Regression Promises and Aggregation Bias Illusions

The Application of Market Delineation to Land Valuation Models

by Matthew C. Trimble, MAI

Abstract

Regression is one of the best tools for consistently deriving market-based adjustments in the appraisal of real estate. There are limitations in regression, however, and the potential for misleading results must be recognized. A principal violation of the validity of a regression model is aggregation bias, which has received limited attention in appraisal literature but is discussed here. This article shows how aggregation bias may creep into a regression model, and how professional appraisers are equipped to avoid it with the tools of market delineation and segmentation. There is a pervasive misunderstanding that a large data sample will minimize the negative impact of inappropriate or incorrect data points (comparables). In truth, the quality of data is as important in large regression modeling data sets as it is in small data sets in the conventional sales comparison approach. This article offers a case study of vacant industrial land to illustrate the misleading results of over aggregation (aggregation bias) and demonstrates how aggregation bias can be avoided through market delineation and segmentation. Only after a data set has been delineated and segmented in accordance with the market can issues related to model specification be effectively addressed.

Introduction

Appraisers can readily measure distance to highways, employment, and other linkages with Google Earth and various GIS applications. Demographic, employment, and income data is available for nearly every locality. Sale and transactional data as well as parcel-specific physical data including flood, parcel, topographical, soil, zoning, and utility maps are generally available online. Regression provides a tool for processing large data sets and extracting adjustments in a consistent manner for use in the sales comparison approach. Limitations of regression include insufficient data availability for unique or nonquantifiable property features and over aggregation (aggregation bias). At a minimum, regression (as well as paired sales) requires a sample size sufficiently larger than the number of predictor (independent) variables included in the model. A fundamental assumption of the underlying aggregated data used in regression modeling is that the modeled relationship between the economic variables is homogeneous across all market participants.[1] As the behaviors of economic agents across distinct real estate markets are not the same, data aggregated over different markets can produce misleading results and an invalid regression model. Misleading regression results due to aggregation bias in real estate appraising can be addressed through market delineation and segmentation, which ensure data selection is representative of the market for the parcel or parcels being appraised.

Real estate sales data is typically classified as a nonprobability sample.[2] The two most funda-

1. Thomas A. Garrett, "Aggregated versus Disaggregated Data in Regression Analysis: Implications for Inference," *Economics Letters* 81, no. 1 (2003): 61–65.

2. *The Appraisal of Real Estate*, 15th ed. (Chicago: Appraisal Institute, 2020), 253.

mental assumptions with regression in appraising are validity and representativeness,[3] which require an appraiser's professional judgment. Validity, the most fundamental assumption in valuation modeling, is the assumption that the regression model describes a real-world relationship.[4] The application of regression in real estate appraising should not run contrary to standard market delineation and segmentation practices of market analysis. The following case study of industrial land valuation using regression will demonstrate the misleading effects of aggregation bias, how aggregation bias can be avoided, and the critical role that market delineation and segmentation play in producing a credible and valid regression model.

Case Study Example

Market Delineation and Segmentation

Market delineation is the process of identifying a specific real estate market. It considers the following factors: property type, property features, market area, available substitute properties, and access to complementary properties.[5] Regression assumes that the modeled relationship between the independent variables (elements of comparison) and the dependent variable (price) is homogeneous across all market participants described by the model; therefore market delineation is a critical step in this assumption. In regression, the goal of market delineation is to identify the competitive market segment,[6] i.e., the set of sales reflective of the market for the appraised property. In some instances, different users may compete for land in a market, and using sales of land acquired for competing uses may be justified for inclusion in the regression model provided the economic behavior underlying competing sales parallels the economic behavior being modeled.

In this case study example, three tracts of land ranging between approximately 30 and 60 acres located in the Southeast Industrial Node of the Oklahoma City metro area are valued.[7] A summary of the pertinent characteristics of the parcels are shown in Exhibit 1.

The Oklahoma City industrial market is characterized by growth and stable demand; it contains three primary industrial areas—the Southwest, the Southeast, and the North. Other smaller industrial areas in the Oklahoma City metro serve as secondary competition to the three major industrial nodes. The North Industrial Node is influenced by a major corridor of newer retail development and the affluent suburban residential areas of north Oklahoma City. The Southwest Industrial Node is concentrated around Will Rogers World Airport, while the Southeast Industrial Node benefits from proximity to Tinker Air Force Base, the largest employer in the state of Oklahoma. Both the Southeast and Southwest Nodes are convenient to middle-income populations, interstate highways, and rail transport and have similar support services.

Initial Data Collection and Regression with Aggregated Data

The initial search for comparable sales included the east, west, central, and southern portions of the Oklahoma City metro area and excluded the North Industrial Node due to demographic and locational differences. Other sales excluded were those with significant building improvements and those from rural type areas. The geographic search boundary is depicted in Exhibit 2. Exhibit 3 shows the twenty-one sales identified in the initial search.

Property features considered for elements of comparison were shape, topography (including drainage and flood), frontage (interior, primary road frontage, dual frontage road), highway expo-

3. *The Appraisal of Real Estate*, 15th ed., 153–154. In order of decreasing importance, the assumptions of regression analysis are validity; representativeness; additivity and linearity; independence of errors; equal variance of errors; and normality of errors. For additional discussion, see *The Appraisal of Real Estate*, 15th ed., appendix B, "Regression Analysis and Statistical Applications," available online at www.appraisalinstitute.org/15th-edition-appendices/, which addresses more complex concepts and considerations in the use of statistical applications like multiple regression analysis.
4. Andrew Gelman, Jennifer Hill, and Aki Vehtari, *Regression and Other Stories* (Cambridge University Press, 2020), 24.
5. *The Appraisal of Real Estate*, 15th ed., 139.
6. George Dell, "Regression, Critical Thinking, and the Valuation Problem Today," *The Appraisal Journal* (Summer 2017): 217–229.
7. While based on actual parcels, some details of the valued parcels were changed or omitted for the purpose of this article.

