Technical Report

Floor Area Data Adjustment for the Parcel Database of Greater Los Angeles Region*

Yizhen Gu and Richard Arnott

yizhengu@berkeley.edu; richard.arnott@ucr.edu

Nov 16, 2011

Abstract: Floor area is an important input in the RELU-TRAN model. However, there are more "0"s and blanks in the parcel database of Greater LA region than expected. A two-step procedure is used for adjusting the floor area data for parcels that have a floor area entered as "0" or blank. First, these parcels are probabilistically categorized as actually being developed or not by estimating a binary development status model. Second, for parcels "identified" as being developed, a floor area is imputed by estimating a floor-to-area ratio (FAR) model. The procedure is partly justified by the negligible variation in the aggregate floor area estimates at the model zone level across various tests.

* We thank Judy Pereira, Huiling Zhang, Ross Guo, and Wenwen Li for their help with this report.

Floor Area Data Adjustment for the Parcel Database of Greater Los Angeles

1. Introduction

In the Regional Economy, Land Use, and Transportation (RELU-TRAN) model, floor area is an important input for calculating the equilibria of floor space markets, as well as for calculating the general equilibrium for the whole model. To apply the RELU-TRAN model to the Greater Los Angeles (LA) region, the parcel database of six counties in the study area is obtained from the Southern California Association of Governments (SCAG). Strictly speaking, the parcel database is compiled and recoded by the SCAG from the raw data provided by the Assessor's Offices in these counties. This parcel database records the land-use type, lot size, floor area, assessed value, etc., for each parcel.

However, a preliminary investigation of floor area data shows that there are more "0"s and blanks in the database than expected. For example, nearly one-quarter of the multi-family residential parcels in Riverside County have a floor area entered as "0" or blank (Table 1). This ratio would be even higher for other land-use types such as office and retail.[1] This indicates the possibility of severe coding errors in the floor area data. As provided by the University of California, Santa Barbara (UCSB) team, ground truth tracking results, which offer the rough floor area using Google Earth and Street View tools, have validated this likelihood. Of 750 randomly chosen parcels in Riverside County with floor area entered as "0" or blank, more than 60% (455 parcels) have actually been developed. Thus, it is necessary to make appropriate adjustments for parcels with floor area data entered as "0" or blank, in order to avoid the underestimation of aggregate floor area at the model zone level.[2]

Table 1. Percentage of parcels in Riverside County with floor area entered as "0" or blank.

Land-use type	% of parcels with floor area entered as	
	"0"	Blank
Single-family residential	2.9	10.5
Multi-family residential	1.8	22.3
Mixed residential	18.9	7.4
Office	3.9	59.5
Retail	5.7	84.8
Other commercial	7.0	73.9

[1] A land use code other than the SCAG code is applied in this project. For the correspondence between these two codes, see Appendix A.

[2] At present, no adjustment will be made for parcels with positive floor area data, although there have been a few cases, found in the ground truth tracking, where parcels with positive floor area data are indeed vacant.

Land-use type	% of parcels with floor area entered as	
	“0”	Blank
Public	9.2	77.0
Warehousing	5.9	92.5
Industrial	10.6	61.5
Transportation/Communication/Utilities	15.8	78.6
Mixed	6.8	85.9
Vacant	40.7	42.9
Other	24.4	63.8

Note: The percentage of parcels with floor area entered as "0" (or blank) is calculated as the number of parcels entered as "0" (or blank) in that land-use type divided by the total number of parcels in that land-use type.

This report documents how we adjust the floor area data using Riverside County as an example. First, by estimating a binary development status model using the 750 random parcels mentioned above, parcels with floor area entered as “0” or blank are probabilistically categorized as actually being developed or not. Second, for parcels “identified” as being developed, a floor area is imputed using the estimation results of a floor-to-area ratio (FAR) model. Applying such a procedure will greatly improve the reliability of aggregate floor area data at the model zone level, which is a main exogenous variable in the LA model.

This report is structured as follows. Section 2 summarizes the procedure for adjusting the floor area data. Section 3 and Section 4 address how this procedure is used to probabilistically categorize “vacant” parcels as actually being developed and then to impute a floor area for those parcels identified as being developed, respectively. Section 5 discusses the “robustness” of this procedure, and Section 6 concludes the report.

2. Procedure

Adjusting the floor area data for the LA model is a little more complicated than simply interpolating missing data, because it is not known which parcels that actually do have a floor area are missing, due to the coding errors. Therefore, the first step is to figure out which parcels are actually developed but have a floor area incorrectly entered as “0” or blank.

Since there are many parcels with a floor area entered as “0” or blank, it is not feasible to do ground truth tracking for each of these parcels. Thus, a binary choice model, which analyzes the effect of accessibility measures on development status (developed or not), is first estimated using some randomly chosen parcels in Riverside County. Then, using the estimation results, the probability that parcels with a floor area entered as “0” or blank are actually developed is calculated. Since the calculated probability for each parcel characterizes a 0-1 Bernoulli distribution, we draw one realization from each distribution, to categorize each parcel as developed or not.

Using only one realization is problematic; nonetheless, this method is still adopted because a "real" floor area, instead of an expected value for each parcel in the database, is needed to build the LA model. Certainly, multiple realizations could be drawn to see the variation in the estimation results of the LA model, but then the results should be carefully interpreted if the model is applied to make policy evaluations. A preferred way is to use the expected floor area for these parcels (interpolated floor area multiplied by calculated development probability) if only the floor area at an aggregate level, e.g., model zone, is required.

The second step is to impute a floor area for parcels categorized in the first step as having been developed but with the floor area entered as "0" or blank. A model that regresses FAR on land area and accessibility measures for each land-use type is estimated using only parcels with positive floor area, as entered in the database in Riverside County. Then, the estimated model is used to interpolate a floor area for those "developed" parcels.

An implicit assumption here is that the expected FAR of a parcel conditional on being identified as developed, but with floor area entered as "0" or blank, is the same as a parcel conditional on having a positive floor area entered, *ceteris paribus*.[3] Also, note that the floor area obtained through truth tracking cannot be used for adjustments due to the unsystematic upward bias in that data.

The Office (OF) parcels and the Single-family residential (RS-SF) parcels are taken as examples, to illustrate how the procedure works. The estimation results for other land-use types in Riverside County are presented in Appendix B. The above procedure will not be applied to the Transportation/Communication/Utilities (TCU) parcels, the Vacant (V) parcels, and the Other (O) parcels. For these land-use types, we take parcels with floor area entered as "0" or blank as indeed vacant at this stage because: 1) the assessors claimed that the land use type provided in the parcel database was accurate so vacant parcels should really be vacant; 2) the TCU parcels and the Other parcels play a relatively unimportant role in the LA model; and 3) no truth tracking has been done for these three land-use types.

