SCAG Parcel Database Validation Report on Accuracy of Total Floor Space per Parcel by Ground truth Trekking

Michael F. Goodchild, Wenwen Li, Alex Schild and Nate Royal

Department of Geography; Center for Spatial Studies
University of California, Santa Barbara 93106-4060
{good, wenwen}@geog,ucsb.edu

Summary

This work aims at evaluating the reliability of the floor space per parcel recorded in the SCAG parcel database. Meanwhile, we tried to provide guidance (and a factor) to adjust the floor space data field with missing values in the parcel database. Floor space is an important indicator for urban economics, for achieving the equilibrium among floor space demand, transportation demand and other factors. One of the tasks for LA-project is to compute the total floor space and FAR (Floor Area to Land Ratio) per model zone in the Great Los Angeles region. However, we found that a substantial amount of parcels are listed as having no buildings (data field "IMPSQRT"='0') or with the values missing. The imperfect raw data in the SCAG database will influence the accuracy of the simulation of RELU-TRAN model. Therefore, ground truth trekking on sample data was conducted for parcel database within Riverside County, Imperial County and Orange County, of which there tends to be more "errors".

1 Procedure

1.1 Data preparation

All the parcels in the study area (Riverside County, Imperial County and Orange County) with missing floor values are selected. Parcels that have a zero in the improvement field or a blank in the improvement field (Field Name: IMPSQRT) are both considered as such parcels. The following query

SELECT * FROM <u>riverside_county</u> WHERE "IMPSQFT"=0 on the SCAG 2007 database (ftp://mrpi.geog.ucsb.edu/data2010/parcel/GIS/) were used.

1.2 Random selection

The selection procedure is to generate random numbers for ALL parcels with missing floor space value in the three counties (Riverside, Imperial and Orange) being studied. Here “ALL” means parcels that are identified as residential (land use code LU08: 11xx), commercial (land use code LU08: 12xx) and industrial parcels (land use code LU08: 13xx). Then the parcels are sorted in the ascending order per model zone and the top 50 parcel in each model zone are selected for ground truth trekking. This algorithm is capable of avoiding duplicate selection of parcels and the python code is listed in Appendix I.

As the parcel records are organized by counties, the parcel-model zone relations need to be identified before random numbering. The dataset indicating parcel and model zone containment can be found at (ftp://mrpi.geog.ucsb.edu/data2010/parcel/MZparcel_imputed_GIS/). Meanwhile, the GIS data and attribute data of the SCAG parcel 2007 dataset need to be joined spatially by field “SCAGXYID” in order to link the floor space field “IMPSQFT” (short for improvement measured in square feet) to the GIS data.

After the sample parcels are selected, the boundaries of these parcels in the GIS file are converted to Google kmz file, which can be overlaid seamlessly with imageries in Google Earth for ground truth trekking. The trekking process is as follows:

a. Identify the centroid coordinates of each parcel being sampled;

b. Type in the latitude and longitude to locate the parcel position by the coordinates obtained in step “a”;

c. Through the satellite image of the parcel in Google Earth, the footprint of any buildings within that parcel can be measured using the polygon tool provided by Google Earth Pro. Version.

d. After identifying the building footprint, the number of floors needs to be calculated. This needs the support of Google Street View tool. Within the environment of Google Earth, double click the nearest street of a building, the street view will appear. Then we can count the number of floors for that building;

e. The total floor space equals to the product of number of buildings, footprint area of a building and number of floors.

Note that some parcels have no street view in Google Earth, so it is difficult to measure the actual floor space. In order to overcome these issues, field work in the study areas was done to obtain the floor space physically. Another issue for using this method is systematic overestimate of floor area since it includes garage area (and probably patio area) which would not be counted as floor area in the parcel database. To resolve this potential problem, we sampled from the parcel database and selected parcels with positive floor area, estimated the floor space for the same parcels in Google Earth, and compared the results.

2 Results

2.1 Results for Riverside County

There are two set of parcels selected in Orange County: 750 parcels with floor space missing and 200 parcels with positive floor space values. The number of 750 is obtained by multiplying 50 random selected parcels for each model zone in Orange County and the total 15 model zones in the same county. The selection of 200 parcels with positive floor space has no model zone constraint.

2.1.1 Statistics on parcels with missing values

By ‘ground truth trekking’ in Google Earth, we found that of 750 randomly selected **non-vacant** (parcels with land use code 11xx, 12xx and 13xx are considered. LU08 3xxx indicating vacant land is not included. For SCAG land use classification table, refer to ftp://mrpi.geog.ucsb.edu/data/11_Land_use/LU_CODE/ and ftp://mrpi.geog.ucsb.edu/data/10_Parcel/GP_LU_Correspondent.doc)

Output: The generated results for these analysis can be found at:

ftp://mrpi.geog.ucsb.edu/data/15_FAR/ground_truth_trecking/Riverside/RVimp_with_adjusted_floor_space_result_ToYiZhen.xlsx

The GIS data is under the folder:

ftp://mrpi.geog.ucsb.edu/data/15_FAR/ground_truth_trecking/Riverside/GIS

For parcels with improvement data (Parcel has improvement if there are constructions on it. The improvement is measured by the total floor space in that parcel) missing or equal to 0,

(1) 295 parcels have no improvement, therefore, the accuracy is at about 39.2%

	has actually non-improvement	has improvement	Subtotal
Single-Family Residential	177	243	420
Multi-Family Residential	73	98	171
Commercial	34	91	125
Industrial	11	23	34
SubTotal	295	455	750

Table 1. Statistics of parcels on the existence of improvement per land use type

(2) Ratios of parcels have or have not improvement among the selected parcels.

Ratio	has actually non-improvement	has improvement	Subtotal
Single-Family Residential	42%	58%	56.0%
Multi-Family Residential	43%	57%	22.8%
Commercial	27%	73%	16.7%
Industrial	32%	68%	4.5%
SubTotal	39%	61%	100.0%

Table 2. Ratios of parcels having improvement and having no improvement per each land use type.

