Evaluating Coverage and Accuracy of Census List of Households using New IT

Sun-Woong Kim[1], Eun-Seok Kim[2], Young-Je Woo[1], Ha-Na Lee[1], Bo-yoon Choi[1]

[1]Dongguk University Survey & Health Policy Research Center, [2]GDS Korea

dongguk UNIVERSITY SHPRC

OUTLINE

1. Research Background
2. Study Design & Methodology
3. Study Results
4. A New Map Service “Web Blocker”
5. Concluding Remarks

1 Research Background

1. Research Background

- In Korea, the Census list of households is commonly used as sampling frames for small areas (e.g., enumeration districts (EDs)), randomly selected in national household surveys.

- It enables many surveys to be done cheaper and faster than is possible with on-site enumeration for building area sampling frames.

1. Research Background (Cont.)

- But since the Census is conducted every five years, the list of households is often out of date in many regions, especially due to the construction for high-rise residential and apartment buildings.

- However, the coverage or accuracy of the list has never been estimated systematically in a national or provincial level.

1. Research Background (Cont.)

- We evaluate the coverage and accuracy of the 2010 Census list of households based on highly advanced **information technology** (**IT**) including ***Road Name Address Information System*** (new address Internet information service) and ***online map service*** (daum, naver, etc.) in ***2013*** as well as a sophisticated sample design.

- Also, we illustrate an alternative, which uses a new map service called “Web Blocker” and may be more useful to measure undercount or overcount of the Census list of households or housing units in the near future.

2 Study Design & Methodology

2. Study Design & Methodology

● Sample Design

- Stratified two-stage unequal-probability sampling method is used to estimate the coverage or accuracy of the Census list of households.

1) The first stage of sampling involves dividing Korea into **226** primary sampling units (PSUs). The PSUs are then grouped into **48** strata on the basis of survey information. Two or three PSUs are sampled in each stratum with πPS sampling.

2) In the second stage, a sample of EDs is drawn with probability proportional to size sampling with replacement. The number of the selected EDs is **4,644**. These EDs have about **60** households on average.

2. Study Design & Methodology (Cont.)

● Sample Design (Cont.)

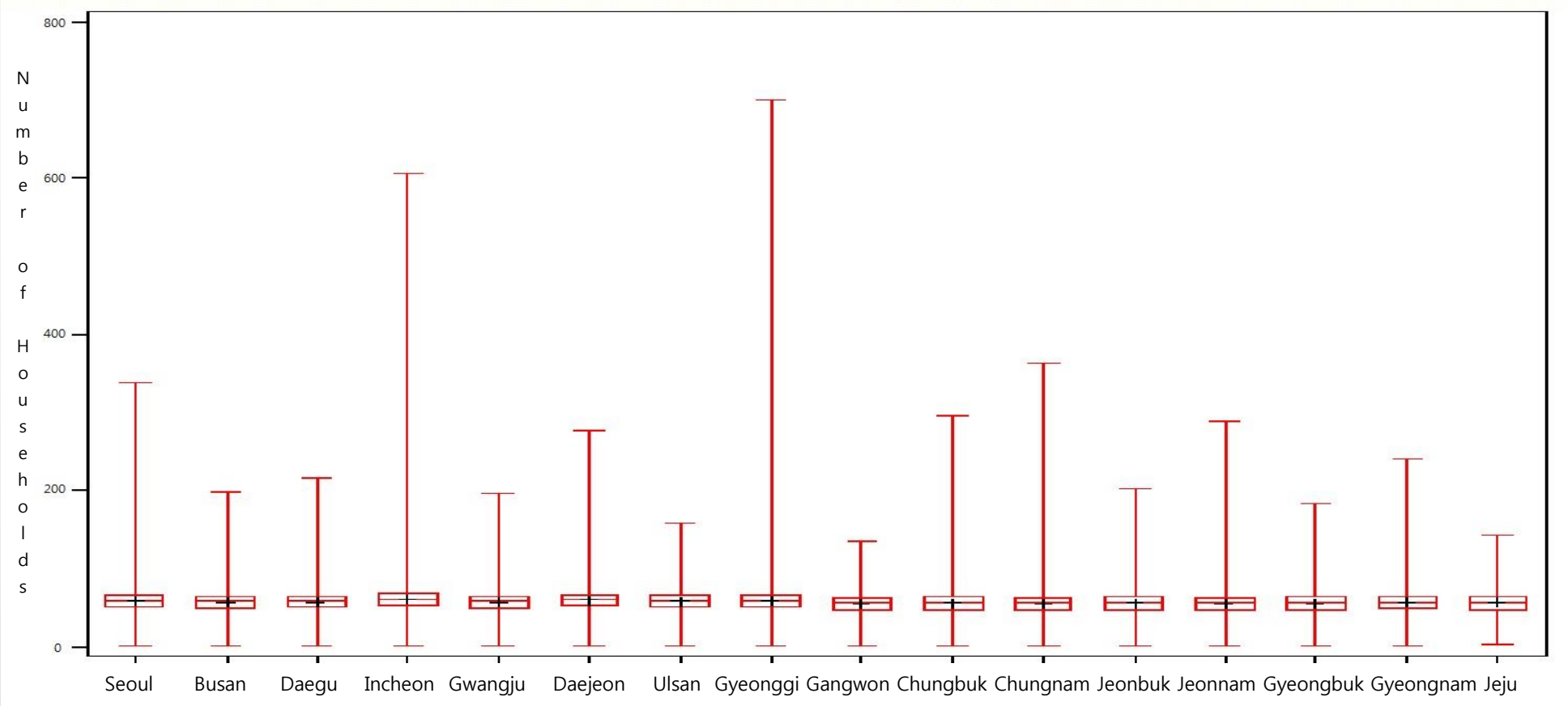
- Selection Procedures for Two-Stage Area Sampling

	First stage	Second stage
Sampling Unit	City or County	Enumeration District (ED)
Sampling Method	πPS sampling(Sampford, 1967) for selecting 2 or 3 cities from each stratum	PPS without replacement sampling for selecting several EDs from each selected city

2. Study Design & Methodology (Cont.)

● Sample Design (Cont.)

- Size Distribution of EDs by Province



2. Study Design & Methodology (Cont.)

- **Using New IT**
  - **The detailed maps with new address information** on the Internet services showing the location of individual households open to the public **are matched to the Census list of households** in each selected ED.
  - Additionally, a variety of online map services are used.

2. Study Design & Methodology (Cont.)

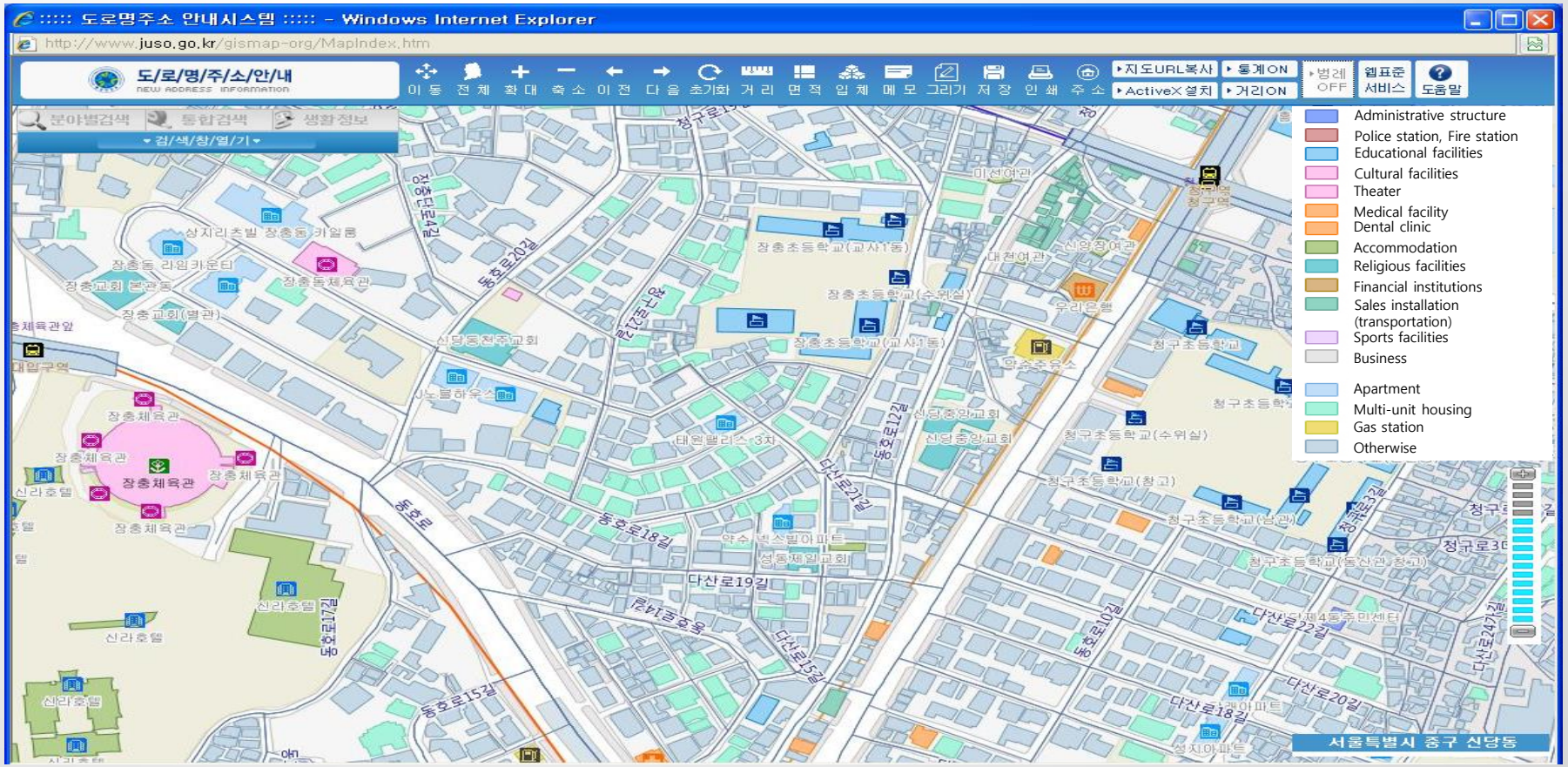
- **Advantages of Using New IT**
  - New IT provides useful information such as

