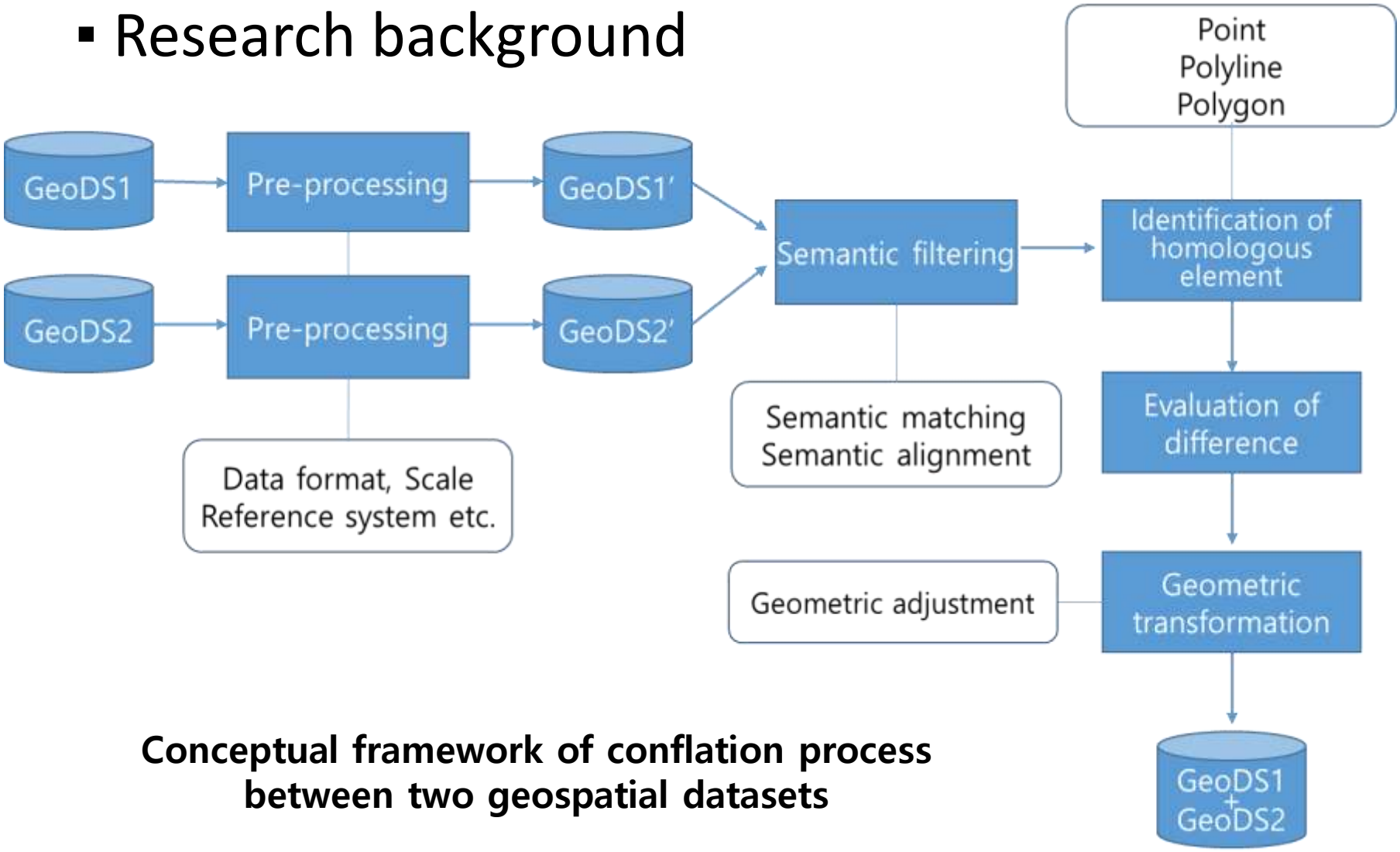
SPATIAL INFORMATION
RESEARCH INSTITUTE

Geospatial feature class integration of Urban information system with latent semantic analysis

