Compstak Analysis Progress

William Clinton Co
Department of Economics, University of British Columbia

April 27, 2025

Abstract

In this project, we conducted an initial review of the CompStak dataset to assess its suitability for analyzing commercial properties in the United States. Each field was individually evaluated, with a focus on 'Property Type' and 'Property Subtype'. Our analysis revealed that while fill rates for 'Property Type' are high (95–99%), 'Property Subtype' fields exhibit higher missingness (up to 28%), potentially complicating downstream analysis.We verified that property IDs are generally stable across time and regions, although inconsistencies in property type and subtype classifications were noted in a small fraction of records ($<$0.02%). To benchmark the dataset's coverage, we compared the number of properties captured in CompStak against external (unverified) estimates of U.S. commercial properties. Our findings indicate non-uniform coverage across property types and states. Specifically, industrial, office, and retail sectors are relatively well-represented, whereas sectors like land and multifamily properties show significant undercoverage. Geographically, West Coast states such as California show higher representation, and regression analysis suggests that coverage rates increase with the number of commercial properties in a given state, indicating increasing returns to scale in data collection. Challenges remain, particularly in mapping category definitions across different datasets, such as the DOE and CoStar sources. Future work will focus on improving benchmark estimates using the DOE dataset and establishing a reliable mapping of property categories between datasets to ensure meaningful comparisons.

1 Initial Analysis

Each field was reviewed individually to identify potential patterns, and all fields appeared relevant for analysis. The broader data set includes several types of industry classifications, such as 'Property Type', 'Property Sub type', 'Space Type', 'Tenant SIC Code/Description', and 'Tenant NAICS Code/Description'. However, there is no clear documentation explaining the differences between these classifications or indicating which should be preferred. In our case, only 'Property Type' and 'Property Sub type' are available, so our analysis will focus on these variables. It is important to note that these classifications are not consistently available across both lease and sales records, which may introduce challenges to our analysis. Additionally, NAICS and SIC codes are not included in our data set because they are a premium add-on. Finally, 'Space Type' is only available for lease data, which may also pose limitations.

We observe "fill rate" (Table 1), wherein a low fill rate would correspond to a large portion of the entries having Nan or missing values. Unique values shows us that the categories used for property type and property sub type are consistent.

Table 1: Data set Initial Analysis

Data set	Type	Fill Rate	Unique Values
Sales	Property type	95%	8
Lease	Property type	99%	8
Sales	Property Subtype	75%	56
Lease	Property Subtype	66%	56

We also investigate whether property IDs are consistent. We expect that a given commercial property would exhibit stable characteristics, with its location and industry classification remaining unchanged over time. Our analysis confirms this expectation: property IDs are stable and pass state consistency checks. However, it is important to note that property types and sub types are not consistent. See Figure 1 and Table 5

```
Property Id 646606 has multiple Property Types: [nan, 'Land']
Property Id 665103 has multiple Property Types: ['Other', 'Office']
Property Id 1019665 has multiple Property Types: [nan, 'Other']
Property Id 1235355 has multiple Property Types: ['Retail', 'Industrial']
Property Id 1251623 has multiple Property Types: ['Land', 'Office']
Property Id 1375074 has multiple Property Types: ['Land', 'Multi-Family']
Property Id 1470728 has multiple Property Types: ['Retail', 'Office']
Property Id 1702654 has multiple Property Types: ['Retail', 'Multi-Family']
Property Id 1721284 has multiple Property Types: ['Land', 'Industrial']
Property Id 1721935 has multiple Property Types: ['Retail', nan]
Property Id 1725734 has multiple Property Types: ['Office', 'Retail']
Property Id 1768020 has multiple Property Types: ['Land', 'Retail']
Property Id 1783304 has multiple Property Types: ['Land', nan]
Property Id 1818335 has multiple Property Types: ['Retail', 'Multi-Family']
Property Id 1825935 has multiple Property Types: ['Land', 'Multi-Family']
Property Id 1832887 has multiple Property Types: ['Land', nan]
Property Id 1973729 has multiple Property Types: ['Other', 'Retail']
Property Id 1981969 has multiple Property Types: ['Retail', 'Industrial']
Property Id 1992092 has multiple Property Types: ['Land', 'Retail']
Property Id 2017235 has multiple Property Types: ['Retail', 'Mixed-Use']
Property Id 2023360 has multiple Property Types: ['Retail', 'Office']
Property Id 2106390 has multiple Property Types: ['Retail', nan]
Property Id 2126846 has multiple Property Types: [nan, 'Retail']
Property Id 2186407 has multiple Property Types: [nan, 'Other']
Property Id 2295396 has multiple Property Types: [nan, 'Land']
Property Id 2310705 has multiple Property Types: ['Retail', 'Office']
Property Id 2423644 has multiple Property Types: ['Retail', 'Other']
Property Id 2720291 has multiple Property Types: ['Land', 'Retail']
Property Id 3141000 has multiple Property Types: ['Office', 'Other']
Property Id 3587132 has multiple Property Types: ['Retail', 'Mixed-Use']
```