Exhibit 1 Summary of Subject Parcels

Parcel	A	B	C
Area / Submarket	SE	SE	SE
Zoning	I-2 (Industrial)	I-2 (Industrial)	I-2 (Industrial)
Shape	Irregular / Functional	Irregular / Functional	Rectangular / Functional
Net Acres	32.50	57.65	45.25
Topography	Rolling	Rolling	Level
Frontage	Dual Primary Road	Primary Road	Primary Road
Hwy. Exposure	No	Yes	No
Dist. to Interstate Hwy. (miles)	1	0.1	1.5
Surrounding Dev.	Medium	Medium	Medium

sure, distance to interstate highway system, and surrounding development (quality and density). Excluded variables were zoning, parcel size, and sale date. Only the distance to interstate highway variable was transformed. As zoning changes are relatively common in expansion areas of the Oklahoma City metro area, zoning was not considered a significant value influence. Size adjustments are not a given in real estate but a function of supply and demand. In this market within the size range modeled, large parcel demand from industrial end users offsets the conventional size adjustment. Alternative regression models that included size as a variable indicated size was not economically or statistically significant. All the sales were considered reflective of current market conditions, and no market condition adjustments were indicated. As the value effect of distance to highway logically lessens with each unit increase in distance (nonlinear), the distance to interstate highway variable was transformed using a basic square root function.

A statistically inclined analyst with limited knowledge of real estate market behavior and insufficient geographic familiarity might proceed to input the initial results into a regression model without further market delineation as shown in Exhibit 4. Upon market delineation and segmentation, however, it is revealed that this initial data set contains mixed markets with economic agents that respond differently to various property attributes. In other words, the behaviors of the economic agents used for input into the aggregated model are not homogeneous enough to approximate a market response to industrial land and its economically relevant attributes. As this initial data set suffers from aggregation bias, the most fundamental assumptions relevant to valuation modeling (validity and representativeness) are violated. Consequently, the regression results (Exhibit 5) are misleading and detached from the reality the model is attempting to measure.

Exhibit 2 Geographic Search Boundaries

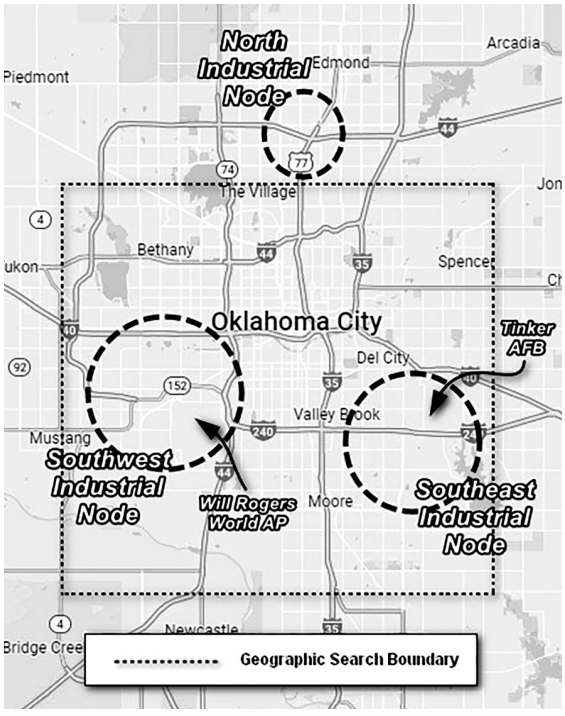

Exhibit 3 Initial Sales Search Results (Aggregated)

Sale	Zoning	Size Acres	$/SF	Shape	Topography	Frontage	Highway Exposure	Distance to Interstate Hwy. System (miles)	Surrounding Development Intensity
1	I-2	22	$1.36	Irregular	Level	Primary road	Yes	0.4	Medium
2	A-2	29	$0.72	Highly irregular	Drainage and pond	Interior	No	0.4	High
3	R1	134	$0.55	Highly irregular	Numerous ponds	Interior	No	2	Medium
4	A-2	57	$1.11	Irregular	Drainage, sloping, ponds	Two primary roads	Yes	0.8	Medium
5	AA	30	$0.52	Irregular	Level	Interior	No	3.25	Low
6	AA	38	$0.42	Irregular due to oil pad near corner	Drainage	Two primary roads	No	2	Low
7	I-3	30	$0.49	Generally rectangular	Level	Interior	No	0.8	Low
8	PUD-1705	74	$0.65	Irregular	Drainage, sloping	Two primary roads	No	0.5	Medium
9	PUD-902	79	$0.52	Generally rectangular	Level	Primary road	No	1.8	Medium
10	AA	38	$0.39	Irregular due to oil pad near corner	Drainage	Two primary roads	No	2	Low
11	C-3, R-1	39	$1.35	Irregular	Level	Two primary roads	Yes	1.5	High
12	I-2	22	$1.07	Generally rectangular	Rolling (small pond drained)	Primary road	No	0.64	High
13	I-2	40	$0.46	Highly irregular	Rolling	Primary road	No	1.2	Medium
14	R-1	17	$0.45	Irregular	Rolling	Primary road	No	2.2	Medium
15	I-2	35	$1.06	Generally rectangular	Level	Primary road	No	1	High
16	I-2	22	$0.88	Highly irregular	Drainage with flood	Two primary roads	Yes	0.25	Medium
17	AA, SPUD-854	26	$1.15	Generally rectangular	Rolling	Two primary roads	No	0	Medium
18	AA	54	$0.43	Irregular	Rolling	Interior	No	2.3	Low
19	I-2	26	$0.45	Irregular	Rolling	Two primary roads	Yes	0	Medium
20	R-1	117	$0.49	Irregular	Numerous ponds	Interior	No	1.5	Medium
21	R-1	52	$0.44	Irregular	Rolling	Primary road	No	1.25	Low