    1) Address of every structure or building
    2) Type of structure
    3) Number of dwellings
    4) Outside of a structure and surroundings

2. Study Design & Methodology (Cont.)

● Advantages of Using New IT (Cont.)

- Road Name Address Information System (Ministry of Security and Public Administration)

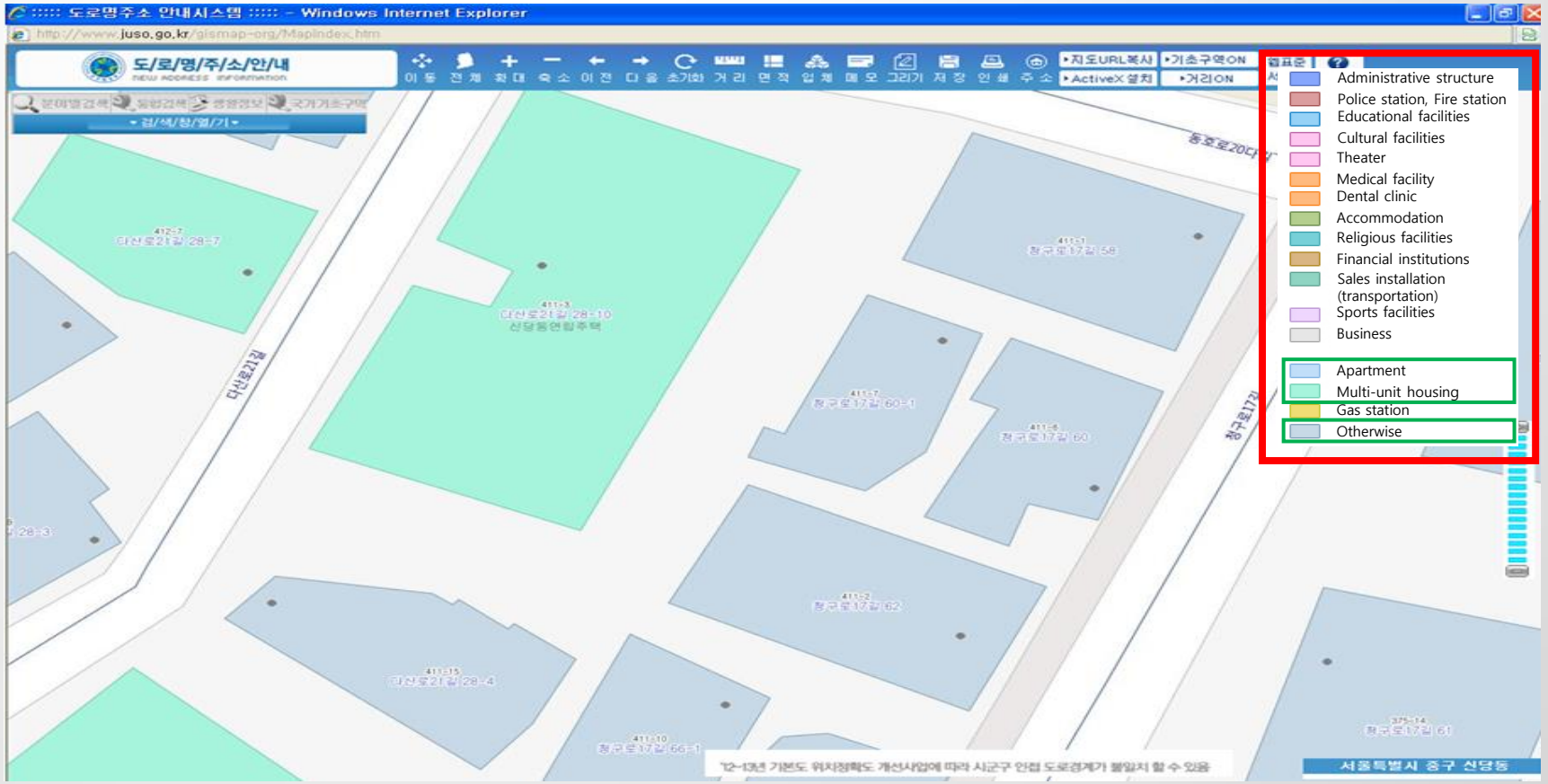


dongguk UNIVERSITY SHPRC

2. Study Design & Methodology (Cont.)

● Advantages of Using New IT (Cont.)

- Road Name Address Information System (Cont.)



2. Study Design & Methodology (Cont.)

- **Advantages of Using New IT** (Cont.)
  - Matching to other online map service

2. Study Design & Methodology (Cont.)

- **Estimation of Proportion of EDs with Non-coverage**
  (Erroneous Enumerations, Omissions, New Constructions, Demolitions)
  - Notation
    - G : The number of all EDs in population
    - H: The number of strata, $h = 1, \dots, H$
    - N_h: The number of all PSUs (Cities or Counties) in stratum h
    - M_{hi}: The number of all EDs in i^{th} PSU of stratum h
    - n_h: The number of selected PSUs in stratum h
    - π_{hi}: The first-order inclusion probability for i^{th} PSU in stratum h of sample
    - π_{hik}: The second-order inclusion probability for i^{th} and k^{th} PSU in stratum h of sample
    - m_{hi}: The number of sampled EDs in i^{th} PSU of stratum h
    - p_{hij}: Relative size for j^{th} ED in i^{th} PSU of stratum h
    - y_{hij}: Value of the study variable for j^{th} ED in i^{th} PSU of stratum h
    - Y_{hi}: Value of the sum of all y_{hij}s in i^{th} PSU of stratum h
    - $\hat{Y}_{hi}$: Estimator of the sum of all y_{hij}s in i^{th} PSU of stratum h

2. Study Design & Methodology (Cont.)

- **Estimation of Proportion of EDs with Non-coverage (Cont.)**

$$\hat{P}_{st} = \frac{1}{G}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\frac{1}{\pi_{hi}}\frac{1}{m_{hi}p_{hij}}y_{hij}$$

$$V(\hat{P}_{st}) = \frac{1}{G^2}\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{k>i}^{N_h}(\pi_{hi}\pi_{hk}-\pi_{hik})\left(\frac{Y_{hi}}{\pi_{hi}}-\frac{Y_{hk}}{\pi_{hk}}\right)^2 + \frac{1}{G^2}\sum_{h=1}^{H}\sum_{i=1}^{N_h}\frac{1}{\pi_{hi}}\left(\sum_{j=1}^{M_{hi}}\frac{{y_{hij}}^2}{p_{hij}}-{Y_{hi}}^2\right)\Big/ m_{hi}$$

$$\hat{V}(\hat{P}_{st}) = \frac{1}{G^2}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k>i}^{n_h}\left(\frac{\pi_{hi}\pi_{hk}}{\pi_{hik}}-1\right)\left(\frac{\hat{Y}_{hi}}{\pi_{hi}}-\frac{\hat{Y}_{hk}}{\pi_{hk}}\right)^2 + \frac{1}{G^2}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{1}{\pi_{hi}}\sum_{j=1}^{m_{hi}}\left(\frac{y_{hij}}{p_{hij}}-\hat{Y}_{hi}\right)^2\Big/ m_{hi}(m_{hi}-1)$$