Figure 1: Property ID Unstable Classifications

Despite concerns about inconsistency, this is not a major issue, as the number of unstable observations is relatively small compared to the total number of properties. Specifically, property type inconsistency is observed in only 39 cases, and property sub type inconsistency in 96 cases. This pales in comparison to the 759,623 unique properties in our data set.

What is more concerning is the Nan surrounding each property type and sub type. Wherein the Nan values for property sub types can be as high as 28% of the data. See Table 2

Table 2: Error Associated with Property Types

	Industry Category		
Error Type	Type	Number	Percentage
Inconsistent Category	Property Type	39 / 759623	0.00513%
Inconsistent Category	Property Subtype	96 / 759623	0.01264%
Nan	Property Type	37020 / 759623	4.87%
Nan	Property Subtype	213255 / 759623	28.07%

2 Strategy

We will begin by matching the number of buildings. Specifically, we observe unique property IDs in the data set, allowing us to estimate the number of commercial properties in the United States. The total number of commercial properties is a relatively stable metric to study compared to more volatile measures such as valuations or square footage.

Using this approach, we can first compare aggregate numbers, starting with the total number of properties in the United States, and then work our way downward to industry-level and state-level comparisons.

We first assess how much of the national commercial property market is captured in our data set. We assume that the CompStak data set represents only a small subsection of the total U.S. market. However, a key concern is whether these observations are uniformly distributed across regions and industries, or if there are systematic biases. For example, are retail properties in California more likely to be reported than warehouses in Michigan?

To begin, we pull external estimates of the total number of commercial properties from publicly available internet sources. Although these external figures are unverified, they provide a useful starting point for benchmarking and analyzing the coverage of our data set. See Table 3 and Table 4

3 Results

The following analysis uses unverified U.S. figures. We observe evidence of non-uniform coverage. If we assume the complete circle represents the true total number of U.S. commercial properties, then the CompStak data represents a subsample of this total. The "Covered" section reflects properties captured in the CompStak dataset, while the "Gap" represents the "missing" properties not covered.

As shown in Figure 2 , office, industrial, and retail sectors appear relatively well covered. Industrial coverage is approximately 53%, while office and retail each have around 22% coverage. Coverage for the remaining sectors is negligible, with less than 5% represented.

In particular see Figure 3 see that land and multi-family have extremely low coverage, at the same time representing a large portion of total US commercial properties.

Using the same data set, I investigated the uniformity of coverage across states. The analysis reveals that the CompStak dataset exhibits a bias toward West Coast states. As

Coverage Analysis by Property Type (Sunburst Visualization)