Yong, Huh

1. Introduction

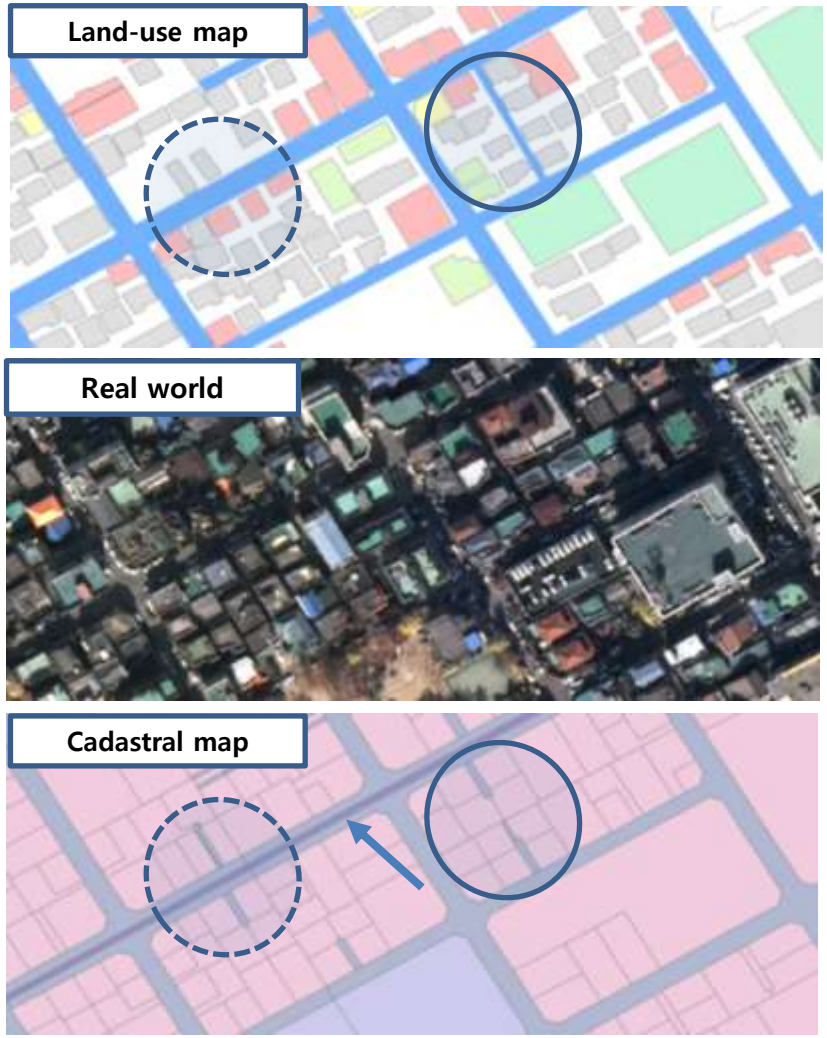
- Research background

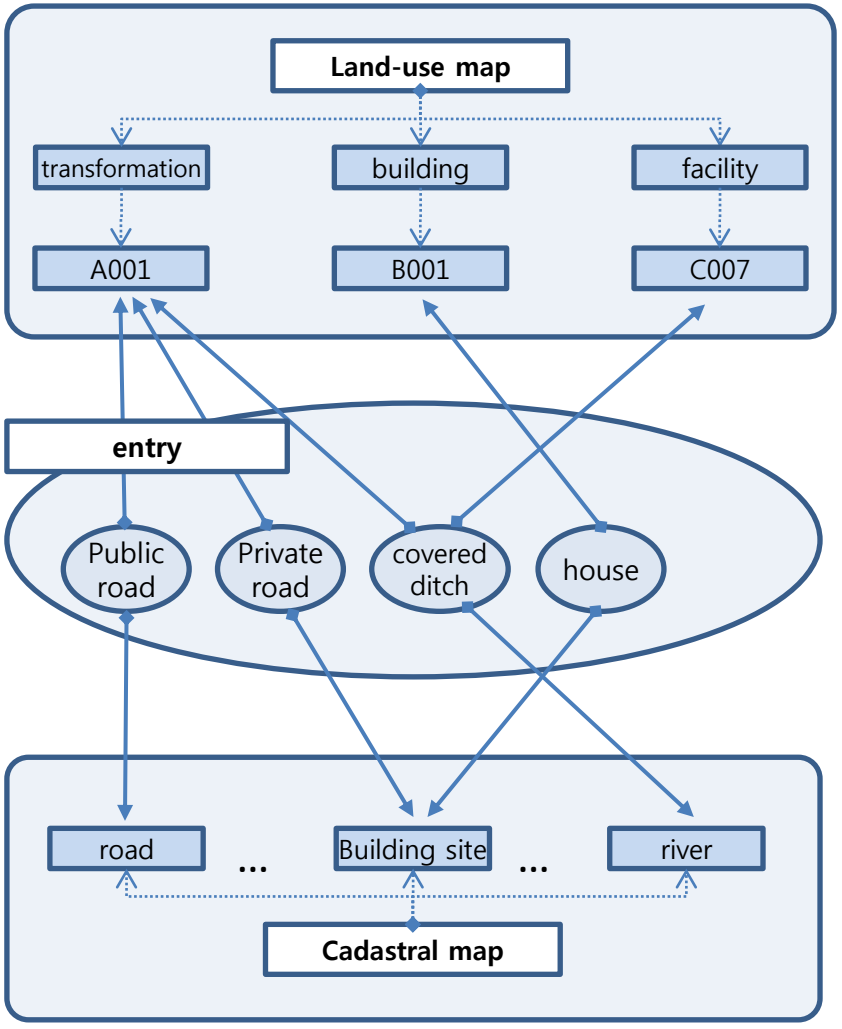


Conceptual framework of conflation process between two geospatial datasets

1. Introduction

- Issue for semantic filtering





1. Introduction

▪ Motivation of this study

❶ (Object based analysis for automation)

- If spatial objects of a certain feature type A in one spatial dataset correspond to spatial objects in another feature type B in the other dataset with a high probability
- Then, there is high semantic similarity between the feature types

❷ (Hierarchical M:N matching for semantic analysis)

- Because of various application view point even for the same real-world, feature type correspondence relations are complicated

- Motivation of this study

❸ **(Information retrieval method)**

- Similar problem in the field of word-document analysis (for web search)

❹ **(Latent semantic analysis)**

- Uses Singular Value Decomposition (SVD) to simulate human learning of word and document meaning

- Represents word and document meaning as high-dimensional vectors in the semantic space

2. Proposed method

- Latent semantic analysis

❶ (Definition)