Appendix C provides the floor area adjustment procedures for other counties, except Imperial County of which the UCSB team is in charge. At present, the ground truth tracking results are only available for sample parcels in Riverside and Orange counties. Therefore, based on our local

[3] For public (P) parcels, the (positive) floor area given in the parcel database is generally lower than the actual value due to tax-exempt. Thus, further adjustment will be made in the next-step work.

knowledge, the development probability is calculated for parcels in San Bernardino County using the estimated Riverside development status model, and for parcels in Los Angeles and Ventura counties using the Orange model. More truth tracking will be done in the next stage, but, for now, it is worthwhile to note the potential bias in the floor area adjustment for San Bernardino, Los Angeles, and Ventura counties.

3. Identifying developed parcels with floor area entered as "0" or blank

This section addresses the development status model specification, data and estimation results, as well as how the estimated model is used to identify "developed" parcels with floor area entered as "0" or blank.

3.1. Model

The available explanatory variables for the development status model are limited: four accessibility measures (distance to the central business district (CBD), distance to the nearest subcenter, distance to the nearest freeway, and distance to the ocean) are calculated by the University of California, Riverside (UCR) team. There are also two dummy variables: 1) whether a parcel has its floor area entered as "0", rather than blank, and 2) whether it is a residential parcel (single-family residential (RS-SF), multi-family residential (RS-MF), or mixed residential (RS-MX)).

Three models are estimated. Model I is a basic binary probit model. Suppose that the latent variable follows

(1)

where represents the explanatory variables, is a vector of coefficients, and is a normally *iid* error term. Instead of observing , we observe only a binary variable indicating the sign of :

(2)

where represents the development status (developed or not). It is easy to see that

(3)

where is the cumulative distribution function (CDF) for the standard normal.

However, the *iid* assumption for the error term may be not valid due to the spatial autocorrelation in the omitted (location) variables. The result of the Moran's *I* test (0.090, significant at the 1%

level) for 750 sample parcels based on an inverse distance weighting matrix validates the existence of spatial autocorrelation in the variable of development status.[4] Therefore, we use spatial econometric techniques to deal with this issue (LeSage and Pace, 2009). Model II is a spatial error probit model as

(4)

where is a vector of error terms, is the spatial weighting matrix, measures the spatial interaction in error terms, and is a vector of normally *iid* disturbances.

Model III is a spatial autoregressive probit model as

(5)

where now directly measures the spatial interaction in the dependent variables. Note that it is vectors that are listed in this equation.

For each model, the estimation results of two specifications are given in the following. The first specification includes only four accessibility measures and the above-mentioned dummy variables. The second one controls for the city fixed effect.

3.2. Data

The actual development status for 750 parcels in Riverside County with floor area entered as "0" or blank was identified through ground truth tracking by the UCSB team (see Figure 1 for the geographic distribution of these parcels). These 750 parcels were selected by first generating a random number for all residential/commercial/industrial[5] parcels in Riverside County with a floor area entered as "0" or blank in 15 model zones, and then sorting them and selecting the top 50 parcels per model zone, to avoid replacement.

Descriptive statistics are given in Table 2. Among all 750 parcels, about 61% have actually been developed. An "average" parcel is 96.17 miles away from the CBD, 5.92 miles away from the nearest subcenter, 1.27 miles away from the nearest freeway, and 52.42 miles away from the ocean. Residential parcels account for nearly 80% of the 750 parcels (industrial parcels less than

[4] Using the same weighting matrix, the Geary's *c* statistic is 0.900, significant at the 1% level.

[5] Here, residential/commercial/industrial is identified by the variable "lu08" in the database and by the SCAG 1993 land-use code (11x/12x/13x), respectively.

5%). This is why we cannot control for each land-use type in the model.[6]

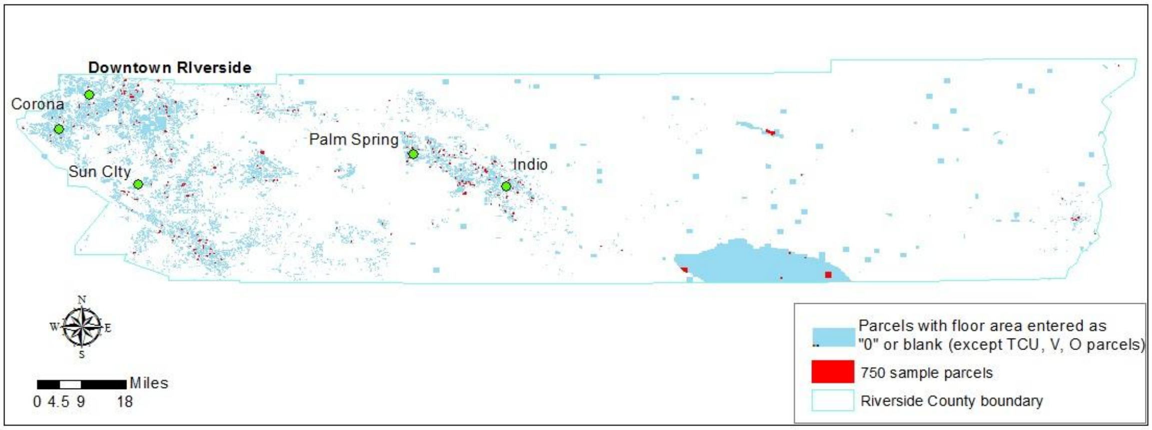

Figure 1. Geographic distribution of 750 sample parcels in Riverside County.

Table 2 also suggests that parcels with floor area entered as "0" are, indeed, more likely to be vacant than those entered as blank. For the former, only about 38% have actually been developed, while, for the latter, this ratio is about 69%. In addition, sample parcels with floor area entered as "0" have seemingly inferior locations to those entered as blank, i.e., significantly farther away from the CBD, the nearest subcenter, and the nearest freeway[7] (t-test values are 2.14, 1.89, and 4.17, respectively, under the unequal variance assumption).

Table 2. Variable definition, data source, and descriptive statistics (I).

Variables	Definition	Source	Parcels with improvements listed as					
			"0" or blank		"0" only		Blank only	
Sample size			750		195		555	
			Mean	SD	Mean	SD	Mean	SD
DEV	Actually developed or not, yes 1, no 0	Truth tracking	0.61	0.49	0.38	0.49	0.69	0.46
CBD	Distance to CBD (mile)	GIS	96.17	37.24	101.18	38.32	94.42	36.73
SUBCENTER	Distance to the nearest subcenter (mile)	GIS	5.92	6.23	8.62	7.36	4.97	5.47
FREEWAY	Distance to the nearest freeway (mile)	GIS	1.27	1.64	1.70	1.70	1.12	1.59
OCEAN	Distance to ocean (mile)	GIS	52.42	29.07	54.10	32.33	51.83	27.84
ZERO	Whether it is a sample parcel with floor area entered as "0", yes 1, no 0		0.26	0.44	1.00	0.00	0.00	0.00
RES	Whether it is a residential parcel, yes 1, no 0		0.79	0.41	0.95	0.22	0.73	0.44

Note: 1) CBD refers to that for the Greater LA region. 2) SD means standard deviation.