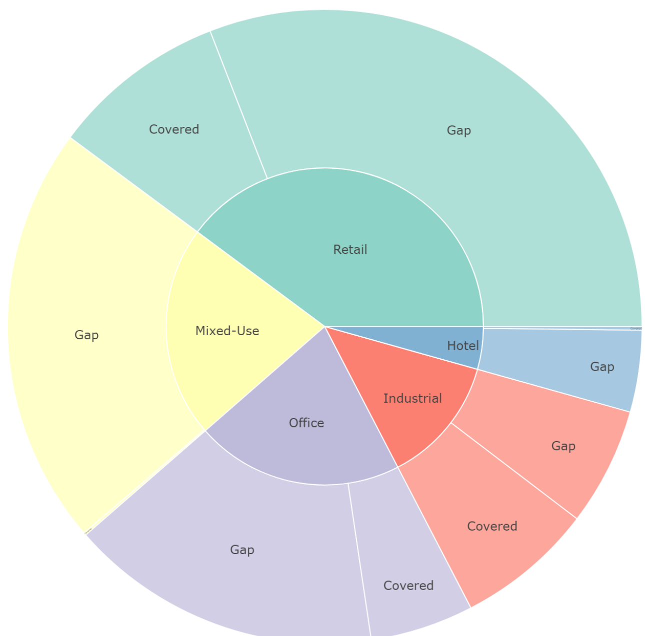


Figure 2

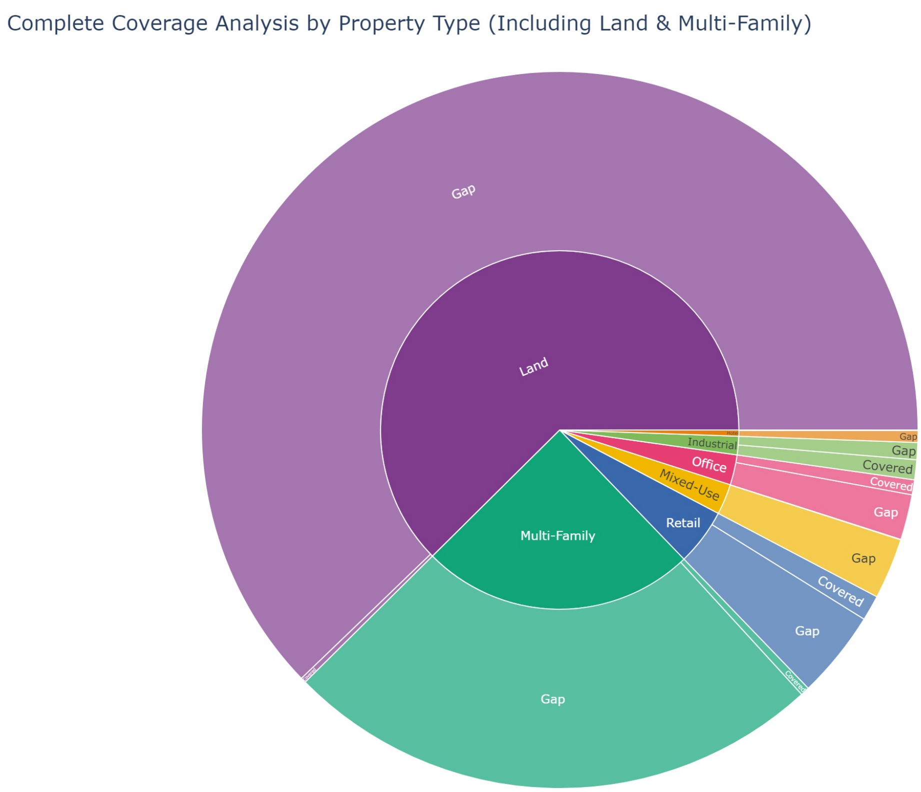

Complete Coverage Analysis by Property Type (Including Land & Multi-Family)
Gap
Land
Multi-Family
Gap
Retail
Gap
Covered
Mixed-Use
Gap
Office
Gap
Covered
Industrial
Gap
Covered
Gap

Figure 3

shown in Figure 4, the coverage rate is calculated by assuming the externally sourced U.S. data set estimate provides the true number of commercial properties, with the CompStak dataset representing a subset. For example, California has an 18% coverage rate, meaning that the number of commercial property observations in CompStak accounts for 18% of the estimated 917,860 commercial properties in California.