Exhibit 4 Initial Aggregated Model Inputs

Observation	$/SF	Highly Irr. Shape	Rolling Topo.	Extreme Topo. (Drainage, Flood, Other)	Primary Road Frontage	Dual Primary Road Frontage	Hwy. Exp.	Dist. to Int. Hwy. (sq. rt. miles)	Med. Surr. Dev.	High Surr. Dev.
1	$1.36	0	0	0	1	0	1	0.632	1	0
2	$0.72	1	0	1	0	0	0	0.632	0	1
3	$0.55	1	0	1	0	0	0	1.414	1	0
4	$1.11	0	0	1	0	1	1	0.894	1	0
5	$0.52	0	0	0	0	0	0	1.803	0	0
6	$0.42	0	0	1	0	1	0	1.483	1	0
7	$0.49	0	0	0	0	0	0	0.894	0	0
8	$0.65	0	0	1	0	1	0	0.707	1	0
9	$0.52	0	0	0	1	0	0	1.342	1	0
10	$0.39	0	0	1	0	1	0	1.414	0	0
11	$1.35	0	0	0	0	1	1	1.225	0	1
12	$1.07	0	1	0	1	0	0	0.800	0	1
13	$0.46	1	1	0	1	0	0	1.095	1	0
14	$0.45	0	1	0	1	0	0	1.483	1	0
15	$1.06	0	0	0	1	0	0	1.000	0	1
16	$0.88	1	0	1	1	0	1	0.500	1	0
17	$1.15	0	1	0	0	1	0	0.000	1	0
18	$0.43	0	1	0	0	0	0	1.517	0	0
19	$0.45	0	1	0	0	1	1	0.000	1	0
20	$0.49	0	0	1	0	0	0	1.225	1	0
21	$0.44	0	1	0	0	1	0	1.118	0	0

Exhibit 5 Initial Aggregated Regression Results

Regression Statistics						
Multiple *R*	0.816567					
R Square	0.666781					
Adjusted *R* Square	0.394147					
Standard Error	0.258177					
Observations	21					
			ANOVA			
	df	***SS***	***MS***	***F***	***Significance F***	
Regression	9	1.467172	0.163019	2.4457	0.08199	
Residual	11	0.733209	0.066655			
Total	20	2.200381				
	Coefficients	**Standard Error**	***t* Stat**	***P*-value**	**Lower 95%**	**Upper 95%**
Intercept	0.7480	0.2783	2.6876	0.0211	0.1354	1.3606
Highly Irregular Shape	–0.1154	0.1976	–0.5841	0.5709	–0.5503	0.3195
Rolling Topography	–0.1758	0.1747	–1.0062	0.3360	–0.5603	0.2087
Extreme Topography	–0.1260	0.2011	–0.6269	0.5435	–0.5686	0.3165
Primary Road Frontage	0.0772	0.1849	0.4178	0.6841	–0.3297	0.4842
Dual Primary Road Frontage	0.0347	0.1857	0.1867	0.8553	–0.3740	0.4434
Highway Exposure	0.2590	0.1672	1.5486	0.1498	–0.1091	0.6271
Distance to Int. Hwy. (sq. rt. miles)	–0.1574	0.1554	–1.0133	0.3327	–0.4994	0.1845
Medium Surr. Dev.	0.1183	0.1900	0.6226	0.5463	–0.3000	0.5366
High Surr. Dev.	0.4382	0.2098	2.0888	0.0608	–0.0235	0.8998

This article primarily focuses on the fundamental assumptions of validity and representativeness. Those assumptions are violated in the aggregated model due to aggregation bias.[8] The aggregated model produced the following three illusory results:

1. The coefficient for drainage and other extreme topographical conditions is irrational as it indicates less of a deduction for extreme topography than for rolling topography. Industrial users are typically averse to creeks, drainage, significant water features, and other topographical features that increase the development cost of land.
2. Industrial users value efficient parcel access. The coefficients for primary and dual road frontage appear small and irrational. Greater economic significance would be expected when compared to interior parcels.
3. Industrial users value accessibility to the interstate highway system. Whether the users are warehousing or manufacturing, interstate highways are a primary mode of transporting their goods to market. Consequently, the coefficient for distance to interstate highway system appears low, as a coefficient of −0.1574 translates into a price-per-square-foot-decrease of approximately $0.11 for a half-mile distance and decrease of approximately $0.16 for a one-mile distance when compared to immediate access.[9]

Data Collection and Regression Results with Market Delineation and Segmentation

The market behavior described in the aggregated model resulting from the initial data set suffers from aggregation bias, and it does not match the market behavior for industrial land that the model is attempting to describe. The initial data set excluded sales in the North submarket, but that was not sufficient for market delineation, and the applicable market segment was not sufficiently identified and isolated. A more detailed and thorough analysis of the sales was conducted and is shown in Exhibit 6. The data set in Exhibit 6 was segmented in accordance with market delineation revealing seven out of the twenty-one sales were not representative of the relevant industrial land market and should therefore be excluded.

As shown in Exhibit 6, analysis of each individual sale indicates the initial data set was commingled with land sales that have negligible, if any, competitiveness with the industrial land parcels whose value is being modeled. The residential market response to ponds, creeks, access to highways, and other features is not consistent (homogeneous) with the response of the industrial market to those features. Therefore, the assumptions of homogeneity of the economic agents being modeled and the validity of the model are violated by aggregation bias. In addition, a non-arm's-length transaction, an interior oil and gas site, and a property unable to connect to sewer services should be excluded as these unique conditions influencing price are beyond the scope of the model. After excluding seven of the twenty-one sales, the data set is reduced to fourteen sales. However, the remaining fourteen sales represent the relevant market segment. Sacrificing sample size for representativeness and validity is a necessary trade-off appraisers must be willing to make in regression modeling. The market delineated and segmented regression inputs and results are shown in Exhibit 7 and Exhibit 8, respectively.