2. Study Design & Methodology (Cont.)

- **Joint selection probabilities π_{hik} for all pairs of selected PSUs in each stratum**

  - $\pi_{hik} = Q_h \lambda_{hi} \lambda_{hk} \sum_{t=2}^{n_h} \left[\{ t - n_h (z_{hi} + z_{hk}) \} L_{h,(n_h - t)} (\overline{ik}) \right] / {n_h}^{t-2}$

    where z_{hi} is the relative size for i^{th} PSU, $\lambda_{hi} = z_{hi} / (1 - n_h z_{hi})$,

    $$Q_h = 1 / \sum_{t=1}^{n_h} \left(t L_{h,(n_h - t)} / {n_h}^t \right),$$

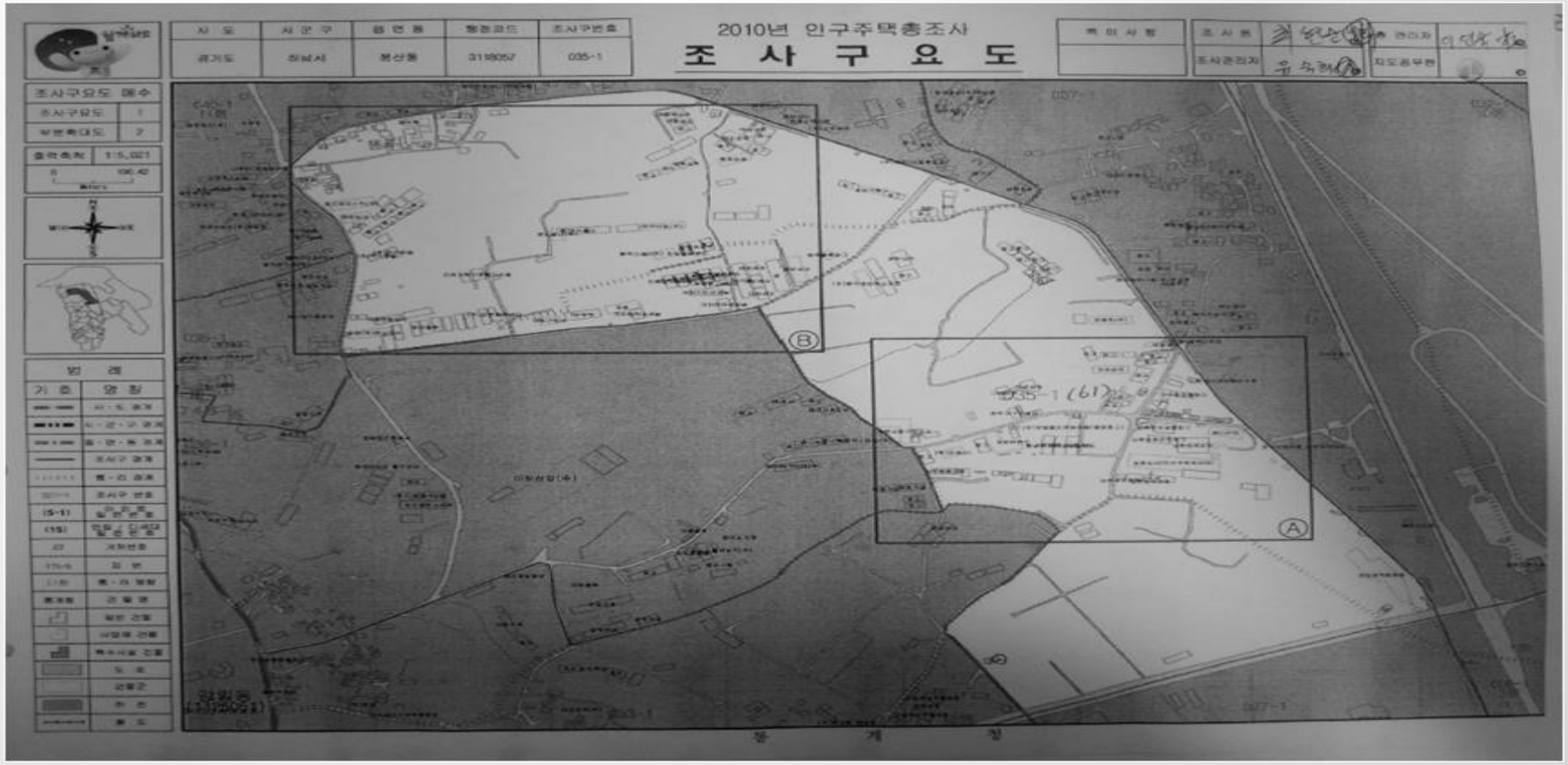
    $$L_{h,v} = \sum_{S_h(v)} \lambda_{hi_1} \lambda_{hi_2} \cdots \lambda_{hi_v}.$$

$S_h(v)$denotes all possible samples of size v, for $v = 1,2, \ldots, N_h$.
The sum $L_{h,v}\ (\overline{ik})$ is defined similarly to $L_{h,v}$ but sums over all possible samples of size v that do not include units i and k.

3 Study Results

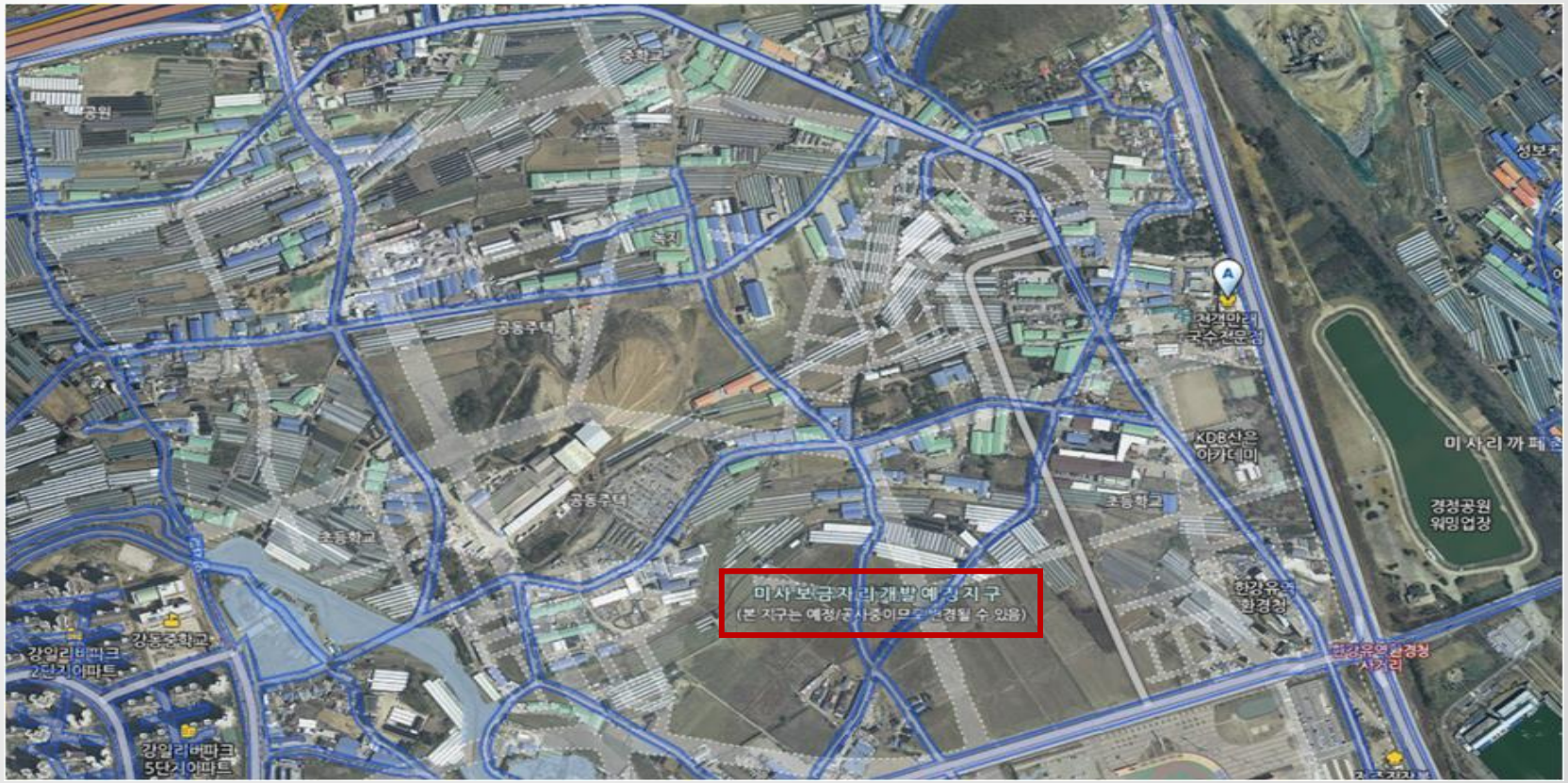
3. Study Results

- **Examples of EDs that need to be replaced**
  - Paper Map of 2010 Census EDs (A, B) (ED 035-1, Pungsan-dong, Hanam-si, Gyeonggi-do)