Note: The determination of land use type is according to the following rules:

#left 3 digits of lu08 code

#0< x <112 --->[0] → Single-family residential parcels

#112<=x<120 --->[1] → Multi-family residential parcels

#120<=x<130 --->[2] → Commercial parcels

#rest --->[3] → Industrial parcels

Specifically, the left 3 digits of land use field "LU08" from 2007 parcel database were selected and a python code was written to classify selected parcels by single residential, multi-residential, commercial and industrial land use types. We observed that for residential parcels, those with no buildings and the parcels with missing values are almost the same; while for commercial and industrial parcels, the majorities are those actually having improvement but listed as missing improvement.

Table 1 and Table 2 illustrate the distribution (actual numbers and ratios) of sampled parcels that fall in different categories of land use. From table 2, we can tell that most (78.8%) of the parcels are residential and only 4.53% of the total sampled parcels are of industrial type.

FAR (Floor space – Land area ratio)

FAR with 0-improvement parcels included				
Model Zone	Single-family Residential	Multi-family Residential	Commercial	Industrial
mz 81	0.22	0.05	0.36	0.38
mz 82	0.14	0.51	0.27	0.21
mz 83	0.14	0.19	0.41	0.03
mz 84	0.04	0.09	0	0
mz 85	0.01	0.04	0.03	0
mz 86	0.22	0.32	0.04	0.31
mz 87	0.13	0.27	0.1	0
mz 88	0.04	0.05	0.17	0
mz 89	0.06	0.15	0.03	0.11
mz 90	0.11	0.13	0.12	null

mz 91	0.26	0.34	0.22	0.62
mz 92	0.19	0.41	0.29	null
mz 93	0.01	0.03	0.08	null
mz 94	0.09	0.24	0.31	0

Table 3. FAR calculated from the sampled parcels (including those actually with 0 improvements).

FAR with parcels have improvement ONLY				
Model Zone	Single-family Residential	Multi-family Residential	Commercial	Industrial
mz 81	0.29	0.06	0.37	0.39
mz 82	0.16	0.59	0.3	0.21
mz 83	0.16	0.26	0.44	0.03
mz 84	0.23	0.1	0.55	null
mz 85	0.1	0.06	0.31	null
mz 86	0.25	0.32	0.26	0.32
mz 87	0.19	0.28	0.38	null
mz 88	0.14	0.08	0.24	0.07
mz 89	0.07	0.18	0.04	0.14
mz 90	0.33	0.29	0.13	null
mz 91	0.41	0.36	0.23	0.75
mz 92	0.38	0.51	0.29	null
mz 93	0.35	0.51	0.09	null
mz 94	0.32	0.64	0.34	null

Table 4. FAR calculated from the sampled parcels (without parcels having actually no improvement).

Table 3 and Table 4 list the averaged floor space – land ratio on sampled parcels on each type of land use. Table 3 compute the average FAR value for every parcel in the sampled parcel dataset; while Table 4 compute the average FAR value for those have improvement only. Figure 1-4 demonstrate comparisons on averaged FAR of each model zone for different land use types. For each model zone, the FAR are computed on both the parcel set in which every parcel has improvement (we call it set A) and the complete sample set in which the parcels may or may not have improvement (we call it set B). It can be told that the FAR computed on the parcel set in which every parcel has improvement is greater than that considering all sampled parcels (parcels that do or do not have improvement). From Figure 1, we can also see that for model zone 84, 85, 93 and 94, the FARs computed on afore mentioned two parcel sets are quite different. Specifically, for single residential parcels in model zone 84, the FAR (0.2267) on set A is much greater than the FAR (0.0426) on set B. Similarly, the FAR for single residential parcels in model zone 85, 93 and 94 on set A are all greater than that on set B. This may mean that the random samples identified as single-family residential types in these zones are more likely to have less buildings (or possibly large area of gardens). For the random selected multi-residential parcels, with the exception of model zones 93 and 94 most of the randomly selected multi-residential parcels have buildings. That's why the FAR on set B and set A are almost the same, as shown in Figure 2. For commercial parcels, the great FAR differences on set A and set B occur in model zone 84-87. However, for industrial parcels, there are no parcels being sampled during the selection procedure in seven model zones, where "null" is listed (in Table 4). This is because industrial parcels are of only a small amount of total parcels and only 23 industrial parcels were selected. Therefore we plan to increase the sample size of parcels in industrial type in the following model zones for further ground truth trekking as next step work.

- Model zone 84
- Model zone 85
- Model zone 87
- Model zone 90
- Model zone 92
- Model zone 93
- Model zone 94

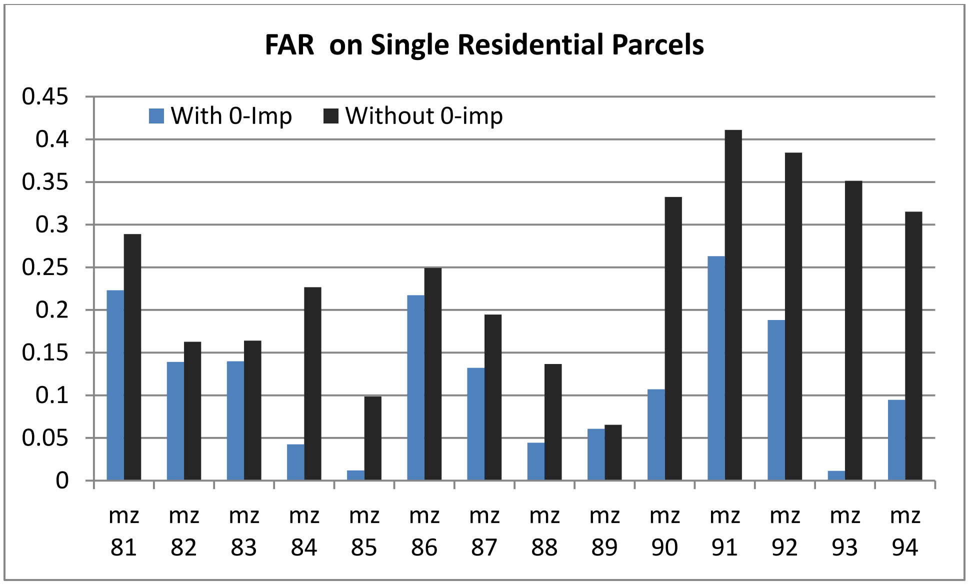


Figure1 FAR comparison for single residential parcels (generated from Table 3 and 4)