The regression results after market delineation and segmentation have a high goodness of fit and illustrate coefficient effects consistent with the market behavior for industrial land. The prior misleading results from the aggregated model have been corrected as follows:

1. The coefficient for drainage and other extreme topographical conditions is now rational, and it indicates a greater value loss than rolling topography. This result is consistent with industrial users, which typically are averse to creeks, drainage, significant water features, and other topographical features that increase the development cost of land.
2. Results now indicate industrial users' value of efficient parcel access. The coefficients for dual and primary road frontage increased substantially compared to the aggregated model and are consistent with known market behavior.
3. Results now indicate industrial users' value of accessibility to the interstate highway system. Whether users are warehousing or

8. Garrett, "Aggregated versus Disaggregated," 61–65.

9. $-0.15742 \times \sqrt{0.5} = -0.11131$

Exhibit 6 Market Delineated Data Set

Observation	Market Delineation Comments	Part of Market Segment (Include in Model)
1	SE industrial land sale.	Yes
4	Zoned for agriculture. Zoning change likely. Industrial uses present in the area. Considered competitive.	Yes
7	Industrial land sale.	Yes
8	SW industrial land sale.	Yes
10	SE land sale. Area includes mix of residential and industrial. Considered competitive.	Yes
11	Purchased for self-storage development. Considered secondarily competitive.	Yes
12	SW industrial land sale with rail access.	Yes
13	SE industrial land sale.	Yes
14	SE industrial land sale. Zoned residential but acquired for materials storage.	Yes
15	SE industrial land sale with rail access.	Yes
16	SW industrial land sale.	Yes
17	SE land sale in an area of mixed residential and industrial. Considered competitive.	Yes
20	SE land sale in an area of mixed residential and industrial. Considered competitive.	Yes
21	SE industrial land sale.	Yes
2	Interior back land site acquired by an oil and gas operator reportedly for a pad site. Not representative.	No
3	Former subdivision golf course surrounded by homes. Acquired for residential infill. Not representative.	No
5	Church land sale to a school district. Surrounded by residential acreage. Not representative.	No
6	Although buyer and seller were under different corporate names, the sale was between related parties. Not representative.	No
9	Residential land purchase located between two residential subdivisions. No competing industrial uses in the vicinity. Not representative.	No
18	Residential land purchase. Surrounded by executive homes on small acreages. Not representative.	No
19	Verification revealed the railroad would not allow a sewer line crossing in this area. Not representative.	No

Exhibit 7 Market Delineated and Segmented Model Inputs

Observation	$/SF	Highly Irr. Shape	Rolling Topo.	Extreme Topo. (Drainage, Flood & Other)	Primary Road Frontage	Dual Primary Road Frontage	Hwy. Exp.	Dist. to Int. Hwy. (sq. rt. miles)	Med. Surr. Dev.	High Surr. Dev.
1	$1.36	0	0	0	1	0	1	0.63	1	0
4	$1.11	0	0	1	0	1	1	0.89	1	0
7	$0.49	0	0	0	0	0	0	0.89	0	0
8	$0.65	0	0	1	0	1	0	0.71	1	0
10	$0.39	0	0	1	0	1	0	1.41	0	0
11	$1.35	0	0	0	0	1	1	1.22	0	1
12	$1.07	0	1	0	1	0	0	0.80	0	1
13	$0.46	1	1	0	1	0	0	1.10	1	0
14	$0.45	0	1	0	1	0	0	1.48	1	0
15	$1.06	0	0	0	1	0	0	1.00	0	1
16	$0.88	1	0	1	1	0	1	0.50	1	0
17	$1.15	0	1	0	0	1	0	0.00	1	0
20	$0.49	0	0	1	0	0	0	1.22	1	0
21	$0.44	0	1	0	0	1	0	1.12	0	0

Exhibit 8 Market Delineated and Segmented Regression Results

Regression Statistics						
Multiple *R*	0.9870					
R Square	0.9742					
Adjusted *R* Square	0.9162					
Standard Error	0.1056					
Observations	14					
			ANOVA			
	df	***SS***	***MS***	***F***	***Significance F***	
Regression	9.000	1.686	0.187	16.802	0.008	
Residual	4.000	0.045	0.011			
Total	13.000	1.730				
	Coefficients	**Standard Error**	***t* Stat**	***P*-value**	**Lower 95%**	**Upper 95%**
Intercept	0.873	0.132	6.621	0.003	0.507	1.239
Highly Irregular Shape	–0.272	0.109	–2.496	0.067	–0.574	0.030
Rolling Topography	–0.065	0.109	–0.595	0.584	–0.368	0.238
Extreme Topography	–0.143	0.101	–1.416	0.230	–0.423	0.137
Primary Road Frontage	0.097	0.119	0.817	0.460	–0.233	0.427
Dual Primary Road Frontage	0.109	0.108	1.007	0.371	–0.191	0.409
Highway Exposure	0.395	0.093	4.225	0.013	0.135	0.654
Distance to Int. Hwy. (sq. rt. miles)	–0.383	0.085	–4.510	0.011	–0.619	–0.147
Med. Surr. Dev.	0.185	0.097	1.910	0.129	–0.084	0.453
High Surr. Dev.	0.461	0.106	4.349	0.012	0.167	0.756

manufacturing, interstate highways are a primary mode of transporting their goods to market. Consequently, the distance to interstate highway system coefficient appears reasonable and indicates a value loss of approximately $0.27/SF for a one-half mile distance[10] and $0.38/SF for a one-mile distance from the interstate highway compared to immediate access.

Model Specification: Regression Trade-Offs, Deficiencies, and Refinements

Aggregation bias is a common deficiency of real estate regression models, but it has received limited discussion in the appraisal literature. Aggregation bias may result in statistically significant value models that are invalid and misleading. An appraiser must ensure the assumptions of validity and representativeness are satisfied. Failing to do so results in misleading outcomes as previously shown in the aggregated model (see Exhibit 5, Initial Aggregated Regression Results). Fortunately, appraisers have a solution to aggregation bias: segmenting the data in accordance with market delineation (shown in Exhibit 6) prior to modeling market behavior using regression (shown in Exhibit 8). Only after the data set has been segmented and delineated can model specification and issues related to significance and fit be effectively addressed. This article addresses potential deficiencies of the delineated and segmented model and suggests solutions.