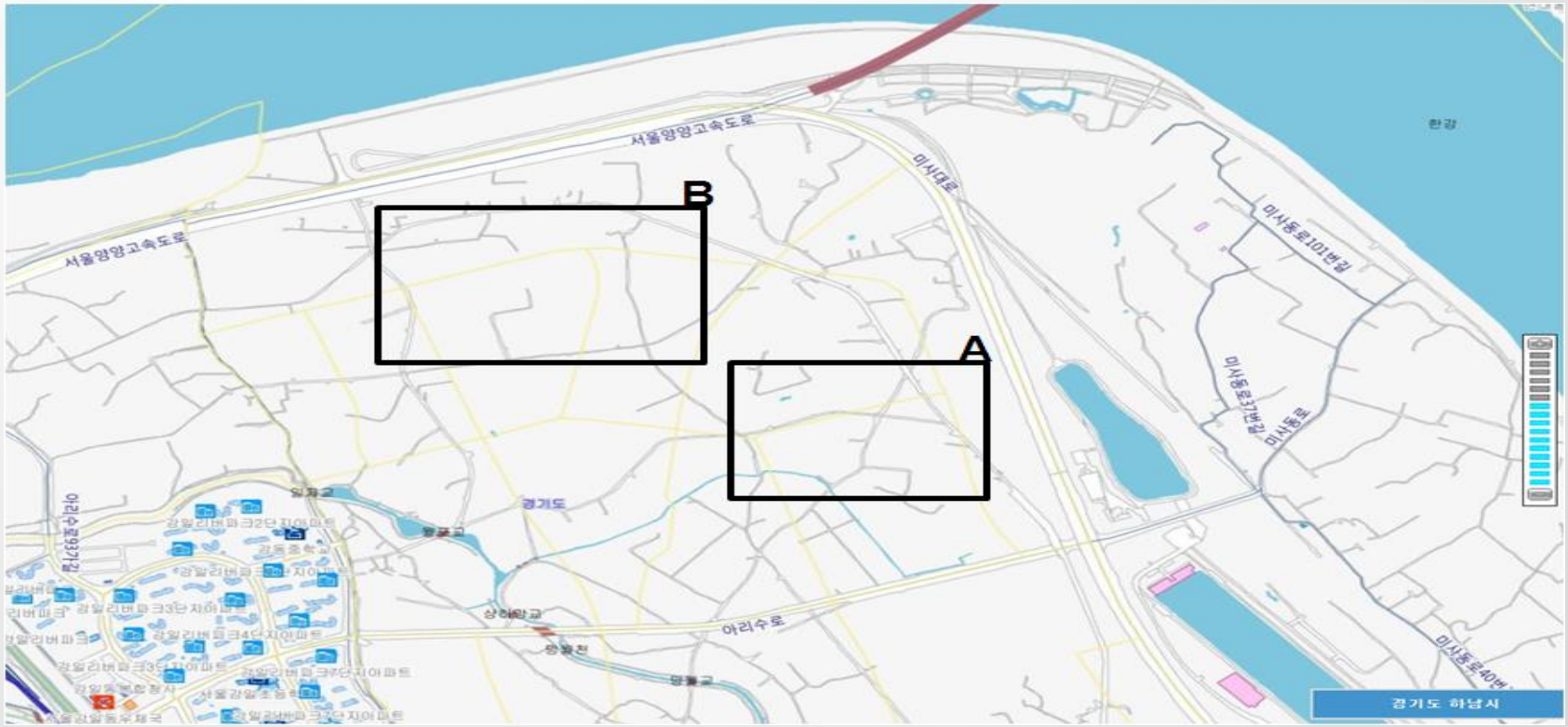
3. Study Results (Cont.)

- **Examples of EDs that need to be replaced** (Cont.)
  - Satellite image in 2013 (ED 035-1)
    : Development-prearranged area (Misa-bogeumzari)



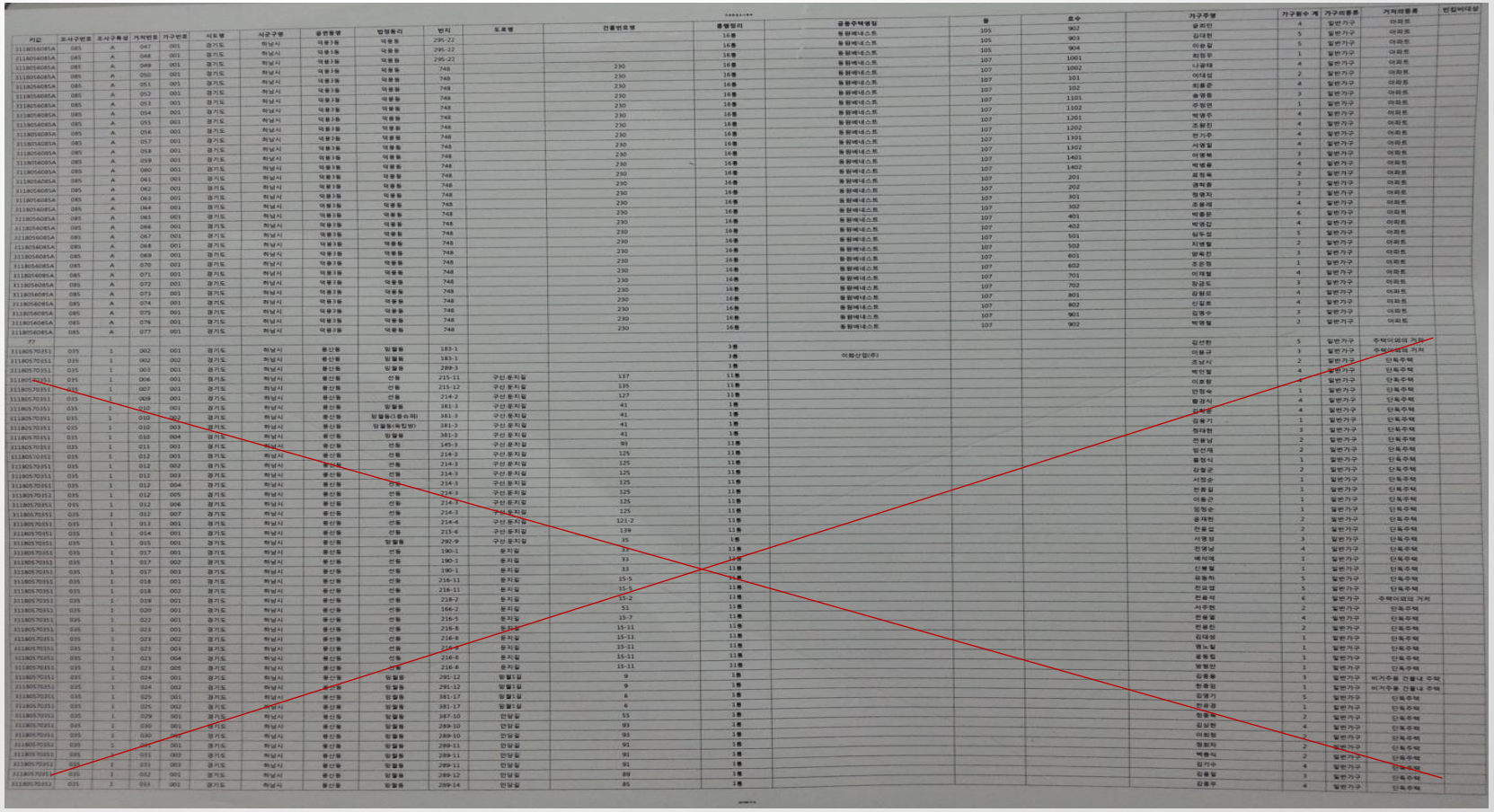
3. Study Results (Cont.)