- a technique in natural language processing, in particular distributional semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.

❷ (Mathematic property)

- A matrix containing word counts per paragraph (rows represent unique words and columns represent each paragraph) is constructed from a large piece of text and a mathematical technique called singular value decomposition (SVD) is used to reduce the number of rows …

2. Proposed method

- Latent semantic analysis

❷ **(Mathematic property)**

- … while preserving the similarity structure among columns.

- Words are then compared by taking the cosine of the angle between the two vectors formed by any two rows. Values close to 1 represent very similar words while values close to 0 represent very dissimilar word

2. Proposed method

- Latent semantic analysis

For an $M \times N$ matrix **C** of rank r there exists a factorization (Singular Value Decomposition = **SVD**) as follows:

$$C = U\Sigma V^T$$

$M \times M$ | $M \times N$ | V is $N \times N$

The columns of ***U*** are orthogonal eigenvectors of $\boldsymbol{CC^T}$.

The columns of ***V*** are orthogonal eigenvectors of $\boldsymbol{C^TC}$.

Eigenvalues $\lambda_1 \ldots \lambda_r$ of $\boldsymbol{CC^T}$ are the eigenvalues of $\boldsymbol{C^TC}$.

$$\sigma_i = \sqrt{\lambda_i}$$

Singular values.