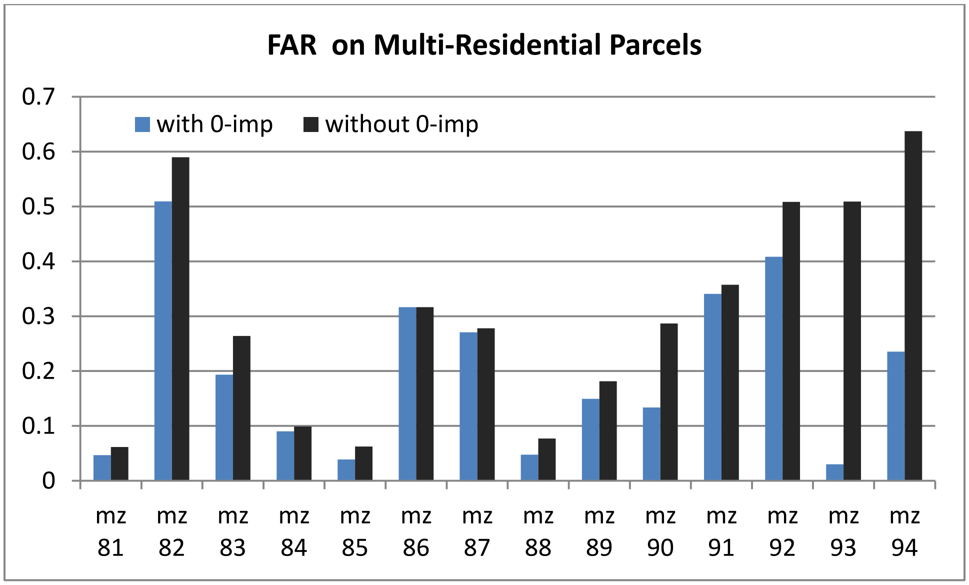


Figure 2 FAR comparison for multi-residential parcels (generate from Table 3 and Table 4)

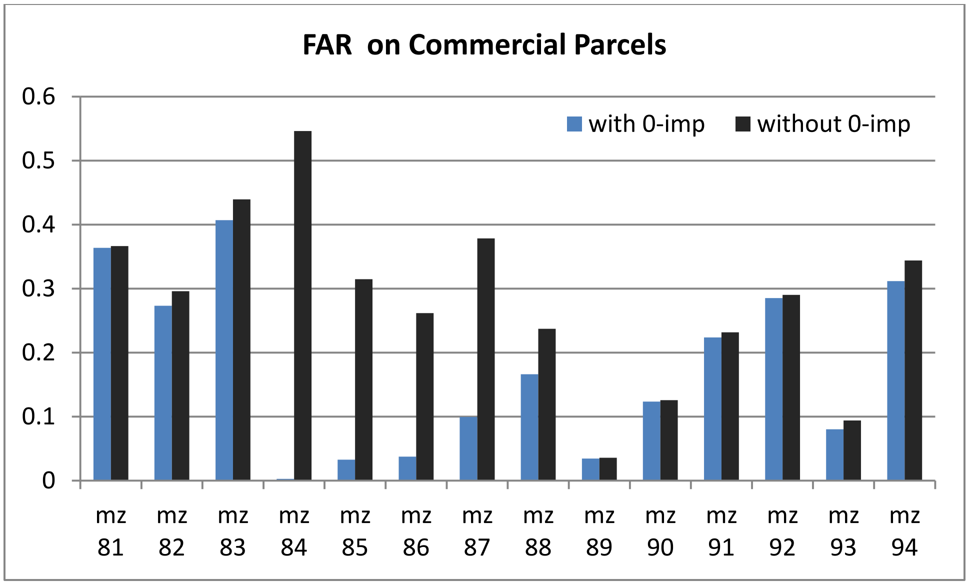


Figure 3. FAR comparison for commercial parcels (generated from Table 3 and Table 4)

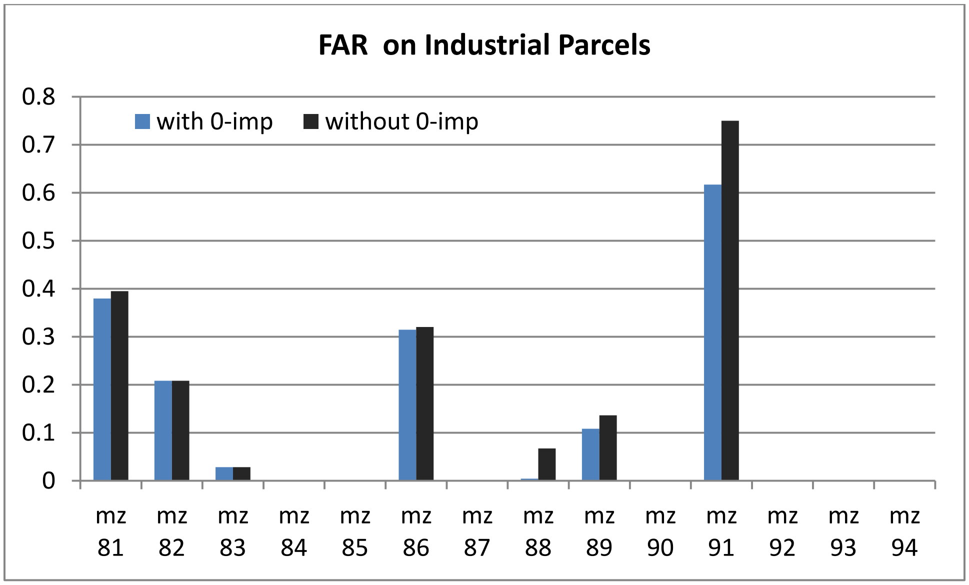


Figure 4. FAR comparison for industrial parcels (generated from Table 3 and Table 4)

2.1.2 Statistics on parcels with positive floor space values

There is a concern about the “accuracy” of Google Earth estimation of floor space: there is potentially a systematic overestimate of floor area using Google Earth method since it includes garage area (and probably patio area) which would not be counted as floor area in the parcel database. Since this is the case, 200 parcels with positive floor spaces in the SCAG 2007 parcel data were randomly selected, the floor space were estimated using Google Earth and Google Street View and the values were compared to those listed in the parcel database.