- **Examples of EDs that need to be replaced** (Cont.)
  - 2013 Road Name Address Information System (ED 035-1)
    - No dwellings in EDs (A, B)
    - Need to be replaced with new sample EDs



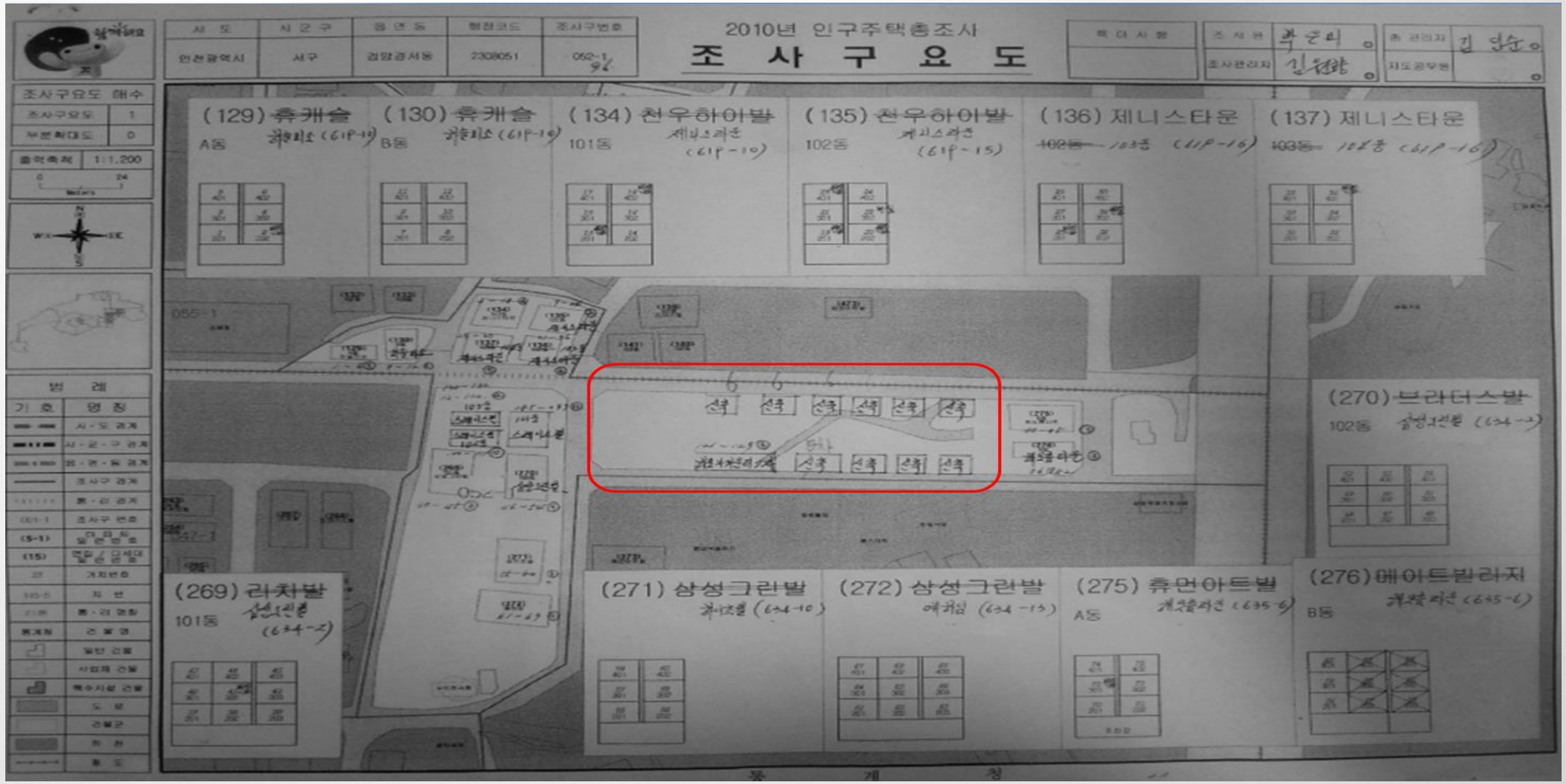
3. Study Results (Cont.)

- **Examples of EDs that need to be replaced** (Cont.)
  - Census list of households in EDs (ED 035-1)



3. Study Results (Cont.)

- **Examples of new constructions in ED**
  - Paper Map of 2010 Census ED (ED 052-1, Geomam Gyeongseo-dong, Seo-gu, Incheon)



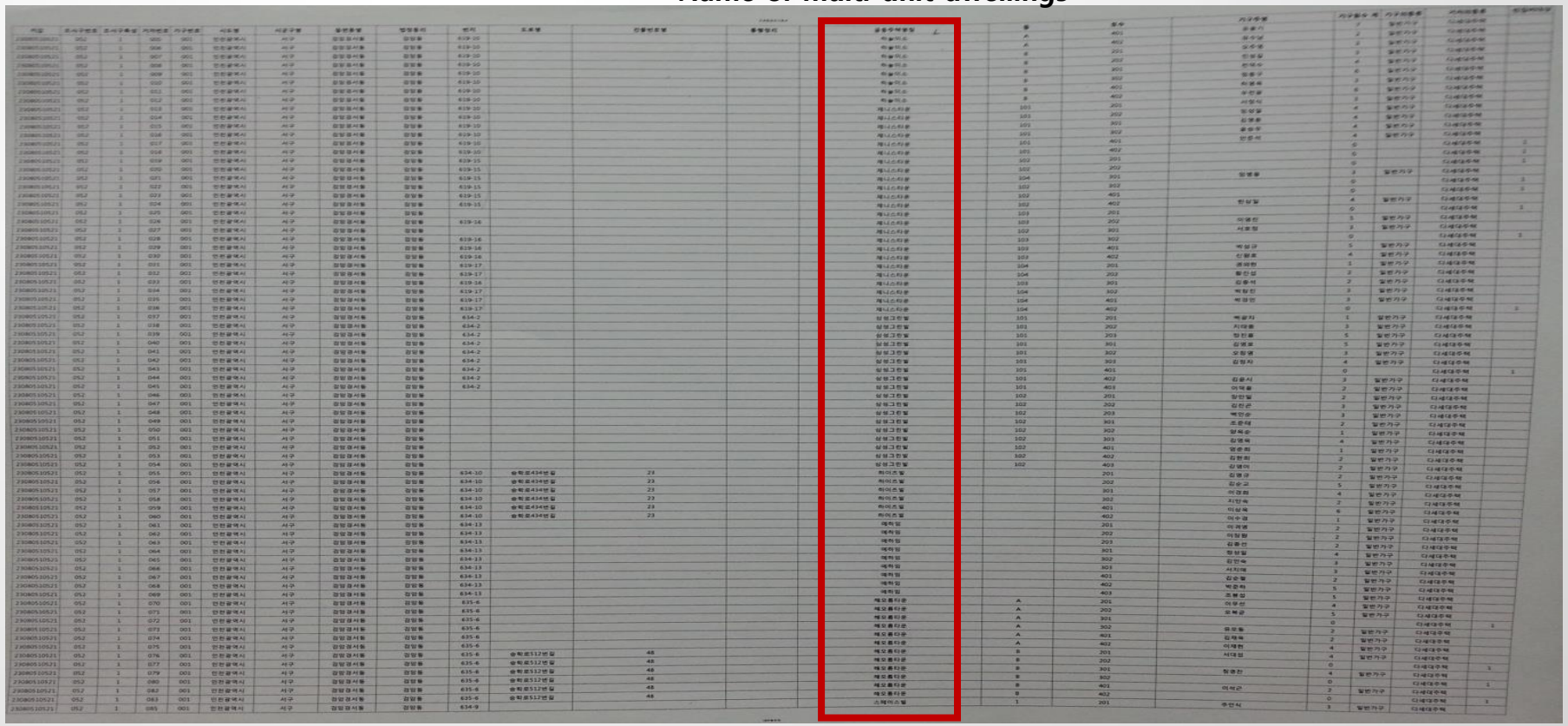
3. Study Results (Cont.)

- **Examples of new constructions in ED** (Cont.)
  - 2013 Road Name Address Information System (ED 052-1)
    - Some new constructions in ED
    - Need to update the list of households

3. Study Results (Cont.)

- **Examples of new constructions in ED** (Cont.)
  - Census list of households in ED (ED 052-1)

Name of multi-unit dwellings



3. Study Results (Cont.)

● Examples of new constructions in ED (Cont.)