[6] The only peculiarity in the following interpolation is that mixed (M) parcels are taken as non-residential although the truth tracking does not include such parcels.

[7] There is no significant difference in distance to the ocean between these two samples (t-test value 0.87).

3.3. Estimation results

The estimation results are given in Table 3 (Columns 2, 4 and 6: fixed effect models). The coefficients of distance to the CBD and distance to the nearest freeway are significantly negative in all models, which conforms to our expectation. Freeway has the largest effect in magnitude on development status among all accessibility measures. The coefficients of distance to the nearest subcenter are significantly negative in models that do not take into account fixed effects, and not surprisingly, become insignificant in fixed effect models. However, the effect of distance to the ocean, though significant, is contrary to our expectation. We suspect that this variable may actually capture other types of location effects, e.g. Palm Springs.

The negative coefficients of ZERO in all these models validate our previous hypothesis that a parcel with its floor area entered as "0" is, indeed, more likely to be vacant than that entered as blank. The insignificant coefficients of RES across models partly indicate that the coding error in floor area is not specific to a certain land-use type.

Table 3. Estimation results of the development status model in Riverside County.
(Dependent variable: DEV)

Variables	Model I		Model II		Model III	
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	1.736 (0.220)***	2.144 (0.523)***	1.679 (0.277)***	1.892 (0.612)***	0.996 (0.204)***	1.549 (0.474)***
CBD	-0.023 (0.005)***	-0.027 (0.011)**	-0.025 (0.007)***	-0.039 (0.014)***	-0.013 (0.005)***	-0.027 (0.010)***
SUBCENTER	-0.022 (0.009)**	-0.018 (0.015)	-0.021 (0.012)**	-0.010 (0.018)	-0.011 (0.009)*	-0.006 (0.014)
FREEWAY	-0.119 (0.039)***	-0.099 (0.046)**	-0.100 (0.045)**	-0.110 (0.054)**	-0.076 (0.035)**	-0.101 (0.048)**
OCEAN	0.025 (0.006)***	0.033 (0.013)**	0.028 (0.009)***	0.045 (0.017)***	0.015 (0.006)***	0.032 (0.013)***
ZERO	-0.545 (0.116)***	-0.653 (0.126)***	-0.662 (0.137)***	-0.785 (0.149)***	-0.542 (0.117)***	-0.661 (0.128)***
RES	-0.180 (0.129)	-0.081 (0.135)	0.104 (0.138)	0.053 (0.149)	0.121 (0.140)	0.060 (0.130)
			0.361 (0.059)***	0.334 (0.072)***	0.374 (0.064)***	0.190 (0.075)***
City fixed effect		YES		YES		YES
N	750	748	750	750	750	750
Log likelihood	-447.0	-415.7	-1352.2	-1361.3	-1328.3	-1359.3

Note: 1) Standard deviations are in parentheses.
2) * p<0.10, ** p<0.05, *** p<0.01.
3) Model II and Model III use the five nearest neighbors weighting matrix.
4) The estimation results for Models II and III are Bayesian estimates obtained using Gibbs sampling.[8]

[8] Therefore, it makes no sense to compare the log likelihood value of non-spatial models with that of spatial models.

The significant, positive [illegible]’s in all spatial models confirm the existence of spatial autocorrelation in the sample data, which indicates that the estimated coefficients in non-spatial models could be biased. And controlling for the city fixed effect helps reduce the spatial autocorrelation in the data. Therefore, the estimation results of spatial fixed effect models will be used in the following for interpolating development status, although fixed effect models do not help increase the log likelihood value.

The spatial error model rather than the spatial autoregressive model is chosen for interpolation, for two reasons. First, Longhi and Nijkamp (2007) suggest that modeling spatial autocorrelation in the residuals produces, on average, the best (forecasting) results, although this is based on their special case. Second, as discussed below, using the spatial error model for interpolation is relatively convenient in this case.

It is also worth discussing the choice of the spatial weighting matrix. Models II and III use the five nearest neighbors weighting matrix instead of other forms, such as inverse distance. First, since the city fixed effect has been controlled for, spatial autocorrelation, if it exists, may only exist locally. Second, the geographic distribution of 750 sample parcels is much more sparse than that of parcels with the development probability to be interpolated (Figure 1). In this case, a distance-based threshold matrix (e.g., 0.5 mile) may capture nothing for some parcels, which may complicate the estimation. And defining more neighbors is not helpful in capturing locally spatial autocorrelation. Further, suppose that the spatial interaction intensity has a relatively flat curve within the “local” area, then, in this case, using the five nearest neighbors weighting matrix will not cause significant bias.

3.4. Development status interpolation

The estimation results of Model 4 in Table 3 are used for interpolating development probability because, in the spatial error model, $E(u) = E((I - \rho W)^{-1}e) = 0$, [illegible] is an unbiased and efficient estimate for development probability.[9] Note that such an estimate would be not efficient when the dependent variable is continuous. In this case, the efficiency could be improved by incorporating the estimated error information of observations.[10]

---

[9] In the spatial autoregressive model, $E(\rho W y) \neq 0$, so $\Phi(x\hat{\beta})$ is a biased estimate. In this case, an iterative process should be used to interpolate the development probabilities.

[10] For the discrete choice model, the latent variable is not observed so there is no way to estimate the error terms.

As mentioned before, since the calculated probability for each parcel characterizes a 0-1 Bernoulli distribution, one realization is drawn from each distribution, to categorize each parcel as developed or not. Then, we compare the calculated probability with a draw from [0,1] uniform distribution, and set the development status as 1 if the former is greater than the latter, otherwise 0.

The descriptive statistics for development status using only one realization are given in Table 4. Of all parcels (except TCU, V, and O parcels), about 65% are identified as having been developed, using the above procedure. This interpolated development ratio is lower for parcels with floor area entered as "0" than that for those entered as blank.

Table 4. Descriptive statistics for the interpolated development status.

Riverside County	Floor area entered as	Sample size	Mean
All parcels (except TCU, V, O parcels)	"0" or blank	112,585	0.65
	"0"	28,274	0.41
	Blank	84,311	0.73
Residential parcels (RS-SF, RS-MF, RS-MX)	"0"	26,097	0.41
	Blank	61,753	0.72
Non-residential parcels (OF, RF, OC, P, W, I, M)	"0"	2,177	0.44
	Blank	22,558	0.75

4. Interpolating the floor area

This section addresses the FAR model specification, data and estimation results, as well as how the estimated model is used to interpolate the floor area for those parcels identified as "developed", using the above procedure. The office parcels (relatively small sample size) and the single-family residential parcels (large sample size) are used as examples, to illustrate the interpolation process. The estimation results for other land-use types are reported in Appendix B.