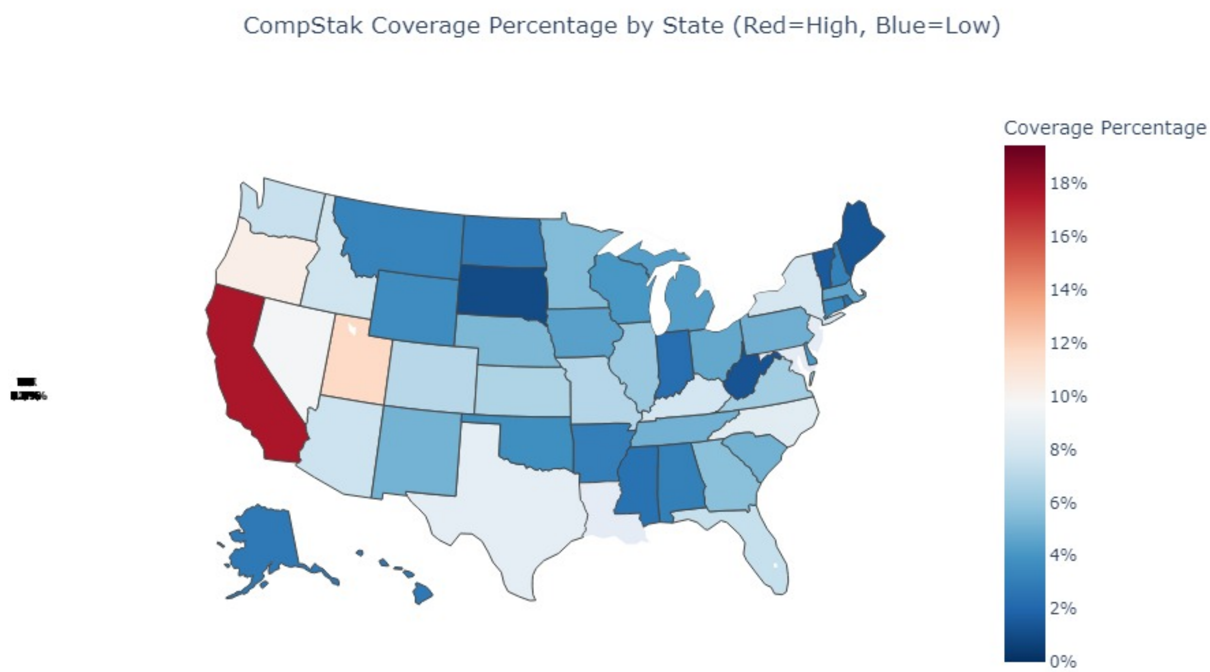


Figure 4

Next, we examine the determinants of this observation. Specifically, I assess whether the coverage rate is a function of the total number of commercial properties in each state, again assuming that our U.S. estimate represents the true total. As shown in Figure 5, there is evidence to support this theory. States with a greater number of commercial properties tend to have higher coverage rates in the CompStak dataset, as indicated by the linear regression results in Figure 5. This pattern suggests that there are fixed costs associated with data collection in each location, and that data providers are more likely to specialize in areas with larger commercial markets. As a result, the dataset exhibits increasing returns to scale, which introduces bias into our observations.