Output:
ftp://mrpi.geog.ucsb.edu/data/15_FAR/ground_truth_trecking/Riverside/RVhasImpCombinedResult200.xls

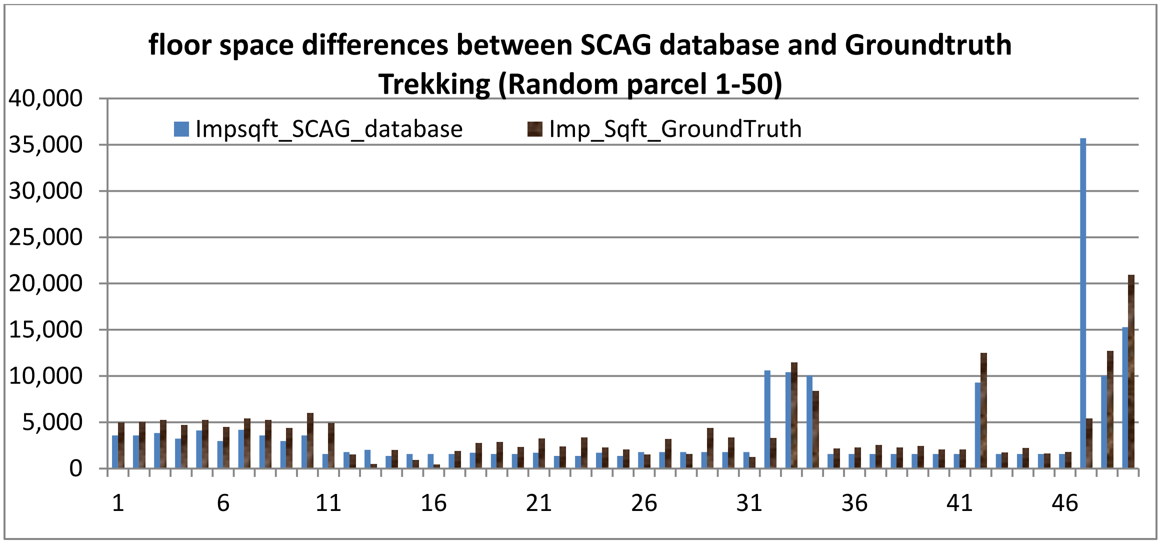


Figure 5 (a). Differences of floor space values between SCAG database and ground-truth trekking (random parcel 1-50)

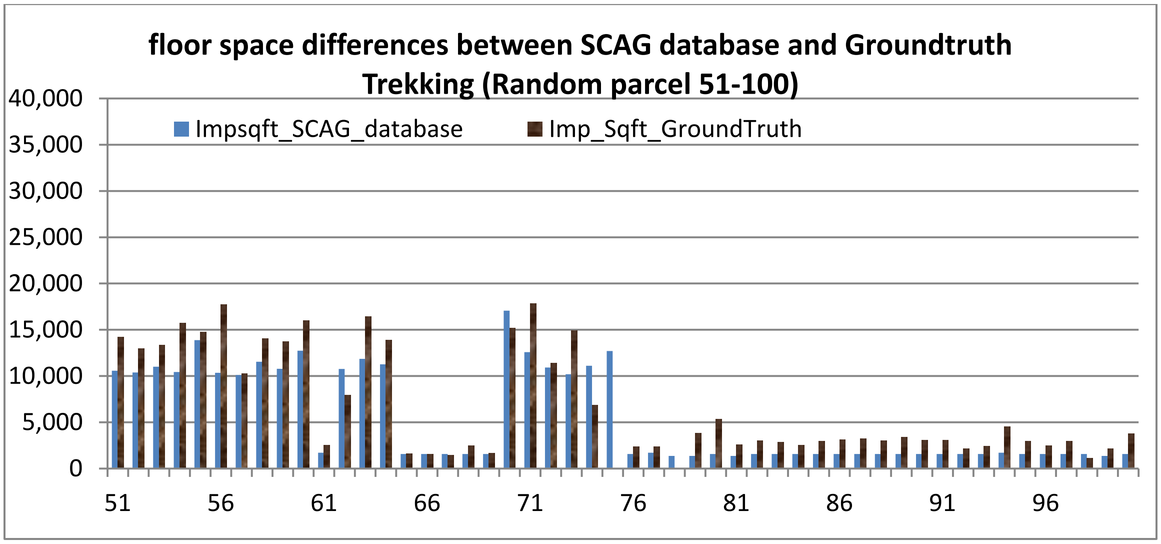


Figure 5 (b). Differences of floor space values between SCAG database and ground-truth trekking (random parcel 51-100)

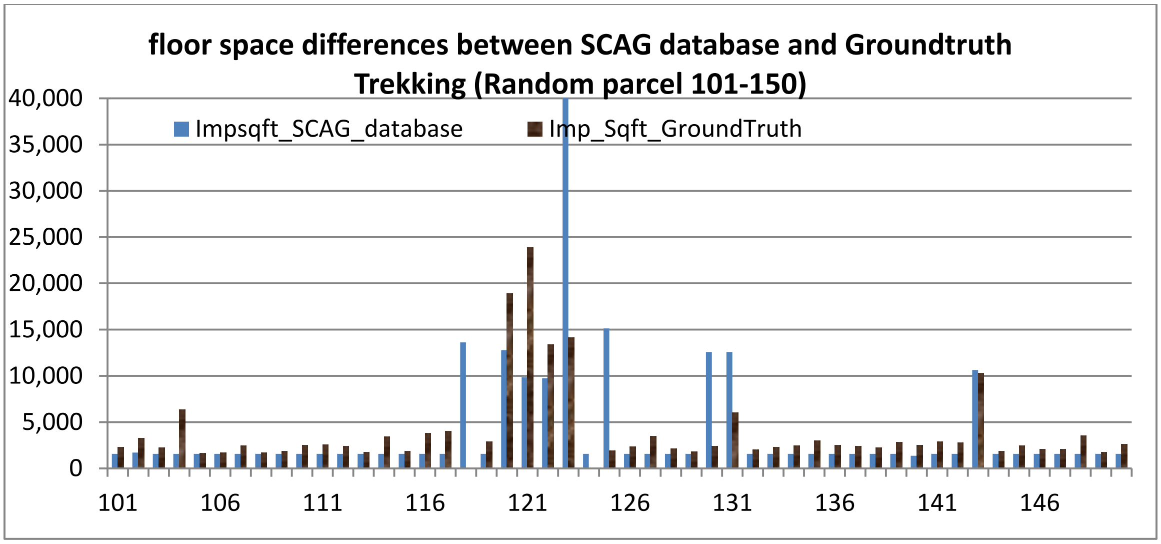


Figure 5 (c). Differences of floor space values between SCAG database and ground-truth trekking (random parcel 101-150)