4.1. Model

The natural logarithm of FAR is used as the dependent variable, to avoid the negative forecast of FAR. Following the urban economic theory, the lot size (in the form of a natural logarithm) and four accessibility measures enter as explanatory variables in the FAR model.

Two models are estimated.[11] Model I is a non-spatial ordinary least squares (OLS) model as

(6)

where represents a vector of explanatory variables, is a vector of coefficients, and is a

[11] The spatial autoregressive model is not tested here due to the reasons given in Section 3.

vector of normally *iid* error terms.

Model II is a spatial error model as

(7)

where  measures the spatial interaction in error terms,  is the spatial weighting matrix, and is a vector of normally *iid* disturbances.

Similarly, for each model, two specifications are tested: a basic one and a fixed effect one. The FAR models are estimated for parcels of each land-use type, respectively. However, sometimes two or more land-use types may be combined when there are only a few parcels of a certain land-use type. In this case, dummy variables indicating the land-use type are introduced into the FAR model.

Now a distance-based threshold matrix (0.5 mile) is used in the FAR model, since, generally, there are many more observations, compared with the development status model. Also, assuming that spatial interaction works in a fixed-size geographic area is more acceptable than defining an equal number of neighbors for all parcels, especially when the geographic distribution of parcels with the same land-use type is not uniform within the whole county.

4.2. Data

Only parcels with a positive floor area, as entered in the database, are used for FAR model estimation. Descriptive statistics are given for two samples in Table 5. There are 1,665 office parcels with positive floor area in Riverside County. An "average" office parcel has a FAR of 0.20, less than 6,000 m^2 lot size, is about 85 miles away from the CBD, 5.60 miles from the nearest subcenter, 1.22 miles from the nearest freeway, and about 43 miles from the ocean. For single-family residential parcels, the sample size is much larger, and the average lot size is less than that of office parcels.

Table 5. Variable definition, data source, and descriptive statistics (II).

Variables	Definition	Source	Parcels with positive floor area			
			Office		Single-family residential	
Sample size			1,665		417,846	
			Mean	SD	Mean	SD
FAR	Floor-to-area ratio		0.20	0.16	0.26	0.12
LOTSIZE	Lot size (1,000 m^2)		5738.74	19247.98	1009.11	1639.44
CBD	Distance to CBD (mile)	GIS	84.92	28.28	82.18	25.85
SUBCENTER	Distance to the nearest subcenter (mile)	GIS	5.60	4.16	5.89	4.19
FREEWAY	Distance to the nearest	GIS	1.22	1.20	1.22	0.97

	freeway (mile)					
OCEAN	Distance to ocean (mile)	GIS	42.58	21.63	41.49	19.35

Note: FAR is calculated as impsqft*0.092951/shape_area, where impsqft is the floor area in sq. ft. and shape_area is the lot size in m^2, as provided in the database.

4.3. Estimation results

The estimation results for office parcels are presented in Table 6. The coefficients of the logarithm of lot size are significantly negative in all specifications. The 1% increase in lot size is associated with a decrease in FAR by about 1%. The distance to the nearest subcenter has negative effects on the FAR, which conforms to our expectation, and the effect is smaller in spatial models than in non-spatial models. However, the coefficients of distance to CBD are significantly positive in the spatial models, possibly indicating the job suburbanization trend in the U.S. The coefficients of the distance to the nearest freeway are significantly positive in the spatial models, due to the negative effects of freeways, such as noise.[12] The coefficients of the distance to the ocean now are significantly negative in the spatial models.

Table 6, Estimation results of the FAR model for office parcels (Dependent variable: ln(FAR)).

Variables	Model I		Model II	
	(1)	(2)	(3)	(4)
Constant	-1.856 (0.158)***	-2.350 (0.574)***	-1.827 (0.060)***	-1.733 (0.182)**
ln(LOTSIZE)	-0.886 (0.018)***	-0.951 (0.027)***	-1.018 (0.009)***	-1.015 (0.009)***
CBD	0.005 (0.002)	-0.001 (0.005)	0.008 (0.001)***	0.009 (0.002)***
SUBCENTER	-0.040 (0.006)***	-0.030 (0.011)***	-0.016 (0.004)***	-0.017 (0.005)***
FREEWAY	0.142 (0.021)***	0.029 (0.034)	0.080 (0.012)***	0.087 (0.012)***
OCEAN	-0.009 (0.004)***	0.002 (0.008)	-0.012 (0.002)***	-0.015 (0.003)***
			0.313 (0.008)***	0.269 (0.008)***
City fixed effect		YES		YES
N	1,665			
Adjusted R^2/ Log likelihood	0.922	0.938	-296.0	-230.7

Note: 1) Standard deviations are in parentheses.
2) * p<0.10, ** p<0.05, *** p<0.01.
3) The estimation results for Model II are maximum likelihood estimates.

The estimation results for single-family residential parcels are similar to those for Office parcels

[12] A quadric form of distance to the nearest freeway will later be added, to test the positive effect of freeways in improving accessibility.

(Table 7). Now the 1% increase in lot size is associated with a 0.8% decrease in FAR. Again, the significantly positive coefficients of distance to CBD are consistent with the suburbanization trend in the U.S. The coefficients of distance to the nearest subcenter are negative, though not significant.

Table 7. Estimation results of the FAR model for single-family residential parcels in Riverside County (Dependent variable: ln(FAR)).

Variables	Model I		Model II	
	(1)	(2)	(3)	(4)
Constant	-1.895 (0.008)***	-1.897 (0.022)***	-1.972 (0.052)***	-2.083 (0.141)***
ln(LOTSIZE)	-0.776 (0.003)***	-0.799 (0.003)***	-0.768 (0.003)***	-0.768 (0.003)***
CBD	0.008 (0.000)***	0.008 (0.000)***	0.008 (0.001)***	0.011 (0.001)***
SUBCENTER	-0.006 (0.000)***	-0.004 (0.001)***	-0.004 (0.003)	-0.007 (0.004)
FREEWAY	0.046 (0.002)***	0.066 (0.002)***	0.098 (0.005)***	0.102 (0.008)***
OCEAN	-0.009 (0.000)***	-0.013 (0.000)***	-0.010 (0.001)***	-0.015 (0.002)***
ρ			0.889 (0.004)***	0.881 (0.006)***
City fixed effect		YES		YES
N	41,785		41,785	
Adjusted R^2/ *Log likelihood*	0.578	0.617	-8689.6	-8891.0

Note: 1) standard deviations are in parentheses.
2) * p<0.10, ** p<0.05, *** p<0.01.
3) The estimation results for Model II are maximum likelihood estimates.