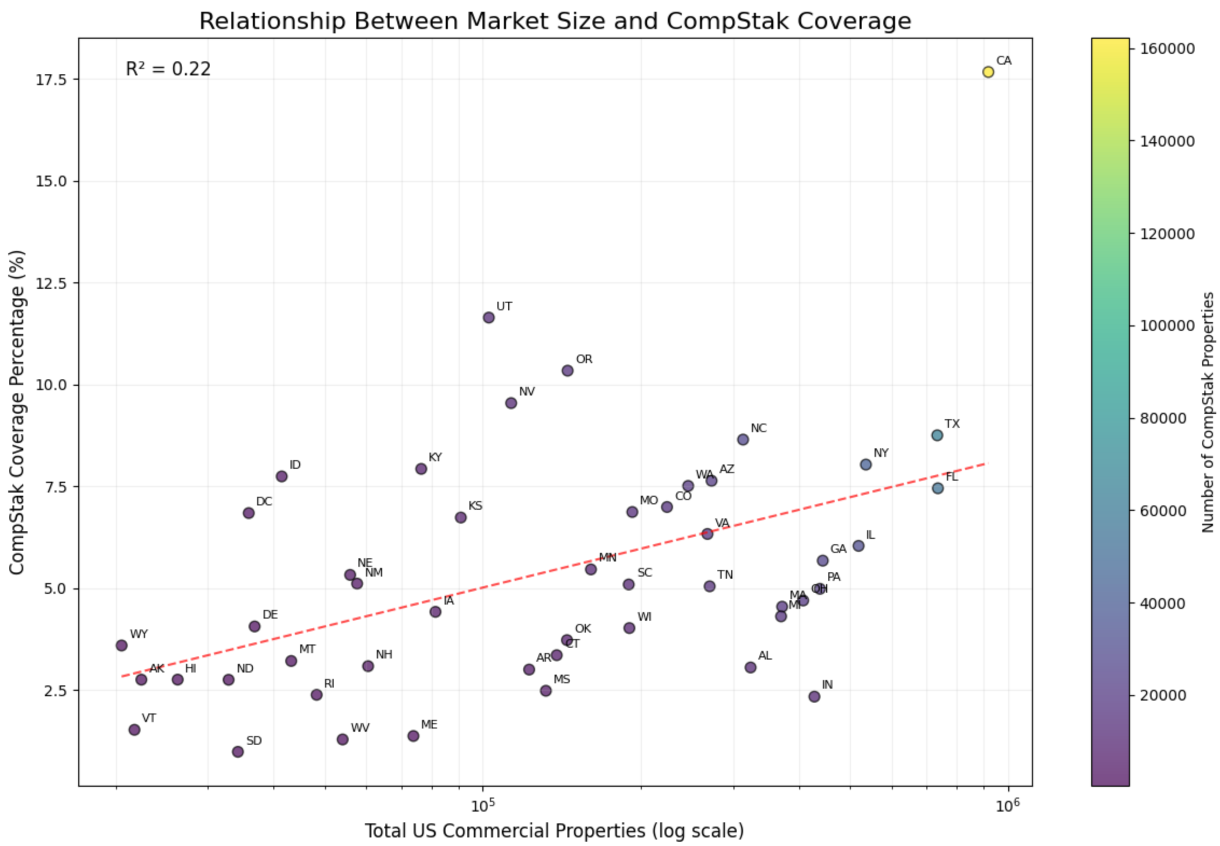


Figure 5

A similar analysis was conducted using property types, but several challenges were encountered. Specifically, the top two outliers significantly skewed the regression line, suggesting potential issues with category interpretation or data quality. Further investigation is warranted to better understand these anomalies. However, it is noteworthy that when these outliers are excluded, the regression line again shows a positive relationship: property types with more commercial properties tend to have higher CompStak coverage rates. The corresponding graph can be found in the appendix. Figure 6 Figure 7

4 Concluding Thoughts

NAICS and SIC codes may be worth considering, given their standardized format and consistency across multiple datasets, depending on our analytical needs. We also realized that our observations may be biased due to unverified numbers in our current data set. Therefore, to ensure greater accuracy, we will be studying the DOE dataset of estimated numbers of commercial properties in the United States (CoStar Glossary, DOE Dataset).

While this is a good starting point, it will require significant effort to reliably reproduce our unverified US estimate data using the DOE dataset. Determining the appropriate mapping and definitions takes time. For example, the DOE categorizes commercial properties under terms such as “sports and entertainment” or “specialty,” which may not directly correspond to the categories used in the CompStak dataset. The CoStar Glossary provides definitions for some of these categories, and a comparison with Table 3 in our dataset highlights these differences. This underscores the need for careful mapping and interpretation when aligning categories across datasets.

A careful review of category definitions and thoughtful value judgments will be necessary to establish equivalencies and ensure meaningful comparisons in our analysis. Additionally, it would be beneficial to determine whether a comprehensive glossary of categories exists for the CompStak dataset. To the best of my knowledge, there is currently no CompStak glossary that is directly comparable to the CoStar Glossary.