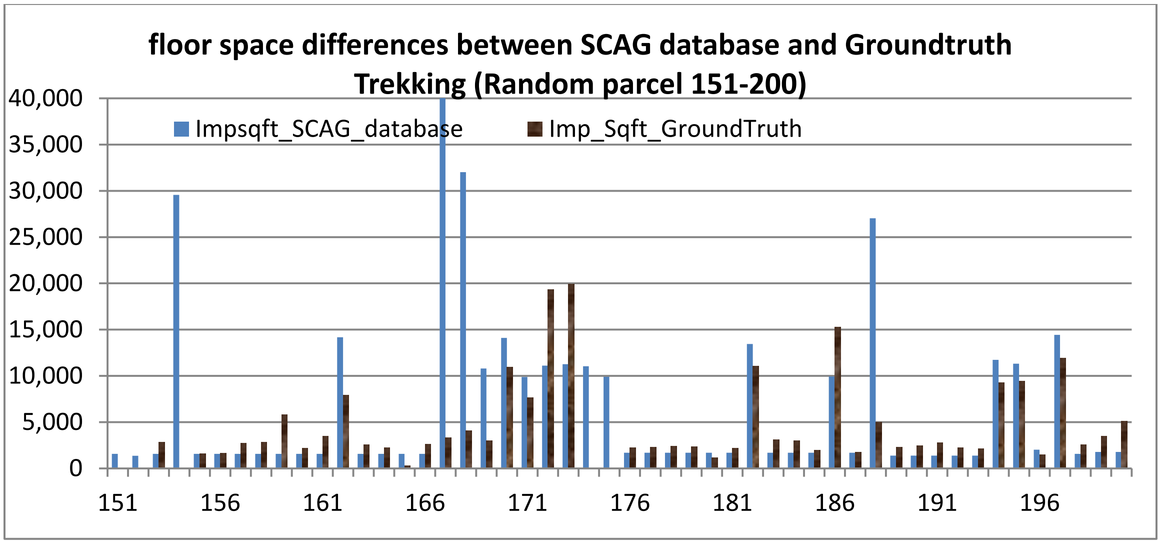


Figure 5 (d). Differences of floor space values between SCAG database and ground-truth trekking (random parcel 151-200)

Figure 5 (a)-(d) show the comparison of floor space recorded in SCAG database and that obtained from ground truth trekking. We can tell that results from ground truth estimation are slightly higher than the value in SCAG database because ground truth estimation will include the area of garage. On average, the FAR over 200 parcels with positive improvement is 0.30 and the floor space weighted FAR is 0.25695. Either of these ratios could be used to estimate the “true” FAR to adjust the bias/error in the SCAG database. In some records the actual floor space obtained from ground truth trekking is substantially smaller the values recorded in the parcel database. It is probably that this is due to the deconstruction of buildings on the parcel since 2000 or errors in the SCAG parcel database. To overcome the issues of overestimation, some procedure needs to be developed to adjust the estimation to value with higher confidence level.It would be very helpful if an architecture expert could help us with this.

The following tables show the average FAR on selected parcels (200 parcels) that have positive improvement in SCAG parcel database in Riverside County:

(1) FAR got from ground truth trekking

FAR				
Model Zone	Single-Residential	Multi-Residential	Commercial	Industrial
mz 81	0.29	0.03	null	0.04
mz 82	0.23	null	null	0.09
mz 83	0.18	0.41	null	null
mz 84	0.09	0.03	null	0.03
mz 85	0.06	0.07	null	0.03
mz 87	0.13	0.06	null	null
mz 88	0.18	0.05	0.04	0.01
mz 89	null	0.14	null	0
mz 90	0.29	0.07	null	0.03
mz 91	0.53	0.13	null	0.03
mz 92	0.37	null	null	0.13
mz 93	0.28	null	null	0.05
mz 94	0.31	null	null	null

Table 5. FAR computed from 200 randomly selected parcel with positive improvement

(2) FAR got from SCAG parcel database

FAR				
Model Zone	Single-Residential	Multi-Residential	Commercial	Industrial
mz 81	0.2	0.06	null	0.04
mz 82	0.12	null	null	0.14
mz 83	0.36	0.31	null	null
mz 84	0.2	0.01	null	0.02

mz 85	0.13	0.05	null	0.03
mz 87	0.26	0.03	null	null
mz 88	0.18	0.03	0.03	0
mz 89	null	0.06	null	0.02
mz 90	0.38	0.13	null	0.03
mz 91	0.71	0.21	null	0.02
mz 92	0.24	null	null	0.13
mz 93	0.28	null	null	0.04
mz 94	0.23	null	null	null

Table 6. FAR computed from the same 200 parcels with improvement values from SCAG database.

Table 5 and Table 6 demonstrate the FARs computed on the 200 parcels with positive improvement in SCAG database by different methods (ground truth trekking and SCAG database). Because of the inclusion of garage area in the ground truth trekking method, the FAR in Table 5 is slightly higher than the value in Table 6 for almost every model zone in Riverside County. This finding is consistent with that found from Figure 5. On average, the FAR over the 750 parcels is 0.24 and the floor space weighted FAR is 0.4385, these are slightly different from the ratios calculated from the 200 sampled parcels discussed in the previous paragraph. This is probably due to the unequal distribution of parcels of different land use types in the two sample sets. For further comparison, we need to keep the distributions about the same.

Additionally, we found that among the 200 selected parcels, commercial parcels are only at a very small fraction. This means that either the amount of total commercial parcels in Riverside County is very small or most of the commercial parcels in the parcel database are missing floor space value; therefore very few records were selected during the random selection process. This conclusion will need further proof which can be accomplished through close examination of the parcel database.