- List of added buildings (ED 052-1)

Name of multi-unit dwellings

행정구번호	조사구번호	시도명	시군구명	읍면동명	법정동리	번지	도로명	건물번호명	공동주택 명칭	거처의종류
2308051052 1	52	인천광역시	서구	검암경서동	검암동	635-8	승학로450 번길	24	하모니 카운티8차	연립
2308051052 1	52	인천광역시	서구	검암경서동	검암동	635-9	승학로450 번길	26	하모니 카운티8차	연립
2308051052 1	52	인천광역시	서구	검암경서동	검암동	635-10	승학로450 번길	28	하모니 카운티8차	연립
2308051052 1	52	인천광역시	서구	검암경서동	검암동	635-5	승학로512 번길	58	하모니 카운티8차	연립
2308051052 1	52	인천광역시	서구	검암경서동	검암동	635-12	승학로450 번길	30	휴케슬 (가동)	연립
2308051052 1	52	인천광역시	서구	검암경서동	검암동	635-12	승학로450 번길	30	휴케슬 (나동)	연립
2308051052 1	52	인천광역시	서구	검암경서동	검암동	635-13	승학로451 번길 20-3	31	아이빌리지	연립
2308051052 1	52	인천광역시	서구	검암경서동	검암동	635-13	승학로451 번길 20-4	31	아이빌리지	연립
2308051052 1	52	인천광역시	서구	검암경서동	검암동	635-7	승학로450 번길	32	스페이스 빌	연립
2308051052 1	52	인천광역시	서구	검암경서동	검암동	635-11	승학로450 번길	34	정성드림 빌3차	연립
2308051052 1	52	인천광역시	서구	검암경서동	검암동	634-11	승학로434 번길	21-1	잔쉐르빌	연립

3. Study Results (Cont.)

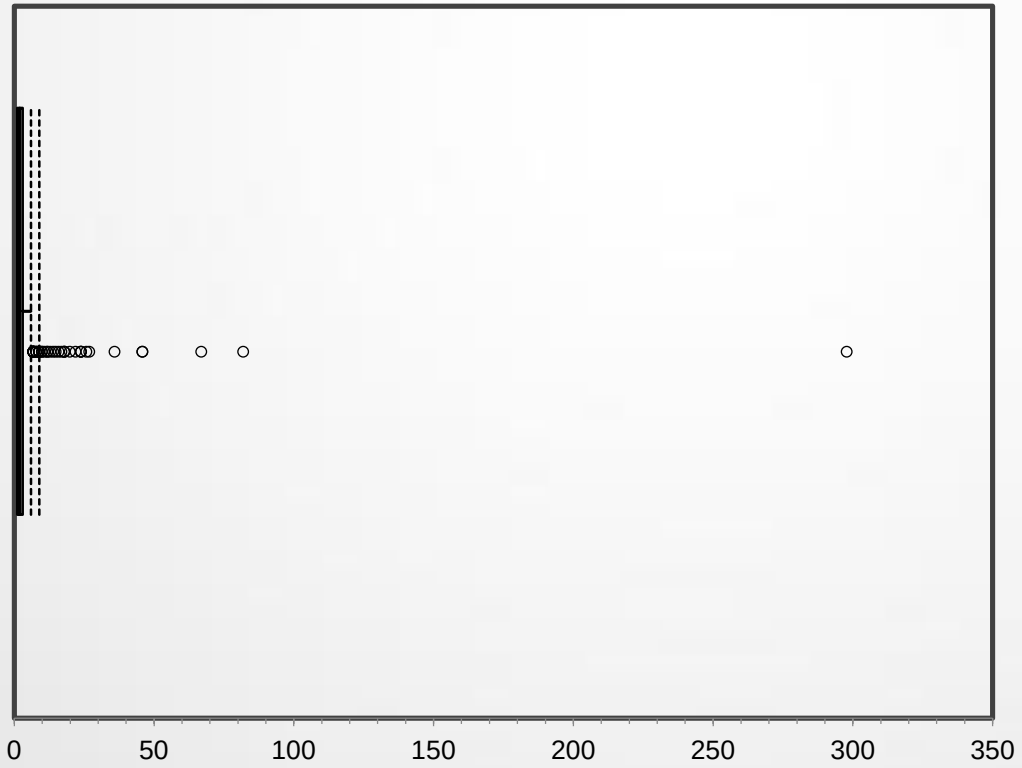
● Distribution of EDs with Undercount or Overcount of HHs

- Erroneous Enumerations, Omissions, New Constructions, and Demolitions

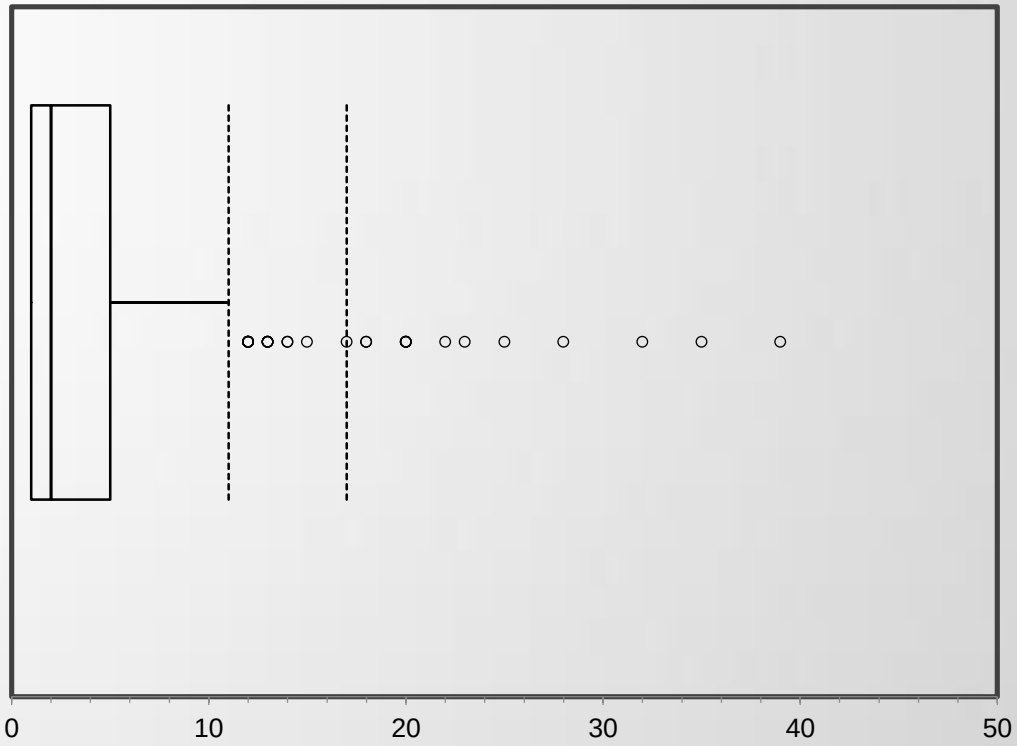
Erroneous Enumerations or Omissions or New Constructions of HHs	No. of EDs
1	225
2	111
3	52
4	26
5	13
6	10
7	10
8-10	11
11-20	12
21-298	12
Total	**482**

Demolitions of HHs	No. of EDs
1	191
2	125
3	79
4	53
5	45
6	29
7	20
8-10	28
11-20	26
21-39	7
Total	**603**