5 Appendix

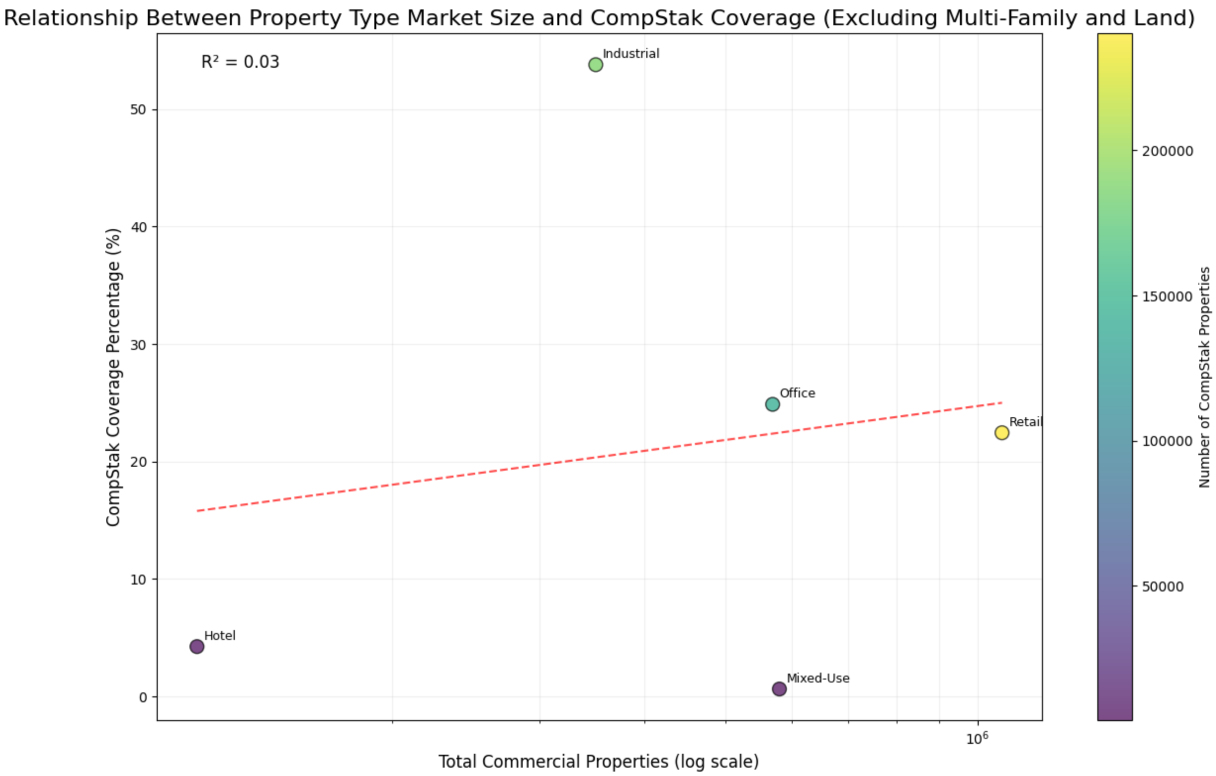
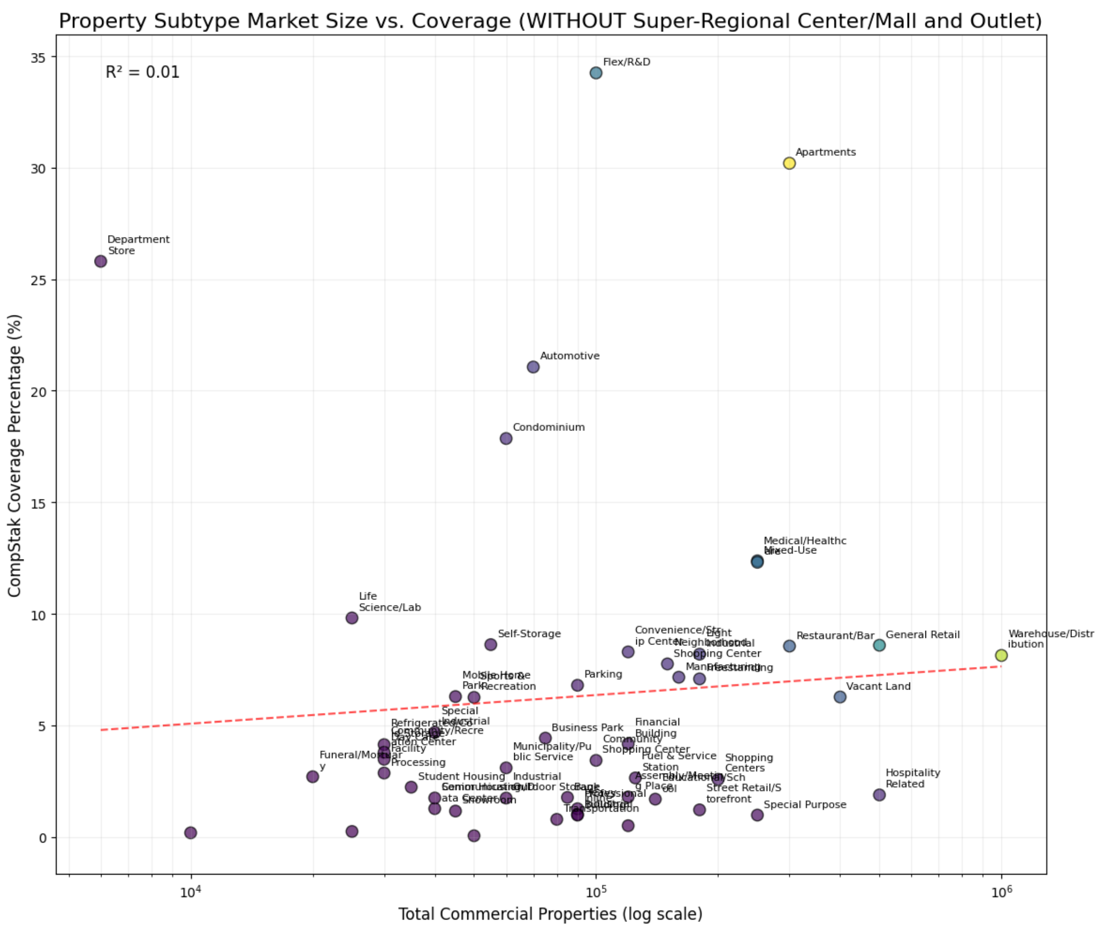