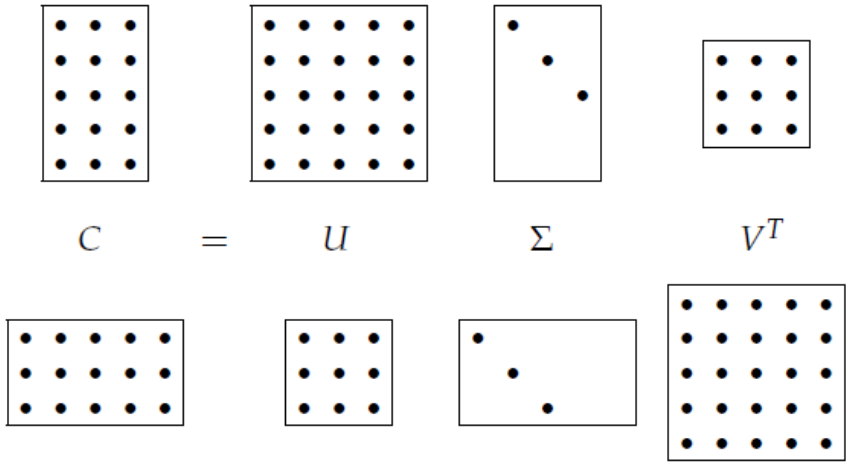
2. Proposed method

- Latent semantic analysis

$$C = U\Sigma V^T$$

$$(\mathbf{t}_i^T) \rightarrow \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix} = (\hat{\mathbf{t}}_i^T) \rightarrow \begin{bmatrix} \begin{bmatrix} \mathbf{u}_1 \end{bmatrix} \cdots \begin{bmatrix} \mathbf{u}_l \end{bmatrix} \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_l \end{bmatrix} \cdot \begin{bmatrix} [& \mathbf{v}_1 &] \\ & \vdots & \\ [& \mathbf{v}_l &] \end{bmatrix}$$

$(\mathbf{d}_j)$ $\downarrow$ $(\hat{\mathbf{d}}_j)$ $\downarrow$



2. Proposed method

- Latent semantic analysis

Index Words	Titles								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
book			1	1					
dads						1			1
dummies		1						1	
estate							1		1
guide	1					1			
investing	1	1	1	1	1	1	1	1	1
market	1		1						
real							1		1
rich						2			1
stock	1		1					1	
value				1	1				

book	0.15	-0.27	0.04
dads	0.24	0.38	-0.09
dummies	0.13	-0.17	0.07
estate	0.18	0.19	0.45
guide	0.22	0.09	-0.46
investing	0.74	-0.21	0.21
market	0.18	-0.3	-0.28
real	0.18	0.19	0.45
rich	0.36	0.59	-0.34
stock	0.25	-0.42	-0.28
value	0.12	-0.14	0.23