2.2 Results for Imperial County

There are two datasets used for ground truth trekking in Imperial County:

(a) ImperialRanSelectedParcels: are the 100 parcels randomly selected from parcels with 0 improvement or missing values. By a checking it was found that most of parcels have actually 0 improvement. Since this dataset is less representative the dataset b) was generated;

The **dataset** is available at:

ftp://mrpi.geog.ucsb.edu/data/15_FAR/ground_truth_trecking/Imperial/ImperialRanSelection100.shx

(b) ImperialRanSelectedParcelsMissing: are 68 parcels from the parcels of which the improvement values are missing.

The **dataset** is available at:

ftp://mrpi.geog.ucsb.edu/data/15_FAR/ground_truth_trecking/Imperial/selectedMissingRecords.xlsx

Output: The resulted dataset could be found at:

ftp://mrpi.geog.ucsb.edu/data/15_FAR/ground_truth_trecking/Imperial

2.2.1 Statistics on parcel set a)

Some basic statistics are applied to the dataset: we counted the number of parcels that have improvement and that do not have improvement from the sample set in which all parcels are listed as 0-improvement parcels in the SCAG database.

By ‘ground truth trekking’ in Google Earth, we found that of 100 randomly selected **non-vacant** (the definition of ‘non-vacant’ is the same as that defined in 2.1.1) parcels with improvement data missing or equal to 0,

(1) 76 parcels have no improvement, therefore, the accuracy is at about 76%

	has actually non-improvement	has improvement	Subtotal
Single-Family Residential	36	17	53
Multi-Family Residential	21	4	25
Commercial	17	3	20
Industrial	2	0	2
Subtotal	76	24	100

Table 7. Statistics of parcels on the existence of improvement per land use type

(2) Ratios of parcels have or have not improvement among the selected parcels.

Ratio	has actually non-improvement	has improvement	Subtotal
Single-Family Residential	68%	32%	53%
Multi-Family Residential	84%	16%	25%
Commercial	85%	15%	20%
Industrial	100%	0	2%
Subtotal	76%	24%	100%

Table 8. Ratios of parcels having improvement and having no improvement per each land use type.

The following tables show the FAR on selected parcels (on dataset a) that have improvement missing or equal to 0 in Imperial County:

Include 0-improvement parcels				
Model Zone	Single-Residential	Multi-Residential	Commercial	Industrial
mz 96	0.002	0.003	0.0000629	0
mz 97	0.019	0.033	0.0702	0

Table 9. FAR calculated from the sampled parcels (including those with 0 improvements).

Not Include 0-improvement parcels				
Model Zone	Single-Residential	Multi-Residential	Commercial	Industrial
mz 96	0.168	0.043	0.27	null
mz 97	0.237	0.125	0.091	null

Table 10. FAR calculated from the sampled parcels (including those with 0 improvements).

* Selection procedure: same as Riverside County, 50 parcels per model zones

* 0 means no improvement found; null means no parcels with that land use code are selected

Table 9 and Table 10 demonstrate the FAR analysis from ground truth trekking. Results show the same trend as that found for Riverside County: the FAR is larger when 0-improvement parcels are included (as Table 9 shows). For commercial parcels, the FAR in both zones are almost 0. In addition, there are non industrial parcels selected for ground truth trekking in both model zones. Therefore, records for industrial and commercial parcels should be increased.

2.2.2 Statistics on parcel set b) (58 records in total)

By ‘ground truth trekking’ in Google Earth, we found that of 58 randomly selected **non-vacant** (the definition of ‘non-vacant’ is the same as that defined in 2.1.1) parcels with improvement data missing or equal to 0,

(1) 27 parcels have no improvement, therefore, the accuracy is at about 46.6%

	has actually non-improvement	has improvement	Subtotal
Single-Family Residential	17	15	32
Multi-Family Residential	1	10	11
Commercial	3	2	5
Industrial	6	4	10
Subtotal	27	31	58

Table 11. Statistics of parcels on the existence of improvement per land use type

(2) Ratios of parcels have or have not improvement among the selected parcels.

Ratio	has actually non-improvement	has improvement	Subtotal
Single-Family Residential	53.1%	46.9%	55.2%
Multi-Family Residential	9.1%	90.9%	19.0%
Commercial	60.0%	40.0%	8.6%
Industrial	60.0%	40.0%	17.2%
Subtotal	46.6%	53.4%	100%

Table 12. Ratios of parcels having improvement and having no improvement per each land use type.

Include 0-improvement parcels				
Model Zone	Single-Residential	Multi-Residential	Commercial	Industrial
mz 96	0.0016	0.1407	0.0019	0.0027
mz 97	0.1498	0.1615	0.0157	0.2984

Table 13. FAR calculated from the sampled parcel set b) (including those with 0 improvements).

Not Include 0-improvement parcels				
Model Zone	Single-Residential	Multi-Residential	Commercial	Industrial
mz 96	0.1058	0.1407	0.0029	0.1431
mz 97	0.1637	0.1621	0.0315	0.3501

Table 14. FAR calculated from the sampled parcel set b) (including those with 0 improvements).

Table 13 and Table 14 list the FARs computed from another sample set in which more records have improvement. We can see the clearly increase on FARs averaged from every record in both sample sets on Table 13 than Table 7. Interestingly, the averaged FARs per model zone at the categories of single family residence and multi-family residence show a decrease in Table 14 than that in Table 8. That is probably because that the parcels in the sample set b) have a very large lot size, which cause the small FAR during the computation.

2.3 Results for Orange County

The dataset used for below analysis can be found at:

ftp://mrpi.geog.ucsb.edu/data/15_FAR/ground_truth_trecking/Orange/

The resulted dataset found be found at:

OUTPUT: ftp://mrpi.geog.ucsb.edu/data/15_FAR/ground_truth_trecking/Orange/result/ RanSelectedParcelsOR.xlsx

By ‘ground truth trekking’ in Google Earth, we found that of 850 randomly selected **non-vacant** (the definition of ‘non-vacant’ is the same as that defined in 2.1.1) parcels with improvement data missing or equal to 0 in Orange County,

(1) 202 parcels have no improvement, therefore, the accuracy is at about 23.7%

	has actually non-improvement	has improvement	Subtotal
Single-Family Residential	99	438	537
Multi-Family Residential	36	94	130
Commercial	51	84	135
Industrial	16	32	48
Subtotal	202	648	850

Table 15. Statistics of parcels on the existence of improvement per land use type

(1) Ratios of parcels have or have not improvement among the selected parcels.