3. Study Results (Cont.)



Erroneous Enumerations or Omissions
or New Constructions of HHs



Demolitions of HHs

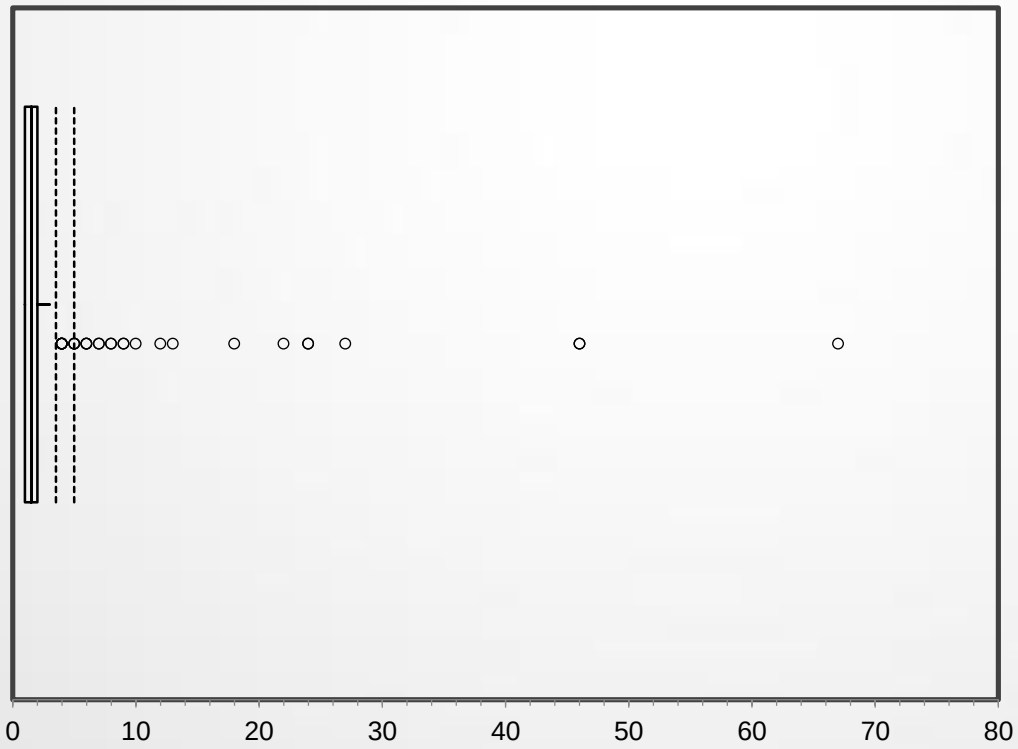
3. Study Results (Cont.)

- **Distribution of EDs with Undercount or Overcount of HHs** (Cont.)
  - Erroneous Enumerations **Only**, Omissions **Only**, New Constructions **Only**, and Demolitions **Only**

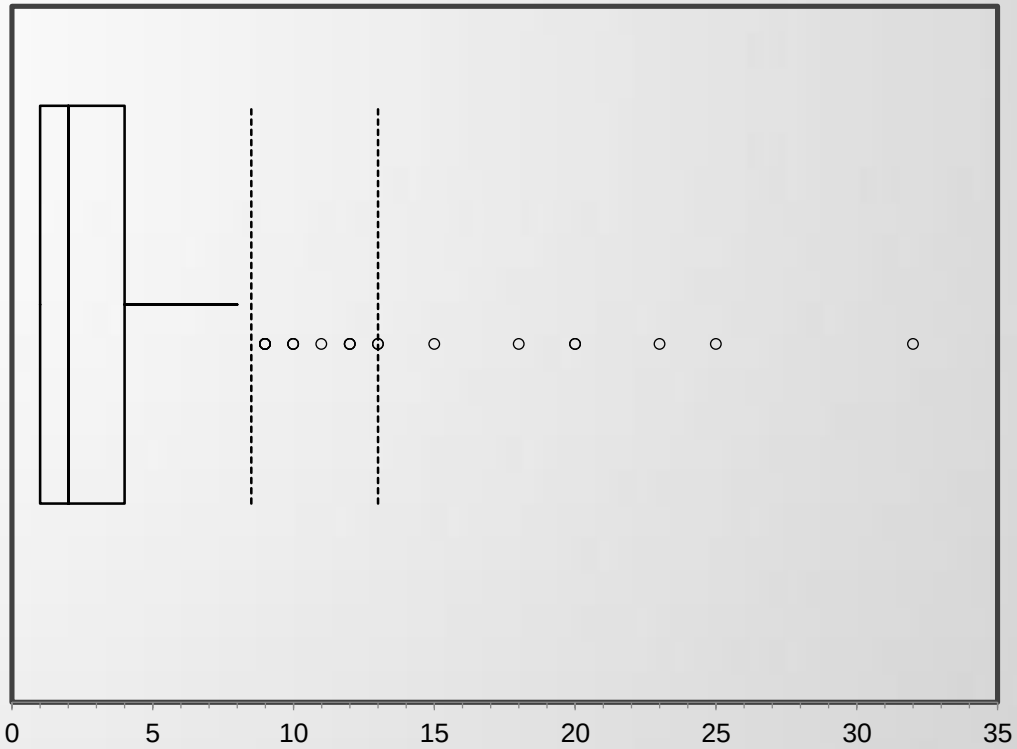
Erroneous Enumerations Only or Omissions Only or New Constructions Only of HHs	No. of EDs
1	97
2	52
3	16
4-6	12
7-20	10
21-67	7
Total	**194**

Demolitions Only of HHs	No. of EDs
1	121
2	75
3	32
4	19
5	18
6	15
7	10
8-10	12
11-20	10
23-32	3
Total	**315**

3. Study Results (Cont.)



Erroneous Enumerations Only or Omissions Only or New Constructions Only of HHs



Demolitions Only of HHs

3. Study Results (Cont.)

● National Level Estimates of Proportion of EDs

	Estimated Percent	Standard Error
1) Erroneous Enumerations or Omissions or New Constructions or Demolitions of HHs	17.75 %	3.51%
2) Erroneous Enumerations or Omissions or New Constructions of HHs	10.22 %	2.53%
3) Demolitions of HHs	13.41 %	2.57%
4) Erroneous Enumerations or Omissions or New Constructions and Demolitions of HHs	5.87 %	1.60%
5) Erroneous Enumerations Only or Omissions Only or New Constructions Only of HHs	4.35 %	1.33%
6) Demolitions Only of HHs	7.54 %	1.41%

3. Study Results (Cont.)

● Estimates by Province

Province	1)	2)	3)	4)	5)	6)
Gangwon	40.64 %	24.79 %	36.34 %	20.49 %	4.30 %	15.85 %
Gyeonggi	10.79 %	6.71 %	6.93 %	2.85 %	3.86 %	4.07 %
Gyeongnam	26.15 %	16.20 %	21.50 %	11.56 %	4.65 %	9.95 %
Gyeongbuk	36.20 %	20.02 %	29.93 %	13.75 %	6.27 %	16.18 %
Busan	12.18 %	5.81 %	8.02 %	1.65 %	4.15 %	6.37 %
Seoul	9.09 %	4.73 %	6.69 %	2.33 %	2.40 %	4.36 %
Jeonnam	37.90 %	23.74 %	28.14 %	13.98 %	9.76 %	14.17 %
Jeonbuk	30.76 %	21.64 %	26.00 %	16.88 %	4.76 %	9.12 %
Chungnam	31.99 %	9.47 %	28.48 %	5.95 %	3.51 %	22.53 %
Chungbuk	28.23 %	16.60 %	21.81 %	10.18 %	6.42 %	11.64 %
Gwangju	9.54 %	7.16 %	2.38 %	0.00 %	7.16 %	2.38 %
Daegu	12.49 %	10.03 %	7.69 %	5.23 %	4.81 %	2.46 %
Daejeon	6.94 %	5.35 %	3.00 %	1.41 %	3.94 %	1.59 %
Ulsan	29.92 %	14.51 %	19.47 %	4.07 %	10.45 %	15.40 %
Incheon	6.08 %	3.75 %	4.66 %	2.33 %	1.42 %	2.33 %
Jeju	29.67 %	14.10 %	21.31 %	5.74 %	8.36 %	15.57 %

3. Study Results (Cont.)

● Errors in Lists of HHs or Paper Maps

Type of Errors	Content
List of HHs	Error of land-lot number
	Error of a road name address
	Error of a residence number in the list
	Error of a name of apartments
Paper Map of EDs	Reversed Locations of apartments
	Shape or location of buildings different from satellite images
	Unable to identify residence number
	No entire map of EDs
Others	List of HHs and the corresponding map of EDs do not match.
	Whether a new building is dwelling or not is not identified.
	The number of HHs in EDs are changed because of new constructions.
	There is no HH in a list but in the map of EDs.

3. Study Results (Cont.)

● Errors in Lists of HHs or Paper Maps (Cont.)