*

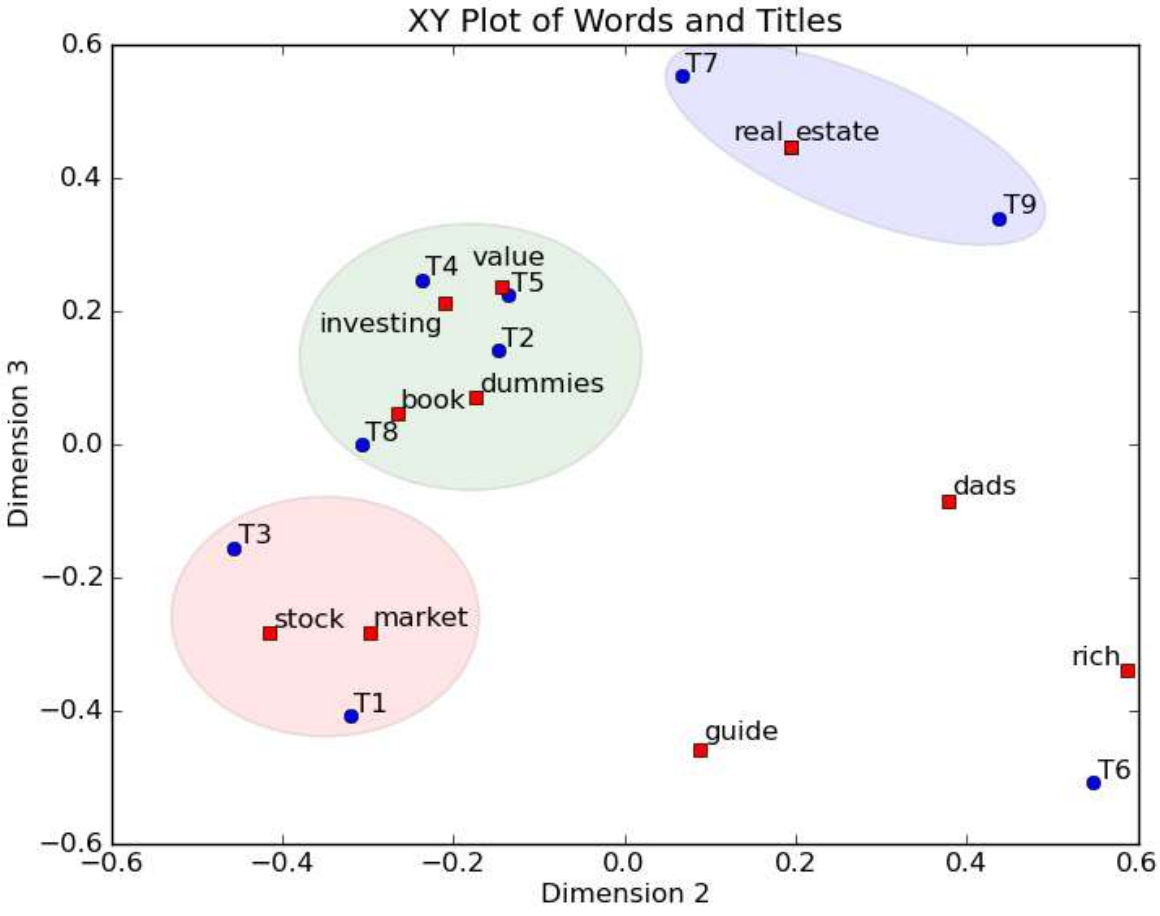
3.91	0	0
0	2.61	0
0	0	2

*

T1	T2	T3	T4	T5	T6	T7	T8	T9
0.35	0.22	0.34	0.26	0.22	0.49	0.28	0.29	0.44
-0.32	-0.15	-0.46	-0.24	-0.14	0.55	0.07	-0.31	0.44
-0.41	0.14	-0.16	0.25	0.22	-0.51	0.55	0	0.34

2. Proposed method

- Latent semantic analysis



3. Experiment

- Dataset(object intersection for about 800,000 parcels in Seoul)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	1,101	213	3	1,329	364,662	6	25	17	3	2	955	272	24	85	126	0	16	11	57	307
2	85	18	0	57	11,364	0	1	1	0	1	45	6	0	0	2	0	0	0	0	23
3	101	28	0	112	69,118	0	1	0	0	2	130	5	2	1	18	0	12	0	2	36
4	215	45	0	382	20,065	9	16	1	1	0	629	29	2	28	30	1	0	9	11	51
5	309	139	0	180	97,089	106	5	35	843	2	313	148	7	32	38	8	7	9	9	281
6	71	47	0	25	11,852	29	8	4	14	1	120	12	1	8	26	0	1	1	5	144
7	145	66	1	57	79,864	12	6	4	2	0	278	31	2	18	25	2	3	0	6	86
8	158	118	1	17	3,870	919	1	1	0	8	52	17	4	14	44	3	0	1	1	89
9	8,392	5,690	3	414	669	0	1	0	0	1	138	15	34	186	123	9	20	3	2	1,167
10	224	22	80	128	4	0	0	0	0	0	1	0	0	0	5	1	0	0	0	2
11	62	3,333	0	7	1	0	0	0	0	0	6	0	4	1	19	5	8	0	0	75
12	17	4	0	161	21	0	0	0	0	0	0	0	0	0	6	0	2	6	0	0
13	428	43	0	10,258	645	2	21	0	0	0	97	4	2	9	61	0	28	172	14	156
14	289	38	0	1,254	395	1	10	0	0	0	24	1	0	8	16	1	1	50	1	208
15	26	6	0	415	35	0	8	0	0	0	5	0	1	2	14	0	9	9	17	22
16	2,732	1,865	1	2,624	30,767	110	197	13	5	2	69,102	1,460	346	1,345	2,977	80	301	291	45	1,764
17	219	261	0	104	65	2	1	0	0	0	120	19	348	1,416	756	59	6	4	0	97
18	119	79	0	247	1,834	1	3	1	1	0	161	30	10	75	38	1	31	2,200	8	109
19	120	89	0	29	114	5	1	0	0	0	44	12	2	14	7	1	1	15	0	14
20	46	104	0	36	381	1	0	387	2	0	45	2	3	6	30	5	0	7	0	52
21	61	10	0	36	144	0	5	0	1	0	19	16	3	157	12	3	3	1	0	166