Ratio	has actually non-improvement	has improvement	Subtotal
Single-Family Residential	18.4%	81.5%	63.1%
Multi-Family Residential	27.6%	72.3%	15.2%
Commercial	37.7%	62.2%	15.8%
Industrial	33.3%	66.6%	5.64%
Subtotal	23.7%	76.2%	100%

Table 16. Ratios of parcels having improvement and having no improvement per each land use type.

The following tables show the FARs on selected parcels that have improvement missing or equal to 0 in Orange County:

Include 0-improvement parcels:				
Model Zone	Single Residential	Multi-residential	Commercial	Industrial
mz 50	0.503	0.823	0.258	null

mz 51	0.658	0.454	0.249	0
mz 52	0.573	0.499	0.514	0.34
mz 53	0.529	0.473	0.206	0.139
mz 54	0.692	0.506	0.278	0.403
mz 55	0.324	0.122	0.385	0
mz 56	0.377	0.302	0.15	0.505
mz 57	0.591	0.412	0.305	0.384
mz 58	0.328	0.12	0.118	0.396
mz 59	0.298	0.066	0	null
mz 60	0.258	0.761	0.428	0
mz 61	0.114	0.367	0.688	0.748
mz 62	0.293	0.173	0.449	null
mz 63	0.47	0.342	0.026	null
mz 64	0.258	0.372	0.152	0.269
mz 65	0.493	0.192	0.053	null
mz 66	0.396	0.487	0.122	0.168

Table 17. FARs computed from ground truth trekking (every sampled record is included)

Not Include 0-improvement parcels:

Model Zone	Single Residential	Multi-residential	Commercial	Industrial
mz 50	0.561	1.022	0.278	null
mz 51	0.695	0.474	0.26	null
mz 52	0.649	0.5	0.533	0.375
mz 53	0.552	0.602	0.236	0.491
mz 54	0.789	0.521	0.28	0.497
mz 55	0.458	0.165	0.385	null
mz 56	0.74	0.997	0.256	0.505
mz 57	0.66	0.578	0.32	0.384
mz 58	0.351	0.123	0.239	0.396
mz 59	0.41	0.721	null	null
mz 60	0.779	0.771	0.437	null
mz 61	0.295	0.429	0.762	0.912
mz 62	0.552	0.896	0.487	null
mz 63	0.683	0.704	0.108	null
mz 64	0.36	0.381	0.187	0.275
mz 65	0.577	0.587	0.06	null
mz 66	0.514	0.507	0.125	1.068516587

Table 18. FARs computed from ground truth trekking (only parcels having improvement are included)

Table 17 and Table 18 demonstrate the comparison of FARs computed from ground truth trekking with 0-improvement parcels included (Table 17) and excluded (Table 18). Comparing to that in Riverside County and Imperial County, we can tell that the FAR ratio in Orange County is higher than the other two Counties. Interestingly, in some model zones, the FARs are higher than 1 (highlighted cells), which are very rare in other Counties. For model zone 50, 51, 55, 59, 60, 62, 63 and 65, more industrial parcels need to be further checked.

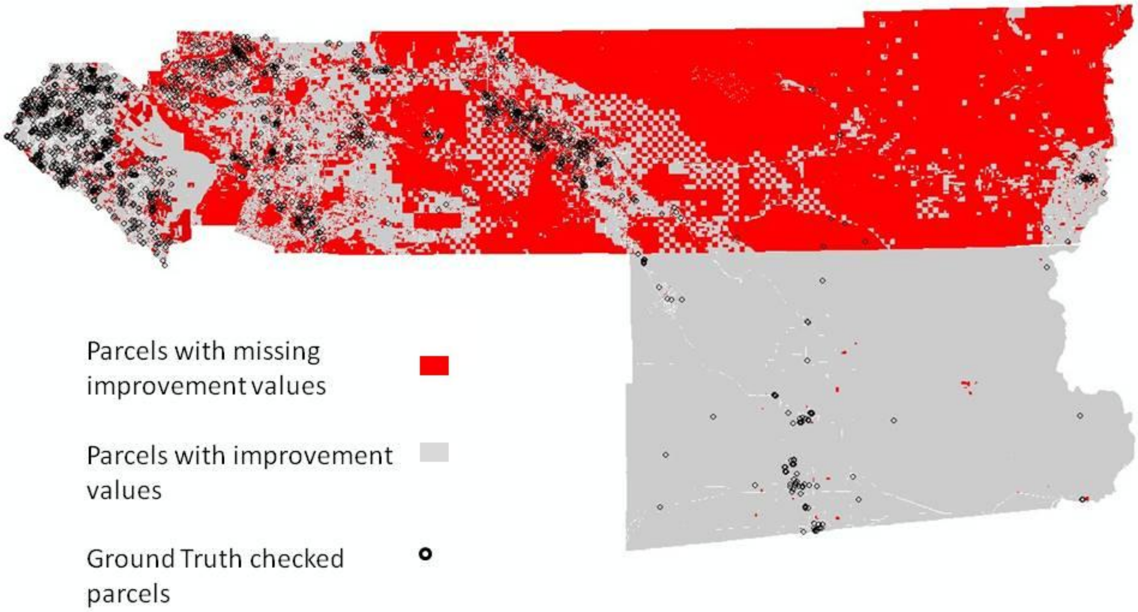
3 Summary and Discussion

We conducted systematic investigation on the 'ground-truth' floor space in Imperial County, Riverside County and Orange County, in which a pre-study showed that the floor space dataset is not very reliable in the SCAG database. The error that exists in the SCAG database is mainly due to a certain amount of parcels which have improvement yet are recorded with zero improvement. This error will significantly influence the final result in computing the averaged FAR in all model zones for RELU-TRAN simulation. By randomly selecting parcels in different land use types and in different model zones, conducting ground truth trekking using Google Earth, Google Street View tools along with field work, we are able to provide the averaged FARs for parcels with improvement listed 0 in the SCAG database. Meanwhile the statistics show that the average FARs for all types of parcels in Imperial County and Riverside County is smaller than those in Orange County. This observation agrees with economy theory that FAR's will be higher in areas with higher land rents and values, as these areas tend to be more developed and have higher densities. The above analysis indicates the feasibility of using ground-truth trekking results to clean up the SCAG parcel database.