Due to the difficulty in calculating the spatial weighting matrix for a large sample size, the estimation results for only a subsample (10% of the whole sample) are presented here in Table 7. There are three possible ways to improve the current results: 1) by using bootstrap techniques (in which case we need to figure out how to deal with the difference in the spatial relationship between the subsample and the whole sample); 2) by using a more powerful machine; and 3) by using other programs such as SAS.

The significant, positive ρ’s confirm the existence of spatial autocorrelation, and similarly, controlling for the city fixed effect helps reduce the spatial autocorrelation among observations. Thus, the spatial fixed effect model is used for interpolating the floor area.

4.4. Interpolation

The estimation results of Model 4 in Table 7 and those in Table B.1 are used for interpolating the floor area for those "developed" parcels identified in the first step. Again, an implicit assumption here is that the expected FAR of a parcel conditional on being identified as developed, but with floor area entered as "0" or blank, is the same as that of a parcel conditional on having a positive floor area entered into the database.

At this stage, a procedure that could improve the efficiency in the floor area estimates, as discussed in Section 3, is lacking in the interpolation. This will be taken into account in the next step.

The descriptive statistics for the new imputed floor area are presented in Table 8. For comparison, we also list those for the original floor area in the database. The average new floor area is lower than the average original one, since the floor area data is updated for only parcels that are supposed to have "inferior" locations.

Table 8. Descriptive statistics for the floor area (sq. feet).

Riverside County	Sample size	Mean	SD
New floor area	698,072	1775.2	1159.9
Original floor area	613,761	1785.7	1179.4

5. "Robustness" check

As mentioned above, since only one realization is drawn from a 0-1 Bernoulli distribution for each parcel to be categorized as developed or not, this may affect the stability of the imputed aggregate floor area at the model zone level. However, the preliminary analysis shows that the aggregated floor area at the model zone level does not change much when the development status assignment process is repeated, which justifies the procedure adopted in estimating the floor area for parcels with the floor area entered as "0" or blank.

Figure 2 presents five different floor area estimates, by drawing five realizations from the 0-1 development status distribution for each parcel with its floor area entered as "0" or blank, as well as the expected floor area estimate at the model zone level. The imputed aggregate floor area estimates at each model zone are higher than the original ones as expected. Also, there is no much variation in the aggregate floor area across these tests.

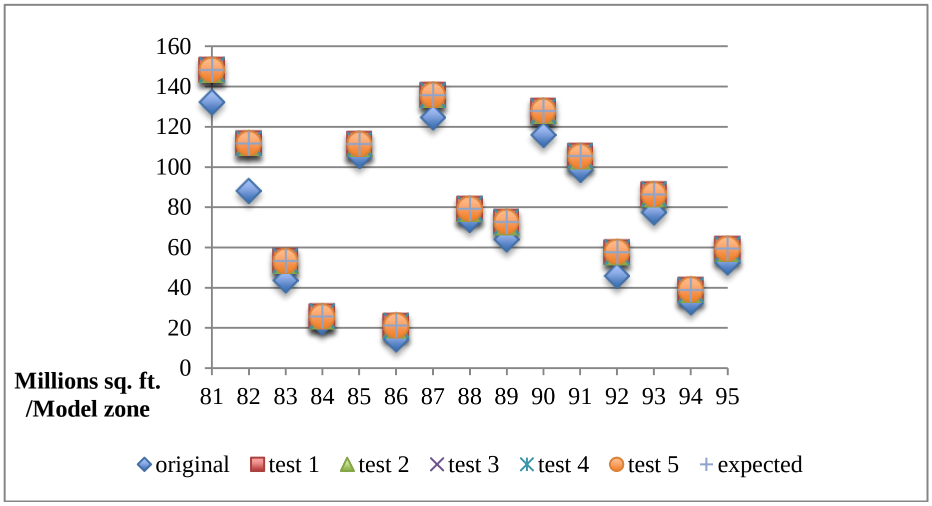

Figure 2. Aggregate floor area at the model zone level in Riverside County (millions sq. ft.).
Note: The "original" refers to that aggregated over the original floor area data, tests 1 to 5 refer to that obtained through repeating the development status assignment process, and the "expected" refers to that calculated using the interpolated floor area multiplied by the development probability.

Figure 3 shows the deviation level of the aggregate floor area from the expected one at the model zone level for all five tests. The maximum deviation level is less than 0.5 per cent for all tests. And the deviation level decreases with the magnitude of the total floor area of the model zone, which indicates that the aggregation procedure (from parcels to model zone) helps reduce the variation.

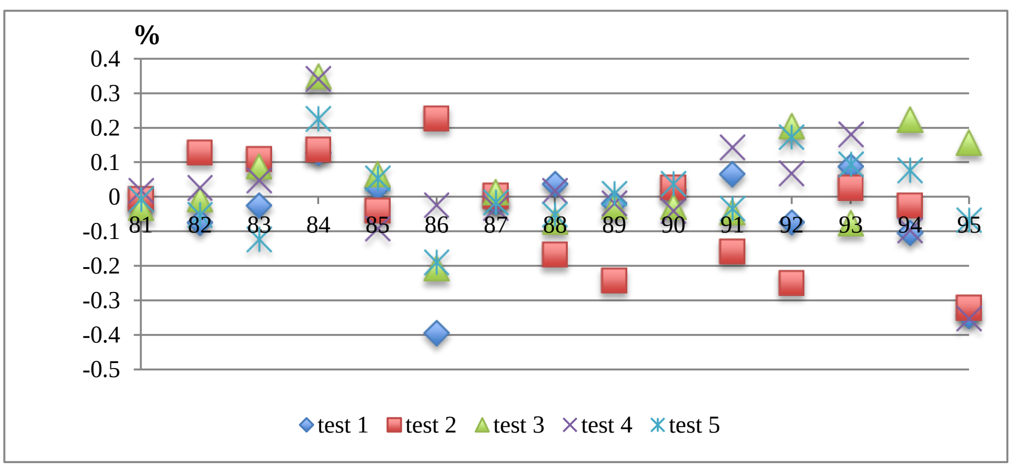

Figure 3. The deviation of the aggregate floor area from the expected one in Riverside County (%).
Note: It is calculated as, (the floor area estimate-the expected floor area)/the expected floor area*100 (%), for each model zone.

6. Conclusions

A two-step procedure is used for adjusting the floor area data in the parcel database of the Greater

Los Angeles region. First, by estimating a binary development status model using some randomly chosen parcels for which the actual development status is obtained through ground truth tracking, parcels that have a floor area entered as "0" or blank are probabilistically categorized as actually being developed or not. Second, for parcels "identified" as being developed, a floor area is imputed using the estimation results of a floor-to-area ratio (FAR) model.