Cadastral dataset

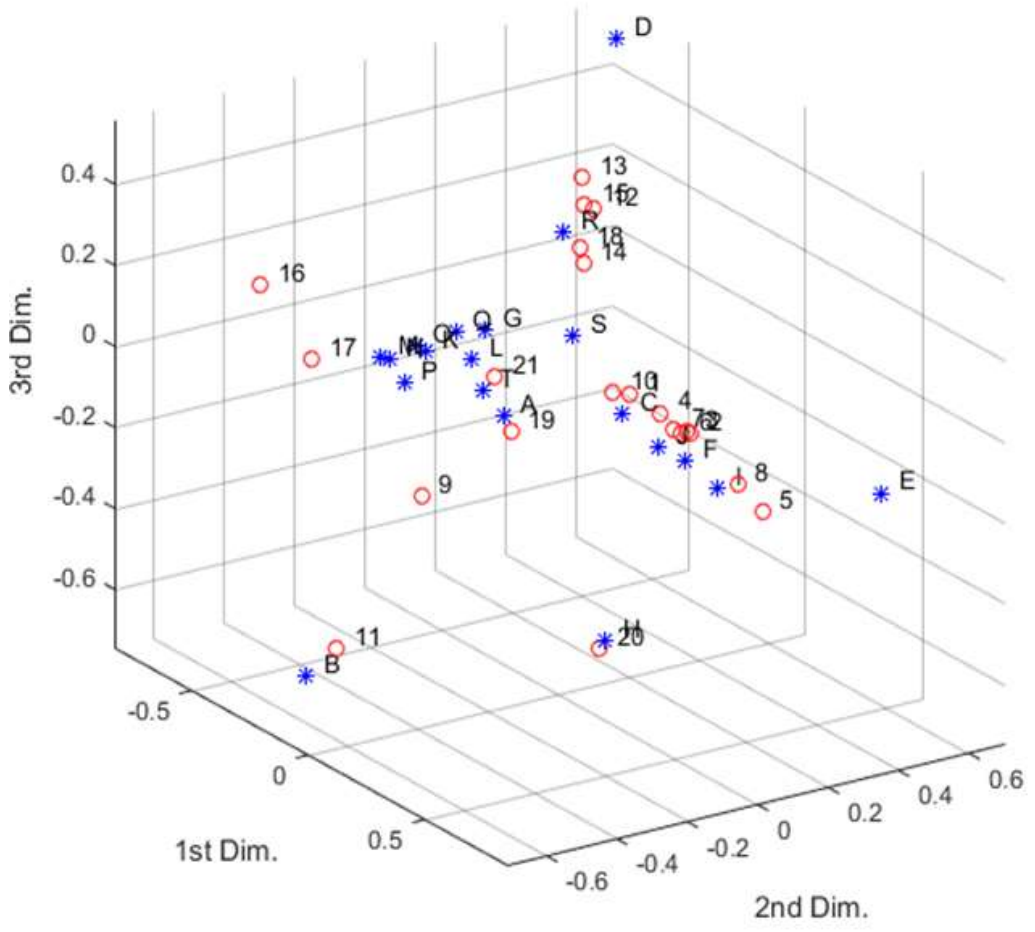
A(Dry paddy field), B(Paddy field), C(Orchard), D(Forestry), E(Building site), F(Factory site), G(School site), H(Parking lot), I(Gas station site), J(Warehouse site), K(Road), L(Railway site), M(Bank), N(River), O(Ditch), P(Marsh), Q(Water supply site), R(Park). S(Gymnasium site), T(Miscellaneous)

Land use dataset

1(Detached house), 2(Row house), 3(Multiplex house), 4(Apartment house), 5(Commercial building), 6(Business building), 7(Multipurpose building), 8(Industrial building), 9(Dry paddy field), 10(Orchard), 11(Paddy field), 12(Forestation field), 13(Natural forest field), 14(Grass field), 15(Bare soil field), 16(Road), 17(River), 18(Park), 19(Gymnasium), 20(Parking lot), 21(Miscellaneous)

3. Experiment

- Semantic space(with latent semantic analysis)