The ground-truth trekking for the above counties was conducted between July to September in 2011, and the imagery data in Google Earth was taken in 2009, while the land use classification code from SCAG database was collected in 2008. Therefore, there is one year difference between the SCAG dataset and ground truth trekking results. If there were a huge construction or demolition occurred between 2008 and 2009, the analysis may have some bias. Luckily, it is known that the construction boom in Riverside County is between 2003 to 2008; therefore, the differences in date shouldn’t make a big difference.

Appendix I:

Map of Imperial, Orange, and Riverside County showing Parcels Floor Space Errors


Parcels with missing improvement values
Parcels with improvement values
Ground Truth checked parcels

Notes: While it appears the majority of Riverside County has parcels with missing improvement values the number is approximately only 1/3 the total number of parcels in the county. For Imperial County the parcels with missing improvement values is ~ 1%; for Orange County it is approximately 15%. Ground truth sampling was accomplished for the parcels that have been circled.

Code I: Python Code for Random Selection

```
import os
import csv
import random


# Function:
# Randomly select 50 parcels per model zone from 0/missing-improvement parcels


# The CSV file is all the parcels that miss or have floor space 0
# There are three fields:
# Field one:   ObjectID
# Field two:   SCAG_XYID
# Field three: NewMZ


# MZ --> left(2) --> integer
# ranked by MZ


datafolder = "E:/Project/MRPI/data/15_FAR/data_anormalies/Imperial/Imperial_0_improvement_new.csv"
#"E:/Project/MRPI/data/15_FAR/data_anormalies/Riverside/data/MissDataForRanSel.csv"



parcels = {}


#parcel[lu code left 3 digit]=[land-0,land-1,FAR]
arr = []
```

```
wb = csv.reader(open(datafolder,'rb'),delimiter=',')
newmz = 0
for row in wb:
    if len(row[2])<3:
        if row[2]!= newmz:
            if len(arr) >0:
                parcels[int(newmz)] = arr
                print newmz, ": length = ", len(arr)
            newmz = row[2]
            arr = []
        arr.append(row[1])
parcels[int(newmz)] = arr
print newmz, ": length = ", len(arr)
for i in parcels:
    #ranindexarr is the randomly selected indexes
    ranindexarr = []
    mzparcels = parcels[i]
    while len(ranindexarr)<50:
        index = random.randint(0,len(mzparcels)-1)
        if not (index in ranindexarr):
            ranindexarr.append(index)
    #print ranindexarr
    for j in ranindexarr:
        print i,"   ", mzparcels[j]
    #for each model zone, write down the parcels (parcelxyid,mz) that have been selected
```

Code II: Statistics on the FAR per land use type

```
import xlrd
import numpy as N


filePath = "E:/Project/MRPI/data/15_FAR/data_anormalies/Orange/coding/"
#result/"
filename = "RanSelectedParcelsOR_input.xls" #"RVhasImpCombinedResult200_groundtruth.xls" #"FAR_on_RandomSample_06302011.xls"


def inita():
    a = []
    for ii in xrange(4):
        a.append([])
        for jj in xrange(2):
            a[ii].append([])
            for kk in xrange(2):
                a[ii][jj].append(0)
    return a
wb = xlrd.open_workbook(filePath+filename)
#Check the sheet names
print wb.sheet_names()


#Get the first sheet either by index or by name
sh = wb.sheet_by_index(0)


# all fie columns:
mz = sh.col_values(0)
```

```
lu08 = sh.col_values(1)

luleft3 = sh.col_values(2)

lotsqft = sh.col_values(3)

impsqft = sh.col_values(4)


mz = N.array(mz,dtype=int)

lu08 = N.array(lu08,dtype=int)

luleft3 = N.array(luleft3,dtype=int)

lotsqft = N.array(lotsqft,dtype=float)

impsqft = N.array(impsqft,dtype=float)


dataPerMZ = {}

#lu code

#0< x <112  --->[0]

#112<=x<120 --->[1]

#120<=x<130 --->[2]

#rest       --->[3]

mzcurrent = -1

#data --> [4][2][2]

#first dimension: four land use type

#second dimenison: 0: sum consider 0-imp; 1: sum not consider 0-imp parcels

#third dimension: 0: lotsqft 1: improvement

mzcurrent = mz[0]

a = inita()

print a

for i in range(len(mz)):
    if mz[i]!=mzcurrent:

        dataPerMZ[mzcurrent] = a

        mzcurrent = mz[i]

        a = inita()

        #print a

    secondD = 3

    if luleft3[i]>0 and luleft3[i]<112:

        secondD = 0

    elif luleft3[i]>= 112 and luleft3[i]<120:

        secondD = 1

    elif luleft3[i]>=120 and luleft3[i]<130:

        secondD = 2

    if impsqft[i] > 0:

        thirdD = 1

        a[secondD][thirdD][0] += lotsqft[i]

        a[secondD][thirdD][1] += impsqft[i]

    thirdD = 0

    a[secondD][thirdD][0] += lotsqft[i]

    a[secondD][thirdD][1] += impsqft[i]

dataPerMZ[mzcurrent] = a

FAR = []

for ii in xrange(len(dataPerMZ)):

    FAR.append([])

    for jj in xrange(4):

        FAR[ii].append([])

        for kk in xrange(2):
            FAR[ii][jj].append(0)

print "Include 0-improvement parcels:"
i = 0
for mzi in dataPerMZ:
    j = 0
    for lui in dataPerMZ[mzi]:
        if lui[0][0]==0:
            FAR[i][j][0] = 'null'
        else:
            FAR[i][j][0] = lui[0][1]/(lui[0][0]*1.0)
        if lui[1][0]==0:
            FAR[i][j][1] = 'null'
        else:
            FAR[i][j][1] = lui[1][1]/(lui[1][0]*1.0)
        j += 1
    print "mz ",mzi,"   ",FAR[i][0][0],"   ",FAR[i][1][0],"   ",FAR[i][2][0],"   ",FAR[i][3][0]
    i += 1

print "Not Include 0-improvement parcels:"
i = 0
for mzi in dataPerMZ:
    print "mz ", mzi, "   ",FAR[i][0][1],"   ",FAR[i][1][1],"   ",FAR[i][2][1],"   ",FAR[i][3][1]
    i += 1

print "all together improve,"
```