Although there is concern about the stability of the floor area estimates obtained using the above procedure, the analysis shows that there is only negligible variation in the aggregate floor area at the model zone level across five tests. This justifies the procedure used for estimating the floor area data.

Reference

LeSage, J., Pace, R.K. 2009. Introduction to Spatial Econometrics. CRC Press, Boca Raton, FL.

Longhi, S., Nijkamp, P. 2007. Forecasting regional labor market developments under spatial autocorrelation. *International Regional Science Review* 30(2), 100-119.

Appendix A

Table A.1. The correspondence between the LA project land use code and the SCAG 1993 code.

LA project land use code			SCAG 1993 code		
Numerical	Alphabetic	Name	3-digit	Name	Note
1	RS-SF	Single-family residential	111	Single-family residential	
2	RS-MF	Multi-family residential	112	Multi-family residential	
3	RS-MX	Mixed residential	110	Residential	For parcels lacking more detailed classification
			113	Mobile homes and trailer parks	
			114	Mixed residential	
			115	Rural residential	
4	OF	Office	121	General office use	
5	RT	Retail	122	Retail stores and commercial stores	
6	OC	Other commercial	120	Commercial and Services	For parcels lacking more detailed classification
			123	Other commercial	
7	P	Public	124	Public facilities	
			125	Special use facilities	
			126	Educational institutions	
			127	Military installations	
8	W	Warehousing	134	Wholesaling and warehousing	
9	I	Industrial	130	Industrial	For parcels lacking more detailed classification
			131	Light industrial	
			132	Heavy industrial	
			133	Extraction	
10	TCU	Transportation/communication/utilities	14x	Transportation, communications, and utilities	Including 140, 141, 142, 143, 144, 145, 146
11	M	Mixed	150	Mixed commercial and industrial	
			160	Mixed urban	
12	V	Vacant	170	Under construction	
			18x	Open space and recreation	Including 180, 181, 182, 183, 184, 185, 186, 187, 188
			2x0	Agriculture	Including 200, 210, 220, 230, 240, 250, 260, 270
			3x0	Vacant	Including 300, 310, 320
13	O	Other	4x0	Water	Including 400, 410, 420, 430, 440, 450
			0, 1900, 1280, 1290, 8888, 9999, or missing		Not included in SCAG code

Note: The original SCAG 1993 code is a four-digit code. Here the three-digit SCAG code drops off the last digit of the original code.

Appendix B

The estimation results of the FAR model for various land-use types in Riverside County are given in Table B.1. Only those of the spatial error model with fixed effects controlled for are reported here, since these are used for interpolating the floor area for parcels identified as developed. Again, a distance-based threshold matrix (0.5 mile) is used in the model.

In addition, due to the small sample size of mixed parcels in Riverside County, they are combined with other commercial parcels. Thus, a dummy variable "Mix", i.e., whether it is a mixed parcel, is added into the model. Similarly, industrial parcels and warehousing parcels are combined and a dummy variable "WARE", i.e., whether it is a Warehousing parcel, is added.

Table B.1. Estimation results of the FAR model for various land-use types in Riverside County. (Dependent variable: ln(FAR))

Variables	Multi-family residential	Mixed residential	Retail	Other commercial & Mixed	Public	Warehousing & Industrial
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	-3.222 (0.153)***	-1.806 (0.100)***	-1.962 (0.226)***	-1.999 (0.679)***	-1.885 (0.227)***	-1.614 (0.163)***
ln(LOTSIZE)	-0.714 (0.003)***	-0.876 (0.003)***	-0.993 (0.013)***	-0.941 (0.028)***	-0.995 (0.016)***	-0.992 (0.010)***
CBD	0.016 (0.004)***	0.007 (0.001)***	0.004 (0.002)*	-0.005 (0.007)	0.001 (0.003)	0.006 (0.002)***
SUBCENTER	-0.009 (0.006)	-0.020 (0.001)***	-0.014 (0.006)***	-0.029 (0.011)**	-0.015 (0.007)**	-0.023 (0.005)***
FREEWAY	0.246 (0.010)***	0.042 (0.004)***	0.086 (0.014)***	0.007 (0.039)	0.059 (0.019)***	0.071 (0.010)***
OCEAN	-0.007 (0.005)	-0.009 (0.001)***	-0.009 (0.004)**	0.007 (0.009)	-0.003 (0.004)	-0.012 (0.002)***
MIX				-0.182 (0.120)		
WARE						()
ρ	0.906 (0.002)***	0.752 (0.007)***	0.216 (0.023)***	0.198 (0.003)**	0.079 (0.066)***	0.205 (0.030)***
City fixed effect	YES					
N	23,042	45,471	1,118	328	669	1,864
Log likelihood	-10213.9	-2316.9	-183.3	-45.9	-156.9	-359.2

Note: 1) Standard deviations are in parentheses.
2) * p<0.10, ** p<0.05, *** p<0.01.
3) The maximum likelihood estimates are presented.

Appendix C

C.1 San Bernardino County

As mentioned earlier, the estimated development status model without city fixed effect for Riverside County (Model 3 in Table 3) is used for interpolating development probability, since no ground truth tracking has been done for San Bernardino County.

Table C.1 gives the estimation results of the FAR model for various land-use types in San Bernardino County, which is used for interpolating the floor area. Similarly, only those of the fixed-effect spatial error model, using a distance-based threshold matrix (0.5 mile), are reported. In addition, industrial parcels and warehousing parcels are combined, and a dummy variable "WARE" is added.

Table C.1. Estimation results of the FAR model for various land-use types in San Bernardino County. (Dependent variable: ln(FAR))

Variables	Single-family residential	Multi-family residential	Mixed residential	Office	Retail
	(1)	(2)	(3)	(4)	(5)
Constant	-1.563 (0.222)***	-1.477 (0.066)***	-2.171 (0.020)***	-1.823 (0.840)**	-0.934 (0.803)
ln(LOTSIZE)	-0.826 (0.003)***	-0.501 (0.002)***	-0.902 (0.003)***	-0.304 (0.015)***	-0.348 (0.014)***
CBD	-0.005 (0.003)	0.008 (0.004)**	-0.016 (0.001)***	-0.044 (0.006)***	-0.023 (0.009)**
SUBCENTER	0.006 (0.004)*	-0.014 (0.004)***	-0.017 (0.001)***	-0.065 (0.006)***	-0.032 (0.008)***
FREEWAY	0.021 (0.045)***	0.068 (0.007)***	0.014 (0.001)***	0.077 (0.014)***	0.039 (0.008)***
OCEAN	-0.002 (0.005)	-0.005 (0.005)	0.023 (0.001)***	0.076 (0.007)***	0.037 (0.013)**
ρ	0.881 (0.008)***	0.710 (0.005)***	0.437 (0.004)***	0.301 (0.018)***	0.360 (0.012)***
City fixed effect	YES				
N	42,326	34,240	28,460	2,711	4,618
Log likelihood	-5174.2	-4240.8	-7147.7	-2758.6	-4749.6
Variables	Other commercial	Public	Warehousing & Industrial	Mixed	
	(6)	(7)	(8)	(9)	
Constant	-1.419 (0.501)***	-2.130 (1.173)*	-0.691 (0.351)**	-7.507 (0.752)***	
ln(LOTSIZE)	-0.226 (0.034)***	-0.578 (0.016)***	-0.508 (0.016)***	-0.076 (0.043)*	
CBD	-0.004 (0.022)	0.010 (0.011)	-0.019 (0.009)**	0.055 (0.031)*	
SUBCENTER	-0.051 (0.021)**	-0.013 (0.013)	-0.046 (0.010)***	-0.097 (0.044)**	
FREEWAY	0.075 (0.020)***	0.011 (0.013)	0.034 (0.010)***	-0.031 (0.010)	