Land use survey

1(Detached house)
2(Row house)
3(Multiplex house)
4(Apartment house)
5(Commercial building)
6(Business building)
7(Multipurpose building)
8(Industrial building)
9(Dry paddy field)
10(Orchard)
11(Paddy field)
12(Forestation field)
13(Natural forest field)
14(Grass field)
15(Bare soil field)
16(Road)
17(River)
18(Park)
19(Gymnasium)
20(Parking lot)
21(Miscellaneous)

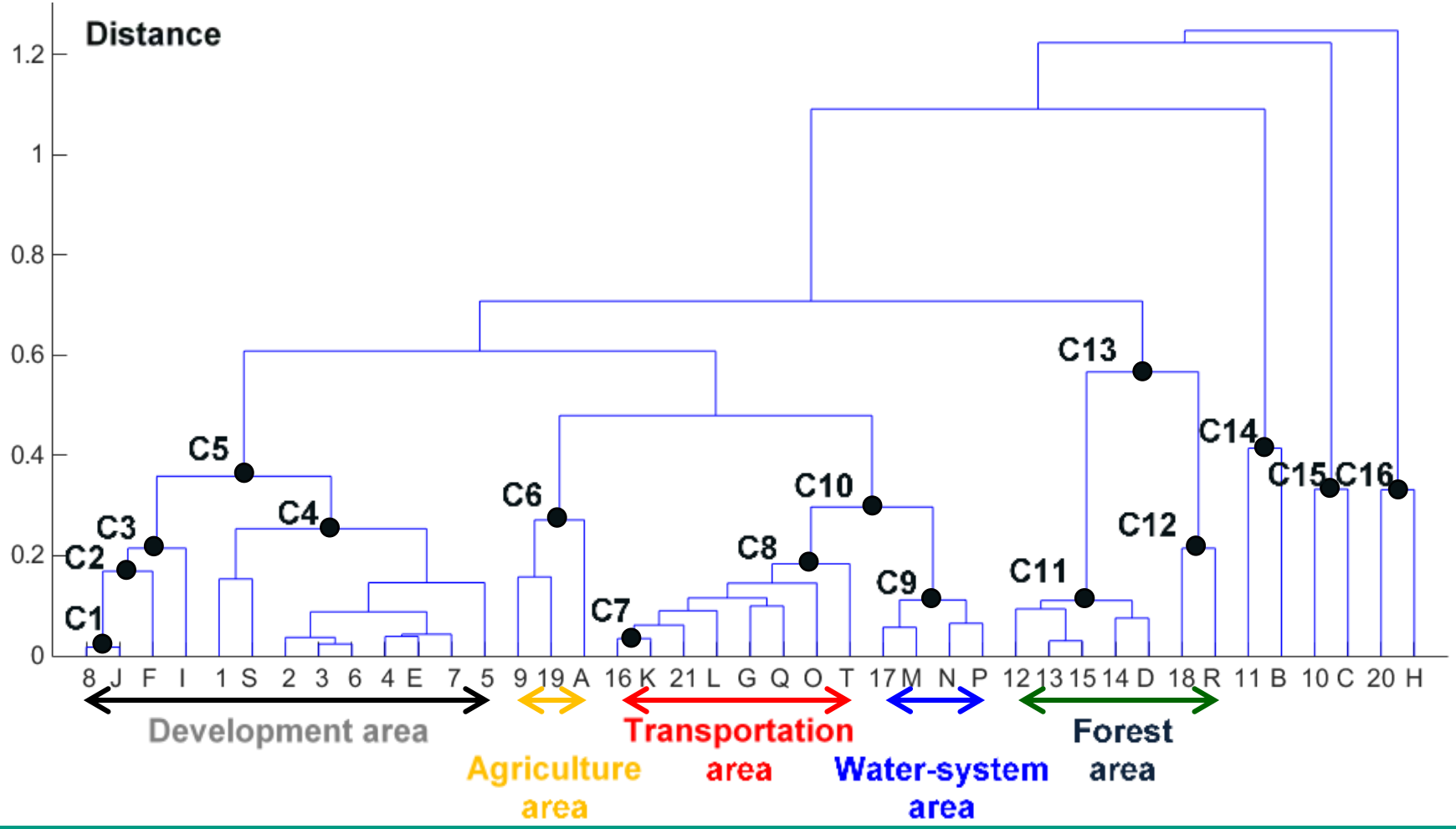
Land category in cadastral map

A(Dry paddy field)
B(Paddy field)
C(Orchard)
D(Forestry)
E(Building site)
F(Factory site)
G(School site)
H(Parking lot)
I(Gas station site)
J(Warehouse site)
K(Road)
L(Railway site)
M(Bank)
N(River)
O(Ditch)
P(Marsh)
Q(Water supply site)
R(Park)
S(Gymnasium site)
T(Miscellaneous)