Economic significance considers whether a coefficient is large enough to matter and has importance in the real-world context. The relative sizes of the coefficients to the modeled price range indicate they are relevant to the market or economically significant. While the magnitude of the coefficients is indicative of their economic significance, a potential deficiency in the delineated and segmented regression model (Exhibit 8) includes generally low *statistical significance* for the individual coefficients. The low statistical significance is partially attributable to the small sample size relative to the number of predictor variables. This is often the cost of satisfying the assumption of representativeness in real estate data. Related to this downside is the way ordinal data (qualitative data that can be ordered on a hierarchical scale) is commonly treated in regression modeling. Topography, frontage, and surrounding development are three ordinal variables used in the model. As there are three levels to each of the three ordinal variables, they are inputted into the regression model as six dummy variables since one level is represented as zero by default. In cases with a high number of dummy predictor variables representing ordinal data, it may be useful to input them as discrete numerical ratings[11] analogous to a typical property productivity analysis. Suggested here is a hybrid multiple regression model where ordinal ratings are regressed as a single variable in the case of linear (near constant) effects while dummy variables are retained when nonlinear effects are indicated.

An upside to regressing ratings as single variables versus numerous dummy variables is that the number of predictor variables decreases relative to the sample size. Consequently, the significance of the model and its coefficients are likely to increase, resulting in greater confidence that the modeled effects are distinguishable from chance. The downside has previously been discussed by A. Ason Okoruwa, who notes that if an ordinal variable is included in the estimated equation as any other discrete or continuous variable, then its coefficient represents a constant impact of a one-unit increase in the ordinal predictor variable.[12] For example, the delineated regression model (Exhibit 8) indicates a value increase of $0.10/SF for primary road frontage and an $0.11/SF value increase for dual primary road frontage compared to an interior parcel. This nonconstant effect would not be captured in a regression model using frontage ratings. However, the

10. $-0.38284 \times \sqrt{0.5} = -0.2707$

11. An extreme form of regressing property productivity ratings known as price-quality regression has been discussed by D. Richard Wincott; in that type of regression model all predictor variables are consolidated into a single weighted rating that is then regressed as a single variable. One of the most elegant features of multiple regression is that the contribution of each predictor variable is given by its coefficient. This feature is lost when all predictors are consolidated into a single rating as in price-quality regression. D. Richard Wincott, "An Alternative Sales Analysis Approach for Vacant Land Valuation," *The Appraisal Journal* (Fall 2012): 310–317.

12. A. Ason Okoruwa, "How to Interpret Regression Coefficients and Calculate Adjustments for Differences in Property Productivity Features," *The Appraisal Journal* (Winter 2018): 68–84.

Exhibit 9 Market Delineated and Segmented Model Inputs (Using Discrete Ratings for Topography and Surrounding Development)

$/SF	Highly Irregular Shape	Topography	Single Primary Road Frontage	Dual Primary Road Frontage	Highway Exposure	Dist. to Hwy. (sq. rt. miles)	Surrounding Development
$1.36	0	0	1	0	1	0.63	1
$1.11	0	2	0	1	1	0.89	1
$0.49	0	0	0	0	0	0.89	0
$0.65	0	2	0	1	0	0.71	1
$0.39	0	2	0	1	0	1.41	0
$1.35	0	0	0	1	1	1.22	2
$1.07	0	1	1	0	0	0.80	2
$0.46	1	1	1	0	0	1.10	1
$0.45	0	1	1	0	0	1.48	1
$1.06	0	0	1	0	0	1.00	2
$0.88	1	2	1	0	1	0.50	1
$1.15	0	1	0	1	0	0.00	1
$0.49	0	2	0	0	0	1.22	1
$0.44	0	1	0	1	0	1.12	0
		0 = Level 1 = Rolling 2 = Extreme					0 = Low 1 = Medium 2 = High

topography and surrounding development variables indicate generally constant (linear) effects between units on the ordinal scale (Exhibit 9); therefore they are good candidates to input as ratings rather than dummy variables. The alternative delineated regression model in Exhibit 10 demonstrates the increased statistical significance when the dummy variables with near linear (constant) effects along the ordinal scale (topography and surrounding development) are replaced by singular discrete rating variables. It is important to emphasize that real estate valuation models should not be specified only by considerations of statistical significance. Models should be built and specified in accordance with market logic. The ratings model in Exhibit 10 describes the data well, has robust predictive power (see Exhibit 10 note), and shows increased statistical significance across all coefficients. These desirable model characteristics are a natural consequence of comprehensive market delineation of the data set to avoid aggregation bias and the reduced number of predictor variables consistent with market logic and statistical practices.

As the comparison in Exhibit 11 illustrates, the regression model using ratings for topography and surrounding development results in greater statistical significance for all predictor variables and the overall model. The increased statistical significance increases confidence that the modeled effects are distinguishable from chance. The advantage of ratings over dummy variables is that the number of predictors relative to the data set is reduced without significant loss of economically relevant information. However, using a discrete ratings scale is only justified when the effects between one unit and the next are generally linear (constant). If the effects are nonlinear, then economically significant information will be lost and dummy variables should be used instead.

Exhibit 10 Market Delineated and Segmented Regression Results (Using Discrete Ratings for Topography and Surrounding Development)

Regression Statistics						
Multiple *R*	0.9859					
R Square	0.9720					
Adjusted *R* Square	0.9393					
Standard Error	0.0899					
Observations	14					
			ANOVA			
	df	***SS***	***MS***	***F***	***Significance F***	
Regression	7	1.6820	0.2403	29.7609	0.0003	
Residual	6	0.0484	0.0081			
Total	13	1.7305				
	Coefficients	**Standard Error**	***t* Stat**	***P*-value**	**Lower 95%**	**Upper 95%**
Intercept	0.8453	0.1039	8.1320	0.0002	0.5909	0.8453
Highly Irregular Shape	–0.2710	0.0915	–2.9618	0.0252	–0.4948	–0.0471
Topography	–0.0874	0.0350	–2.4959	0.0468	–0.1732	–0.0017
Single Primary Road Frontage	0.0962	0.0898	1.0713	0.3252	–0.1236	0.3161
Dual Primary Road Frontage	0.1243	0.0783	1.5862	0.1638	–0.0674	0.3160
Highway Exposure	0.3793	0.0586	6.4774	0.0006	0.2360	0.5226
Dist. to Hwy. (sq. rt. miles)	–0.3638	0.0665	–5.4693	0.0016	–0.5266	–0.2011
Surrounding Development	0.2320	0.0450	5.1572	0.0021	0.1219	0.3421