- Estimates of proportion of EDs with errors by Province

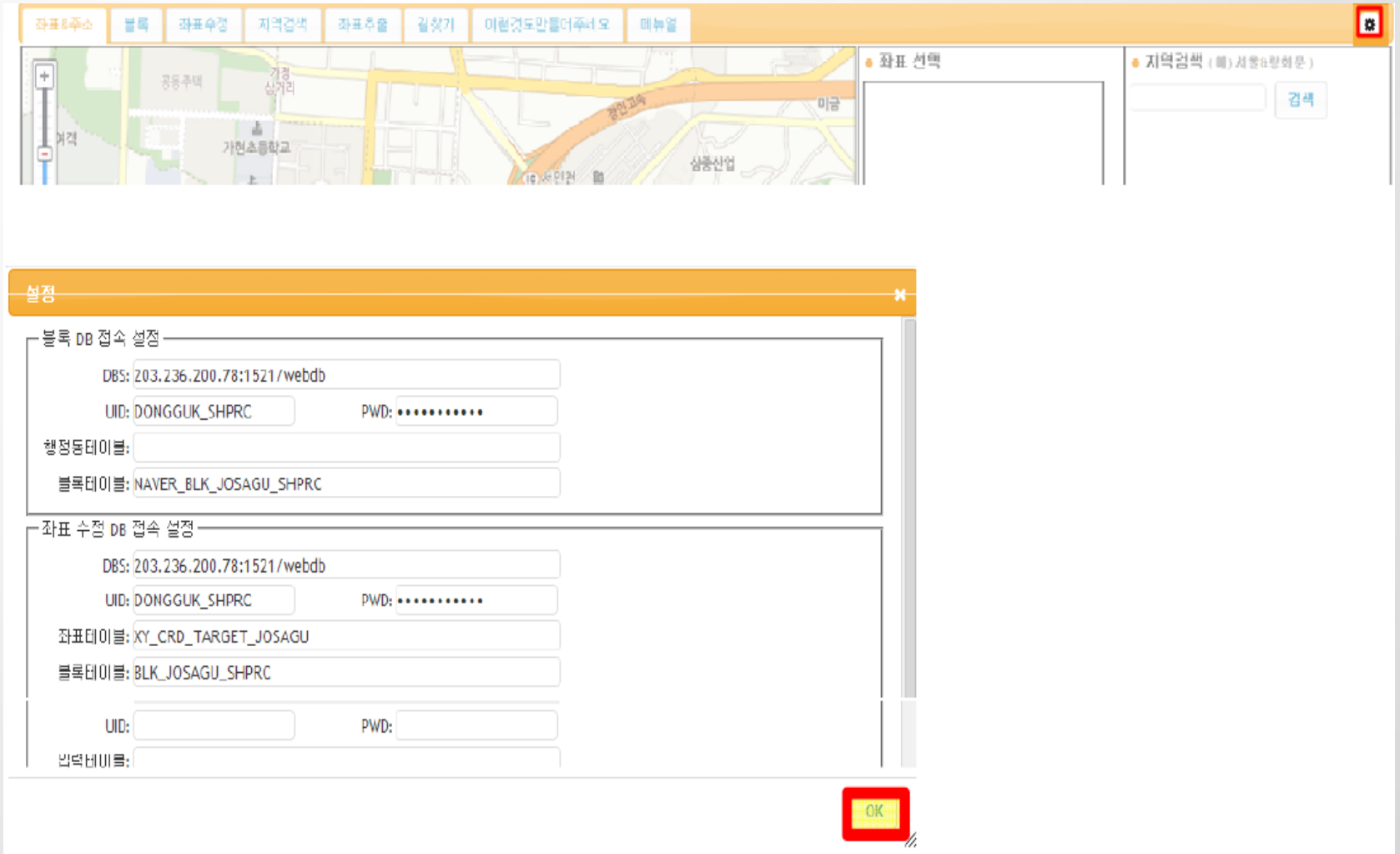
Province	Percent
Gangwon	17.1%
Gyeonggi	10.4%
Seoul	2.6%
Gyeongnam	16.4%
Gyeongbuk	12.1%
Daegu	0.5%
Daejeon	2.4%
Incheon	9.0%
Chungnam	20.7%
Chungnuk	9.6%
Busan	**30.9%**
Gwangju	0.0%
Ulsan	5.6%
Jeonnam	23.4%
Jeonbuk	26.3%
Jeju	13.3%
Total	**12.5%**

4 A New Map Service “Web Blocker”

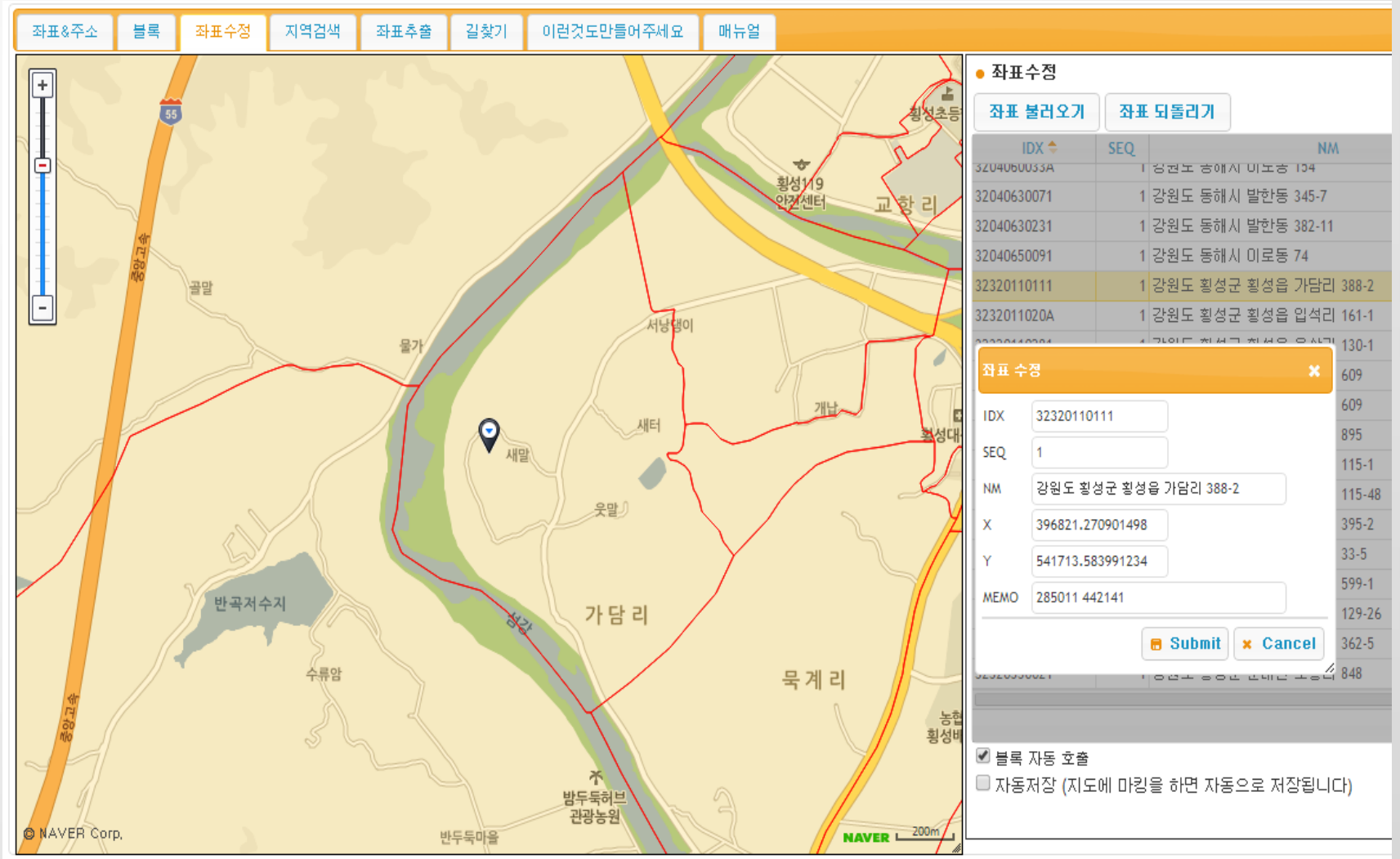
4. A New Map Service “Web Blocker”

- Map service provided by a commercial company
- Residential mailing lists available
- More useful in checking the coverage

4. A New Map Service “Web Blocker” (Cont.)



4. A New Map Service “Web Blocker” (Cont.)



5 Concluding Remarks

5. Concluding Remarks

- New IT offers significant reductions in time and cost of evaluating the Census coverage as compared with a post-enumeration survey based on on-site enumeration.
- Census lists or maps have a large non-coverage that include erroneous enumerations, omissions, new constructions, or demolitions.
- Census lists or maps also have a large number of errors.

5. Concluding Remarks (Cont.)

- Since Census lists or maps will miss many dwellings or housing units, specific remedies or treatments for the non-coverage or inaccuracy should be given or offered.
- We need to develop field procedures which can result in drastic improvements in housing unit coverage.
- In the near future we may use a map service called “Web Blocker” to be more accurate in checking the coverage or to be used as a sampling frame.

THANK YOU

Contact at: sunwk@donggk.edu