Figure 6

Table 3: Number of Commercial Properties by Industry (Unverified Internet Sourced US estimates)

Property Type	Estimated Number of Properties
Retail	1,070,000
Industrial	350,000
Office	569,311
Multi-Family	5,200,000
Hotel	116,873
Mixed-Use	580,000
Land	13,100,000



Property Subtype Market Size vs. Coverage (WITHOUT Super-Regional Center/Mall and Outlet)
R² = 0.01
CompStak Coverage Percentage (%)
Total Commercial Properties (log scale)
Number of CompStak Properties
Flex/R&D
Apartments
Department Store
Automotive
Condominium
Medical/Healthcare
Mixed-Use
Life Science/Lab
Self-Storage
Restaurant/Bar
General Retail
Warehouse/Distribution
Parking
Vacant Land
Business Park
Financial Building
Student Housing
Industrial
Special Purpose
Hospitality Related
Transportation

Figure 7

Other	Not specified

Table 4: Number of Commercial Properties by State (Unverified Internet Sourced US estimates)

State	Commercial Properties
CA	917,860
TX	733,648
FL	735,652
NY	536,608
IL	519,616
PA	438,648
OH	407,557
GA	444,143
NC	313,187
MI	369,983
WA	246,208
AZ	272,797
MA	371,710
VA	267,936
CO	224,418
IN	428,138
TN	270,544
MO	192,733

State	Commercial Properties
WI	190,274
MN	160,773
AL	323,716
SC	189,736
KY	76,415
OR	145,157
OK	144,752
CT	138,387
IA	81,338
MS	131,969
AR	122,634
KS	90,904
NV	113,336
UT	102,769
NM	57,693
NE	55,961
WV	54,143
ID	41,460
HI	26,275
ME	73,831
NH	60,537
RI	48,317
MT	43,219