Note: The predictive *R*-squared is a measure of how well the model predicts the responses for new observations by iteratively holding out each observation and comparing its predicted value to its actual value. A model that overfits the data by describing random noise is generally poor at prediction. The predictive *R*-squared for the Exhibit 10 model was 0.8614, indicating the model is robust for forecasting, is predicting holdout data points, and is not a result of overfitting random noise due to a high number of predictor variables relative to the size of the data set. The predictive *R*-squared of 0.8614 for the delineated model using ratings is also substantially higher than the predictive *R*-squared of 0.6588 for the delineated model using dummy variables presented in Exhibit 8.

Exhibit 11 Comparison of the Delineated Regression Results: Exhibit 8 Dummy Variables vs. Exhibit 10 Discrete Ratings

Coefficient Comparison			*P*-Values Comparison		
Model	**Exhibit 10 (Ratings)***	**Exhibit 8 (Dummy)**	**Model**	**Exhibit 10 (Ratings)**	**Exhibit 8 (Dummy)**
Intercept	0.8453	0.8730	Regression (*F*)	0.0003	0.0077
Highly Irregular Shape	–0.2710	–0.2718	Highly Irregular Shape	0.0252	0.0670
Rolling Topography	–0.0874	–0.0649	Rolling Topography	0.0468	0.5836
Extreme Topography	–0.1749	–0.1428	Extreme Topography	0.0468	0.2298
Single Primary Road Frontage	0.0962	0.0970	Single Primary Road Frontage	0.3252	0.4600
Dual Primary Road Frontage	0.1243	0.1090	Dual Primary Road Frontage	0.1638	0.3707
Highway Exposure	0.3793	0.3947	Highway Exposure	0.0006	0.0134
Dist. to Hwy. (sq. rt. miles)	–0.3638	–0.3828	Dist. to Hwy. (sq. rt. miles)	0.0016	0.0107
Med. Surrounding Development	0.2320	0.1845	Med. Surrounding Development	0.0021	0.1288
High Surrounding Development	0.4640	0.4615	High Surrounding Development	0.0021	0.0122

*The coefficient for extreme topography under ratings is equal to twice the coefficient for rolling topography. The coefficient for high surrounding development under ratings is equal to twice the coefficient for medium surrounding development. The ratings model illustrates substantial improvement in the statistical significance of the model and the individual coefficients.

Interpreting Regression Results and the Sales Comparison Approach

The delineated regression models discussed here fit the underlying data well, and the additive relationship[13] among the predictor variables allows for direct application to the traditional sales comparison approach. Applying the adjustments from regression to the sales comparison approach provides an opportunity for further reconciliation. By selecting sales that the appraiser deems most comparable to the property being valued, the appraiser can further analyze the subject property's position in the market and account for features that may not have been sufficiently captured by the regression model. It is also possible that the sales comparison approach with properly selected comparable sales may partially mitigate the negative effects of aggregation bias compared to direct application of the aggregated model itself and inform an appraiser that a valuation model is deficient. Exhibit 12 compares the adjustments reconciled from the delineated models to the biased adjustments indicated by the aggregated model. In Exhibits 13A, 13B, and 13C, the delineated and aggregated adjustments are applied to the traditional sales comparison approach to further highlight the misleading effects of aggregation bias.

13. The dependent variable, $/SF, was left in its original form and not transformed in this case study. The most common transformation discussed in literature is the natural log transformation. While the advantages of log transformations of the dependent variable have been widely discussed, there are valid reasons not to do so. Linear regression on a log scale mathematically equates to a multiplicative model on the original scale. Rather than compounding percentage adjustments, an additive model was used in this article as it has a more natural interpretation when applied to the sales comparison approach. Additionally, adjustments for elements such as topography are primarily related to increased development costs, which are typically fixed or additive regardless of the values for other elements of comparison.

Exhibit 12 Adjustments Based on Regression Coefficients

Model	Delineated			Aggregated (Biased)
Shape	**Delineated / Dummy Var.**	**Delineated/ Ratings**	**Reconciled Adjustment $/SF**	**Aggregated Adjustment $/SF**
Functional	Base	Base	Base	Base
Highly Irregular	–$0.27	–$0.27	–$0.27	–$0.12
Topography	**Delineated Model**	**Delineated/ Ratings**	**Reconciled Adjustment $/SF**	**Aggregated Adjustment $/SF**
Level	Base	Base	Base	Base
Rolling	–$0.06	–$0.09	–$0.08	–$0.18
Extreme	–$0.14	–$0.17	–$0.16	–$0.13
Frontage	**Delineated Model**	**Delineated/ Ratings**	**Reconciled Adjustment $/SF**	**Aggregated Adjustment $/SF**
Interior	Base	Base	Base	Base
Primary Road	$0.10	$0.10	$0.10	$0.08
Dual Primary Road	$0.11	$0.12	$0.12	$0.03
Direct Highway Exposure	**Delineated Model**	**Delineated/ Ratings**	**Reconciled Adjustment $/SF**	**Aggregated Adjustment $/SF**
No Direct Hwy. Exposure	Base	Base	Base	Base
Direct Hwy. Exposure	$0.39	$0.38	$0.38	$0.26
Distance to Interstate Highway (miles)*	**Delineated Model**	**Delineated/ Ratings**	**Reconciled Adjustment $/SF**	**Aggregated Adjustment $/SF**
0	$0.00	$0.00	$0.00	$0.00
0.25	–$0.19	–$0.18	–$0.18	–$0.08
0.5	–$0.27	–$0.26	–$0.26	–$0.11
0.75	–$0.33	–$0.32	–$0.32	–$0.14
1	–$0.38	–$0.36	–$0.36	–$0.16
1.25	–$0.43	–$0.41	–$0.41	–$0.18
1.5	–$0.47	–$0.45	–$0.45	–$0.19
1.75	–$0.51	–$0.48	–$0.48	–$0.21
2	–$0.54	–$0.51	–$0.51	–$0.22
Surrounding Development	**Delineated Model**	**Delineated/ Ratings**	**Reconciled Adjustment $/SF**	**Aggregated Adjustment $/SF**
Low Density / Older Vintage	Base	Base	Base	Base
Medium	$0.18	$0.23	$0.23	$0.12
High / Newer Commercial	$0.46	$0.46	$0.46	$0.44

*Adjustment based on square root of miles times coefficient.