OCEAN	0.024 (0.036)	-0.012 (0.020)	0.041 (0.015)***	0.075 (0.055)	
WARE			1.484 (0.147)***		
ρ	0.133 (0.007)**	0.228 (0.009)***	0.228 (0.007)***	0.226 (0.019)***	
City fixed effect	YES				
N	717	1,771	3,466	564	
Log likelihood	-800.0	-1490.5	-3770.2	-232.4	

Note: 1) Standard deviations are in parentheses.
2) * p<0.10, ** p<0.05, *** p<0.01.
3) The maximum likelihood estimates are presented.
4) For single-family residential parcels, the model is estimated using a 10% sample.

C.2 Orange County

Table C.2 gives the estimation results of the development status model using 849 randomly chosen parcels in Orange County for which the actual development status is obtained through ground truth tracking. Only those of spatial error model, using a 5 nearest neighbor weighting matrix, are reported here. Two land-use type dummies are added: "COM", whether it is a commercial parcel (SCAG code: lu08=12x in the parcel database); and "IND", whether it is an industrial parcel (SCAG code: lu08=13x in the parcel database). Model 2 (with city fixed effect) in Table C.2 is used for interpolating the development probability for parcels that have a floor area entered as "0" or blank in Orange County.

Table C.2. Estimation results of the development status model in Orange County.
(Dependent variable: DEV)

Variables	Model I	
	(1)	(2)
Constant	1.617 (0.313)***	3.633 (1.114)***
CBD	-0.025 (0.007)***	-0.057 (0.024)***
SUBCENTER	0.053 (0.025)**	0.105 (0.057)**
FREEWAY	0.173 (0.071)***	0.212 (0.093)**
OCEAN	-0.009 (0.011)	0.008 (0.036)
ZERO	-0.087 (0.118)	-0.142 (0.131)
COM	-0.507 (0.135)***	-0.556 (0.144)***
IND	-0.337 (0.210)*	-0.305 (0.224)*
ρ	0.127 (0.059)**	0.118 (0.079)*
City fixed		YES

effect		
N	849	849
Log likelihood	-1526.5	-1561.6

Note: 1) Standard deviations are in parentheses.
2) * p<0.10, ** p<0.05, *** p<0.01.
3) The Bayesian estimates obtained using Gibbs sampling are presented.

Table C.3 gives the estimation results of the FAR model for various land-use types in Orange County, which is used for interpolating the floor area. Only those of the fixed-effect spatial error model, using a distance-based threshold matrix (0.5 mile), are reported. In addition, industrial parcels and warehousing parcels are combined, and a dummy variable "WARE" is added.

Table C.3. Estimation results of the FAR model for various land-use types in Orange County. (Dependent variable: ln(FAR))

Variables	Single-family residential	Multi-family residential	Mixed residential	Office	Retail
	(1)	(2)	(3)	(4)	(5)
Constant	-2.312 (0.088)***	-0.517 (0.218)**	-0.657 (0.964)	-1.026 (0.954)	-2.118 (0.139)***
ln(LOTSIZE)	-0.747 (0.004)***	-0.404 (0.004)***	-0.688 (0.013)***	-0.122 (0.009)***	-0.185 (0.009)***
CBD	0.007 (0.001)***	-0.017 (0.004)***	-0.047 (0.003)***	0.012 (0.009)	0.013 (0.005)**
SUBCENTER	0.005 (0.006)	0.003 (0.007)	0.009 (0.011)	-0.002 (0.012)	-0.020 (0.013)
FREEWAY	0.051 (0.009)***	0.017 (0.013)	0.018 (0.001)	-0.158 (0.025)***	-0.049 (0.023)**
OCEAN	0.009 (0.003)***	0.005 (0.005)	0.009 (0.008)	-0.010 (0.010)	-0.019 (0.009)**
ρ	0.891 (0.004)***	0.649 (0.012)***	0.508 (0.006)***	0.279 (0.048)***	0.374 (0.003)***
City fixed effect	YES				
N	40,869	28,376	6,681	5,563	9,824
Log likelihood	-8234.2	-9999.8	-2849.2	-4148.6	-8261.2
Variables	Other commercial	Public	Warehousing & Industrial	Mixed	
	(6)	(7)	(8)	(9)	
Constant	1.923 (2.287)	0.221 (0.809)	-2.350 (0.172)***	-2.062 (0.430)***	
ln(LOTSIZE)	-0.230 (0.031)***	-0.289 (0.016)***	-0.168 (0.007)***	-0.804 (0.015)***	
CBD	-0.011 (0.038)	-0.014 (0.014)	0.035 (0.008)***	0.019 (0.011)*	
SUBCENTER	-0.145 (0.049)***	0.036 (0.023)	-0.073 (0.014)***	-0.037 (0.029)	
FREEWAY	-0.288 (0.097)***	-0.144 (0.038)***	-0.019 (0.028)	-0.176 (0.036)***	
OCEAN	0.021 (0.034)	-0.029 (0.017)*	-0.045 (0.011)***	-0.011 (0.016)	
WARE			0.242		

			(0.013)***		
ρ	0.078 (0.005)**	0.201 (0.031)***	0.585 (0.002)***	0.444 (0.011)***	
City fixed effect	YES				
N	649	2,110	8,356	2,376	
Log likelihood	-628.0	-1829.7	-5055.9	-2238.0	

Note: 1) Standard deviations are in parentheses.

2) * p<0.10, ** p<0.05, *** p<0.01.

3) The maximum likelihood estimates are presented.

4) For single-family residential parcels and multi-family residential parcels, the model is estimated using a 10% sample and a 50% sample, respectively.

C.3 Los Angeles County

The estimated development status model without city fixed effect for Orange County (Model 1 in Table C.2) is used for interpolating development probability, since no ground truth tracking has been done for Los Angeles County.

