pp.392-398.
- Yasunori, M., Noboru, H. and Katsutoshi, O. 1994. A Study on the Decline of Retail Centers and Location Change of Large-Scale Stores. Papers on City Planning, No. 29, pp.529-534
- Youichi, I., Masako, Y. and Susumu, K., 1994. Research and Valuation on Space Use on Each Floor of Buildings at Mixed Area with Dwelling and Factory Around Kawaguchi Station. Theory and Applications of GIS, Vol.2, No. 1, pp.57-63.

Tatsuro, O., Yuki, A., Takeshi, S. and Ryosuke S. 2007. A Study on Location Changes of Business Establishments by Extensive Urban Dataset and Its Accuracy Validation. In: *The 9th International Cooperative Seminar between KAGIS and GISA*, Busan, Korea, The KAGIS 10th Anniversary International Conference, pp.197.

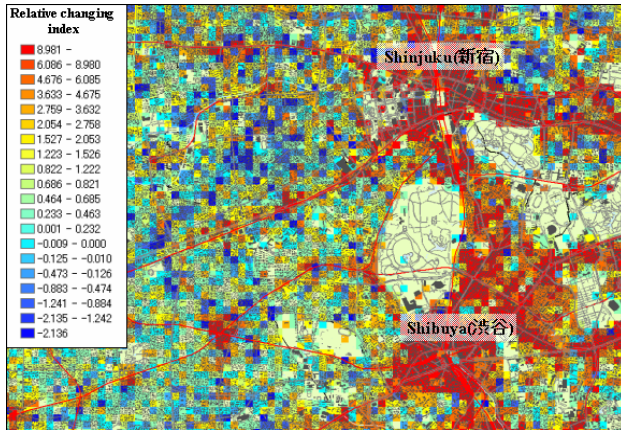


Figure 10. 100 meter grid map of the relative changing index around Shibuya to Shinjuku (2000 to 2005)

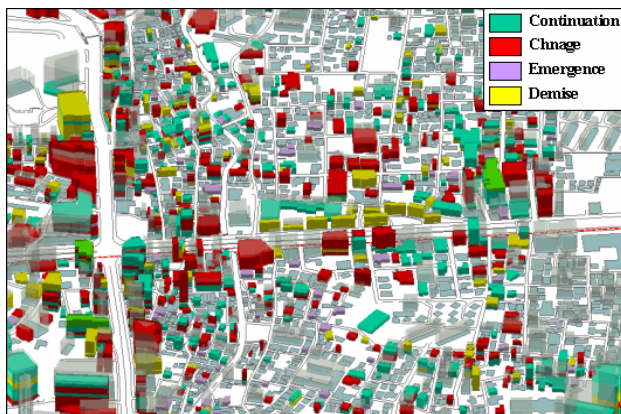


Figure 11. 3D map of tenant variations in Omote-Sando area (2000 to 2005)

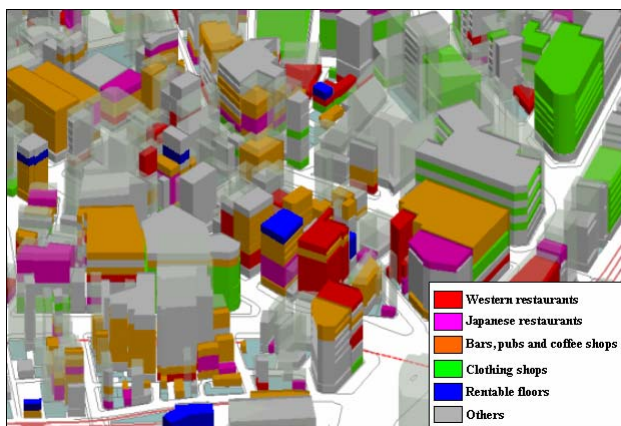


Figure 12. 3D map of restaurants, clothing shops and vacant rooms or floors in the western area of Shibuya station in 2005