Exhibit 13A Sales Comparison Using Reconciled Regression Adjustments: Parcel A Sales Comparison

Parcel / Sale	A	Sale 7	Sale 8	Sale 12
Zoning	I-2 (Industrial)	I-3	PUD-1706	I-2
Highly Irregular Shape	No	No	No	No
Net Acres	32.50	30.14	74.28	21.88
Topography	Rolling	Level	Extreme	Rolling
Frontage	Dual Primary Rd.	Interior	Dual Primary Rd.	Primary Rd.
Direct Hwy. Exposure	No	No	No	No
Dist. to Hwy. (miles)	1	0.80	0.50	0.64
Surrounding Dev.	Medium	Low	Medium	High
$/SF		$0.49	$0.65	$1.07
Delineated Adjustments				
Highly Irregular Shape		$0.00	$0.00	$0.00
Topography		–$0.08	$0.08	$0.00
Frontage		$0.12	$0.00	$0.02
Direct Highway Exposure		$0.00	$0.00	$0.00
Distance to Interstate Highway (miles)*		–$0.04	–$0.11	–$0.07
Surrounding Development		$0.23	$0.00	–$0.23
Total Adjustments		$0.23	–$0.03	–$0.28
Indicated $/SF (Average of Sales)	$0.71	$0.72	$0.62	$0.79
Delineated Model Value $/SF	$0.72			
Delineated-Ratings Model Value $/SF	$0.75			
Aggregated (Biased) Adjustments				
Highly Irregular Shape		$0.00	$0.00	$0.00
Topography		–$0.18	–$0.05	$0.00
Frontage		$0.03	$0.00	–$0.04
Direct Highway Exposure		$0.00	$0.00	$0.00
Distance to Interstate Highway (miles)*		–$0.02	–$0.05	–$0.03
Surrounding Development		$0.12	$0.00	–$0.32
Total Adjustments		–$0.04	–$0.10	–$0.39
Indicated $/SF (Average of Sales)	$0.56	$0.45	$0.55	$0.68
Aggregated Model Value $/SF	$0.57			

*Adjustment based on square root of miles times coefficient.

Exhibit 13B Sales Comparison Using Reconciled Regression Adjustments: Parcel B Sales Comparison

Parcel / Sale	B	Sale 1	Sale 4	Sale 16
Zoning	I-2 (Industrial)	I-2	A-2	I-2
Highly Irregular Shape	No	No	No	Yes
Net Acres	57.65	21.89	57.06	22.09
Topography	Rolling	Level	Extreme	Extreme
Frontage	Primary Road	Primary Rd.	Dual Primary Rd.	Primary Rd.
Direct Hwy. Exposure	Yes	Yes	Yes	Yes
Dist. to Hwy. (miles)	0.1	0.4	0.80	0.25
Surrounding Dev.	Medium	Medium	Medium	Medium
$/SF		$1.36	$1.11	$0.88
Delineated Adjustments				
Highly Irregular Shape		$0.00	$0.00	$0.27
Topography		–$0.08	$0.08	$0.08
Frontage		$0.00	–$0.02	$0.00
Direct Highway Exposure		$0.00	$0.00	$0.00
Distance to Interstate Highway (miles)*		$0.12	$0.21	$0.07
Surrounding Development		$0.00	$0.00	$0.00
Total Adjustments		$0.04	$0.27	$0.42
Indicated $/SF (Average of Sales)	$1.36	$1.40	$1.38	$1.30
Delineated Model Value $/SF	$1.36			
Delineated-Ratings Model Value $/SF	$1.35			
Aggregated (Biased) Adjustments				
Highly Irregular Shape		$0.00	$0.00	$0.12
Topography		–$0.18	–$0.05	–$0.05
Frontage		$0.00	$0.04	$0.00
Direct Highway Exposure		$0.00	$0.00	$0.00
Distance to Interstate Highway (miles)*		$0.05	$0.09	$0.03
Surrounding Development		$0.00	$0.00	$0.00
Total Adjustments		–$0.13	$0.08	$0.09
Indicated $/SF (Average of Sales)	$1.13	$1.23	$1.19	$0.97
Aggregated Model Value $/SF	$0.98			

*Adjustment based on square root of miles times coefficient.

Exhibit 13C Sales Comparison Using Reconciled Regression Adjustments: Parcel C Sales Comparison