3. Experiment

- Hierarchical clustering in the space

(with agglomerative hierarchical clustering)



3. Experiment

- **Dataset**(object intersection for about 800,000 parcels in Seoul)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	1,101	213	3	1,329	364,662	6	25	17	3	2	955	272	24	85	126	0	16	11	57	307
2	85	18	0	57	11,364	0	1	1	0	1	45	6	0	0	2	0	0	0	0	23
3	101	28	0	112	69,118	0	1	0	0	2	130	5	2	1	18	0	12	0	2	36
4	215	45	0	382	20,065	9	16	1	1	0	629	29	2	28	30	1	0	9	11	51
5	309	139	0	180	97,089	106	5	35	843	2	313	148	7	32	38	8	7	9	9	281
6	71	47	0	25	11,852	29	8	4	14	1	120	12	1	8	26	0	1	1	5	144
7	145	66	1	57	79,864	12	6	4	2	0	278	31	2	18	25	2	3	0	6	86
8	158	118	1	17	3,870	919	1	1	0	8	52	17	4	14	44	3	0	1	1	89
9	8,392	5,690	3	414	669	0	1	0	0	1	138	15	34	186	123	9	20	3	2	1,167
10	224	22	80	128	4	0	0	0	0	0	1	0	0	0	5	1	0	0	0	2
11	62	3,333	0	7	1	0	0	0	0	0	6	0	4	1	19	5	8	0	0	75
12	17	4	0	161	21	0	0	0	0	0	0	0	0	0	6	0	2	6	0	0
13	428	43	0	10,258	645	2	21	0	0	0	97	4	2	9	61	0	28	172	14	156
14	289	38	0	1,254	395	1	10	0	0	0	24	1	0	8	16	1	1	50	1	208
15	26	6	0	415	35	0	8	0	0	0	5	0	1	2	14	0	9	9	17	22
16	2,732	1,865	1	2,624	30,767	110	197	13	5	2	69,102	1,460	346	1,345	2,977	80	301	291	45	1,764
17	219	261	0	104	65	2	1	0	0	0	120	19	348	1,416	756	59	6	4	0	97
18	119	79	0	247	1,834	1	3	1	1	0	161	30	10	75	38	1	31	2,200	8	109
19	120	89	0	29	114	5	1	0	0	0	44	12	2	14	7	1	1	15	0	14
20	46	104	0	36	381	1	0	387	2	0	45	2	3	6	30	5	0	7	0	52
21	61	10	0	36	144	0	5	0	1	0	19	16	3	157	12	3	3	1	0	166

Cadastral dataset

A(Dry paddy field), B(Paddy field), C(Orchard), D(Forestry), E(Building site), F(Factory site), G(School site), H(Parking lot), I(Gas station site), J(Warehouse site), K(Road), L(Railway site), M(Bank), N(River), O(Ditch), P(Marsh), Q(Water supply site), R(Park). S(Gymnasium site), T(Miscellaneous)

Land use dataset

1(Detached house), 2(Row house), 3(Multiplex house), 4(Apartment house), 5(Commercial building), 6(Business building), 7(Multipurpose building), 8(Industrial building), 9(Dry paddy field), 10(Orchard), 11(Paddy field), 12(Forestation field), 13(Natural forest field), 14(Grass field), 15(Bare soil field), 16(Road), 17(River), 18(Park), 19(Gymnasium), 20(Parking lot), 21(Miscellaneous)

4. Conclusion

- new method for finding M:N correspondence

- by a search for the corresponding feature type–cluster pairs for two spatial datasets using the overlapping areas of the spatial object sets within the feature types

- the similarities of the feature types are measured and projected onto a lower dimensional vector space after applying latent semantic analysis

- since the feature types of high similarity are distributed close angle to each other in the projection space, clustering analysis is conducted to effectively find the feature types with high similarities.

Thank you