State	Commercial Properties
DE	36,816
SD	34,215
ND	32,846
AK	22,410
VT	21,740
WY	20,549
DC	35,878

6

Table 5: Property Sub type Inconsistency

Property ID	Inconsistent Property Subtypes
21036	General Retail, nan
353844	nan, Vacant Land
354461	nan, Municipality/Public Service
374802	nan, Vacant Land
398906	nan, Vacant Land
415099	General Retail, nan
418514	Apartments, Sports & Recreation
420902	nan, Apartments
422199	Outlet, Vacant Land
433935	nan, Municipality/Public Service

Property ID	Inconsistent Property Subtypes
434400	nan, Vacant Land
443800	Vacant Land, Super-Regional Center/Mall
444319	Apartments, nan
445159	General Retail, nan
448126	Apartments, nan
449199	General Retail, nan
466525	General Retail, nan
476566	Parking, Apartments
485298	nan, Shopping Centers
489216	nan, Parking
490863	nan, Automotive
491210	Apartments, General Retail
491863	Apartments, Vacant Land
496416	nan, Apartments
508849	General Retail, Shopping Centers
520174	nan, Vacant Land
538162	Vacant Land, nan
567596	Super-Regional Center/Mall, Neighborhood Shopping Center
581027	nan, General Retail
581309	Apartments, nan
623357	Vacant Land, nan
624813	Condominium, Apartments
633831	Apartments, nan

Property ID	Inconsistent Property Subtypes
646161	General Retail, nan
669817	nan, Apartments
679028	Apartments, nan
699484	nan, Apartments
702675	Apartments, nan
703449	nan, Apartments
731139	Apartments, Convenience/Strip Center
742440	nan, Apartments
745417	Vacant Land, nan
754857	nan, Apartments
755143	General Retail, nan
757418	Apartments, nan
849172	Flex/R&D, Business Park
1204302	General Retail, nan
1211995	nan, Sports & Recreation
1212507	Special Purpose, nan
1235355	General Retail, Vacant Land
1254651	Self-Storage, nan
1255443	Vacant Land, Condominium
1261872	nan, Mixed-Use
1272015	Freestanding, General Retail
1311302	General Retail, Freestanding
1418012	Warehouse/Distribution, Special Industrial

Property ID	Inconsistent Property Subtypes
1421693	Apartments, nan
1431770	nan, Vacant Land
1448113	nan, General Retail
1449588	Apartments, Financial Building
1451345	nan, Apartments
1684081	nan, Light Industrial
1705515	Manufacturing, Light Industrial
1721935	Day Care Facility, nan
1722765	Super-Regional Center/Mall, Convenience/Strip Center
1724884	Vacant Land, nan
1725903	General Retail, Community Shopping Center
1743582	nan, General Retail
1765432	Apartments, nan
1822722	Vacant Land, nan
1858406	nan, Apartments
1866157	Hospitality Related, Apartments
1922481	nan, Apartments
1929300	Condominium, nan
2049015	General Retail, nan
2050907	Apartments, nan
2054074	nan, Super-Regional Center/Mall
2057688	nan, Vacant Land
2096757	nan, Vacant Land

Property ID	Inconsistent Property Subtypes
2106390	Automotive, nan
2144456	Parking, Restaurant/Bar
2266042	Apartments, nan
2288779	Parking, Warehouse/Distribution
2292241	nan, Warehouse/Distribution
2295396	nan, Vacant Land
2330278	Apartments, nan
2331364	Vacant Land, General Retail
2423080	nan, Vacant Land
2425229	nan, Restaurant/Bar
2720291	Vacant Land, Mixed-Use
3417569	nan, Medical/Healthcare
3417659	nan, Mixed-Use
3418319	nan, Mixed-Use
3464690	Warehouse/Distribution, Manufacturing
3575908	Community Shopping Center, Vacant Land
3587132	Apartments, Mixed-Use

References