Parcel / Sale	C	Sale 8	Sale 15	Sale 16
Zoning	I-2 (Industrial)	PUD-1706	I-2	I-2
Highly Irregular Shape	No	No	No	Yes
Net Acres	45.25	74.28	34.63	22.09
Topography	Level	Extreme	Level	Extreme
Frontage	Primary Rd.	Dual Primary Rd.	Primary Rd.	Primary Rd.
Direct Hwy. Exposure	No	No	No	Yes
Dist. to Hwy. (miles)	1.5	0.50	1.00	0.25
Surrounding Dev.	Medium	Medium	High	Medium
$/SF		$0.65	$1.06	$0.88
Delineated Adjustments				
Highly Irregular Shape		$0.00	$0.00	$0.27
Topography		$0.16	$0.00	$0.16
Frontage		–$0.02	$0.00	$0.00
Direct Highway Exposure		$0.00	$0.00	–$0.38
Distance to Interstate Highway (miles)*		–$0.19	–$0.08	–$0.26
Surrounding Development		$0.00	–$0.23	$0.00
Total Adjustments		–$0.05	–$0.31	–$0.21
Indicated $/SF (Average of Sales)	$0.67	$0.60	$0.75	$0.67
Delineated Model Value $/SF	$0.69			
Delineated-Ratings Model Value $/SF	$0.73			
Aggregated (Biased) Adjustments				
Highly Irregular Shape		$0.00	$0.00	–$0.12
Topography		–$0.13	$0.00	–$0.13
Frontage		$0.04	$0.00	$0.00
Direct Highway Exposure		$0.00	$0.00	$0.26
Distance to Interstate Highway (miles)*		–$0.08	–$0.04	–$0.11
Surrounding Development		$0.00	–$0.32	$0.00
Total Adjustments		–$0.16	–$0.36	–$0.10
Indicated $/SF (Average of Sales)	$0.66	$0.49	$0.70	$0.78
Aggregated Model Value $/SF	$0.75			

*Adjustment based on square root of miles times coefficient.

Conclusion

This article shows how aggregation bias may creep into a regression model, and how professional appraisers are equipped to avoid it with the tools of market delineation and segmentation. No amount of statistical testing or advanced mathematics can cure nonrepresentative data. The "law of large numbers" has become cliché in some circles, a platitude to justify models built on giant data sets that ignore basic assumptions of economic behavior. By increasing the sample size of a nonrepresentative sample, a model may become further removed from that which it purports to measure while, ironically, being shielded by increasing "statistical significance." Such models are illusory, and the appraisal industry should be skeptical of any efforts to hide the underlying data and source of algorithmic valuations behind a proprietary black box. An appraiser's initial opinion of any model should be that the model is descriptive, not predictive or inferential. Rather than asking what a model predicts or what inferences can be made, appraisers should first ask what it describes. If it describes nothing, then it predicts nothing. It is an appraiser's professional market knowledge, interactions with market participants, and application of the tools of market analysis that make the human appraiser uniquely qualified to make the leap from description to inference. If these real assets of professional appraisers are emphasized convincingly, human appraisers will not be replaced by algorithms for the foreseeable future.

The presence of aggregation bias is damaging to the real estate industry. While few would use a city's median home price as an indicator of value for a specific home, there are more subtle forms of aggregation bias disguised by regression and other sophisticated valuation models. A list of inaccurate, algorithmically produced ad valorem tax valuations that purport to be market value would be exhaustive. Other unfortunate examples include over-aggregated data used in litigation settings involving unique events and unique markets. One highly publicized example of aggregation bias involved Zillow, which shut down its algorithm-driven home buying program in November 2021. Despite Zillow having arguably the largest, most-comprehensive data set of single-family homes and consumer behavior, supported by billions in assets and human capital, "its algorithm proved to be overoptimistic, even in a housing boom."[14] In the aftermath, Zillow priced two-thirds of its homes for less than what it paid, lost 15% in market capitalization in a single day, and laid off 25% of its workforce. Zillow was warned of its overvaluations over a decade prior in an *Appraisal Journal* article by Hollas, Rutherford, and Thomson.[15] Those authors' research found Zillow's valuations to be less accurate than those of a typical homeowner, with Zillow overvaluing homes by 10% on average in a market Zillow had reported to be its most accurate. While Zillow's mistakes were limited to a single company, the accelerated growth of automated valuation models makes aggregation bias a market-wide risk. Aggregation bias is not uncommon in the current big-data-driven world, but it has not received sufficient attention. The illusions of big data have been obscured by its promises.

Rather than asking what a model predicts, appraisers should first ask what it describes. If it describes nothing, then it predicts nothing.

As Gelman, Hill, and Vehtari state, "If we do not know what the data actually represent, then we cannot extract the right information. Data analysis reaches a dead end if we have poor data."[16] Market delineation and segmentation practices provide appraisers with the toolset to know what their data represent. Representativeness is necessary for validity. There is no escaping the arguably cumbersome process of market

14. Felix Salmon, "Zillow Abandons Its Home-Flipping Algorithm," Axios, November 2, 2021, https://bit.ly/3HSIkBO.

15. Daniel R. Hollas, Ronald C. Rutherford, and Thomas A. Thomson, "Zillow's Estimates of Single-Family Housing Values," *The Appraisal Journal* (Winter 2010): 26–32.

16. Gelman, Hill, and Vehtari, *Regression and Other Stories*, 23.

delineation and segmentation that may involve detailed confirmation or verification of a plethora of sales. Fortunately, for this hard work the professional human appraiser is uniquely qualified. Market delineation and segmentation should be the first step in valuation modeling, including regression, as it is a fundamental requirement for validity.

About the Author

Matthew C. Trimble, MAI, is the principal of Trimble Valuation, based in Oklahoma City, Oklahoma. He began as a practicing appraiser with Isaacs & Associates in 2008 after receiving his master's degree in mathematics from the University of Oklahoma. In 2019 he received his MAI designation from the Appraisal Institute. He has a broad-based real estate appraisal and consulting practice and is an associate professor of Real Estate Appraising at the University of Central Oklahoma. **Contact: Trimble.okc@gmail.com**

Additional Resources

Suggested by the Y. T. and Louise Lee Lum Library

Appraisal Institute

- **Education**
  - *Quantitative Analysis*
  - *Real Estate Finance, Statistics, and Valuation Modeling*

- **Lum Library Knowledge Base information compilation [Login required]**
  Appraisal Practice—Data, statistics, and statistical analysis

- **Publications**
  - *The Appraisal of Real Estate,* fifteenth edition (Chicago: Appraisal Institute, 2020)
  - *An Introduction to Statistics for Appraisers* (Chicago: Appraisal Institute, 2009)
  - *Practical Applications in Appraisal Valuation Modeling* (Chicago: Appraisal Institute, 2004)
  - *Valuation by Comparison* (Chicago: Appraisal Institute, 2018)