Table C.4 gives the estimation results of the FAR model for various land-use types in Los Angeles County, which is used for interpolating the floor area. Only those of the fixed-effect spatial error model, using a distance-based threshold matrix (0.5 mile), are reported.

Table C.4. Estimation results of the FAR model for various land-use types in Los Angeles County.
(Dependent variable: ln(FAR))

Variables	Single-family residential	Multi-family residential	Mixed residential	Office	Retail
	(1)	(2)	(3)	(4)	(5)
Constant	-1.369 (0.074)***	-1.736 (0.110)***	0.189 (0.067)***	-2.259 (0.550)***	-2.955 (0.131)***
ln(LOTSIZE)	-0.746 (0.004)***	-1.129 (0.004)***	-0.645 (0.006)***	-1.141 (0.005)***	-0.836 (0.005)***
CBD	-0.008 (0.001)***	-0.020 (0.002)***	-0.007 (0.002)***	0.002 (0.001)*	-0.006 (0.002)**
SUBCENTER	-0.007 (0.002)***	0.006 (0.004)	-0.029 (0.003)***	0.009 (0.004)**	-0.016 (0.004)***
FREEWAY	0.000 (0.001)	0.009 (0.002)***	0.002 (0.001)*	0.007 (0.002)***	0.001 (0.002)
OCEAN	0.062 (0.006)***	0.035 (0.012)***	0.039 (0.006)***	0.004 (0.011)	-0.066 (0.012)***
ρ	0.714 (0.007)***	0.334 (0.010)***	0.490 (0.002)***	0.198 (0.023)***	0.228 (0.006)***
City fixed effect	YES				
N	42,088	27,804	32,075	36,485	43,178
Log likelihood	-790.8	-22056.0	-14792.3	43435.16	-51207.3
Variables	Other commercial	Public	Warehousing	Industrial	Mixed
	(6)	(7)	(8)	(9)	(10)
Constant	2.368 (3.665)	-0.691 (0.554)	-0.843 (1.430)	-1.946 (0.330)***	-1.153 (0.031)***

ln(LOTSIZE)	-0.767 (0.016)***	-0.780 (0.007)***	-0.095 (0.016)***	-0.679 (0.006)***	-1.005 (0.004)***
CBD	-0.016 (0.009)*	-0.006 (0.004)*	0.012 (0.012)	-0.039 (0.005)***	-0.020 (0.002)***
SUBCENTER	-0.110 (0.013)***	-0.047 (0.005)***	-0.030 (0.016)*	-0.036 (0.008)***	-0.006 (0.002)***
FREEWAY	0.083 (0.008)***	0.009 (0.003)***	-0.022 (0.012)*	0.021 (0.005)***	0.011 (0.001)***
OCEAN	0.073 (0.042)*	0.126 (0.018)***	-0.219 (0.063)***	0.120 (0.027)***	-0.007 (0.004)*
ρ	0.231 (0.009)**	0.181 (0.005)***	0.472 (0.007)***	0.632 (0.006)***	0.326 (0.006)***
City fixed effect	YES				
N	5,088	22,755	1,793	26,103	37,146
Log likelihood	-6792.8	-29990.5	-1339.0	-31339.4	-30681.2

Note: 1) Standard deviations are in parentheses.
2) * p<0.10, ** p<0.05, *** p<0.01.
3) The maximum likelihood estimates are presented.
4) For Single-family residential, multi-family residential, mixed residential, office, retail and industrial parcels, the model is estimated using a 3%, 10%, 20%, 50%, 50%, and 50% sample, respectively.

C.4 Ventura County

The estimated development status model without city fixed effect for Orange County (Model 1 in Table C.2) is used for interpolating development probability, since no ground truth tracking has been done for Ventura County.

Table C.5 gives the estimation results of the FAR model for various land-use types in Ventura County, which is used for interpolating the floor area. Only those of the fixed-effect spatial error model, using a distance-based threshold matrix (0.5 mile), are reported.

Retail, other commercial, and mixed parcels are combined, with two dummy variables "OC" (whether it is an other commercial parcel) and "MIX" (whether it is a mixed parcel) added. Also, industrial parcels and warehousing parcels are combined, and a dummy variable "WARE" is added.

Table C.5. Estimation results of the FAR model for various land-use types in Ventura County. (Dependent variable: ln(FAR))

Variables	Single-family residential	Multi-family residential	Mixed residential	Office
	(1)	(2)	(3)	(4)
Constant	-1.547 (0.189)***	-1.468 (0.041)***	-0.939 (0.151)***	-0.419 (0.261)
ln(LOTSIZE)	-0.801 (0.003)***	-0.659 (0.002)***	-0.782 (0.011)***	-0.585 (0.025)***
CBD	-0.002 (0.004)	0.012 (0.002)***	-0.006 (0.003)**	-0.020 (0.005)***
SUBCENTER	-0.025	-0.020	-0.012	-0.011

	(0.007)***	(0.003)***	(0.006)**	(0.012)
FREEWAY	0.134 (0.045)***	0.015 (0.007)**	0.040 (0.018)**	0.046 (0.048)
OCEAN	0.015 (0.005)**	-0.010 (0.003)***	-0.012 (0.005)***	-0.052 (0.012)***
ρ	0.936 (0.005)***	0.698 (0.003)***	0.610 (0.026)***	0.065 (0.069)
City fixed effect	YES			
N	33,293	28,419	3,608	601
Log likelihood	-6861.6	-5589.1	-546.8	-245.9

Variables	Retail & Other commercial & Mixed	Public	Warehousing & Industrial	
	(5)	(6)	(7)	
Constant	-0.310 (0.540)	-1.359 (0.460)***	-1.734 (0.219)***	
ln(LOTSIZE)	-0.704 (0.028)***	-0.761 (0.021)***	-0.803 (0.024)***	
CBD	0.008 (0.007)	-0.012 (0.005)**	-0.010 (0.004)***	
SUBCENTER	0.004 (0.017)	0.019 (0.016)	0.003 (0.09)	
FREEWAY	0.107 (0.070)	0.044 (0.038)	0.133 (0.040)***	
OCEAN	-0.005 (0.016)	-0.021 (0.010)**	-0.014 (0.009)	
OC	0.168 (0.081)**			
MIX	-0.173 (0.223)			
WARE			1.484 (0.147)***	
ρ	0.336 (0.058)***	0.232 (0.013)***	0.200 (0.015)***	
City fixed effect	YES			
N	523	514	516	
Log likelihood	-179.8	-15.7	-43.9	

Note: 1) Standard deviations are in parentheses.
2) * p<0.10, ** p<0.05, *** p<0.01.
3) The maximum likelihood estimates are presented.
4) For Single-family residential parcels, the model is estimated using a 20% sample.