

# *ABAG Nets and Costar Data Set Geocoding*

Ben Stabler, Sandeep Puppala, PB, 04/05/11

## **Overview**

PB was asked by ABAG to help process two large data sets for land use modeling. The first data set is the National Establishment Time Series (Nets) business establishment data set of every business location in California over the last 20 years. The second data set is the CoStar data set of buildings in the nine-county bay area. This memo describes the data processing and geocoding that was done to get these data sets into a form useful for land use modeling. A number of data summaries are provided as well.

## **Initial Data Processing**

The raw Nets and CoStar data sets were processed for geocoding in this step. The data processing steps are summarized below.

### ***Nets Data Processing***

The raw Nets data set consists of three files: 1) Nets records for establishments in the MSA (Metropolitan Statistical Area) areas of California, 2) Nets records for establishments outside of MSAs, and 3) move records for each move an establishment made over the last twenty years. For starters, the MSA and non-MSA Nets tables were appended to create one Nets table. This Nets table and the Nets Moves table were then added to a MySQL database. From there, the following queries were run to create Nets tables with only the required fields in order to reduce the memory required for data processing. These simplified tables were then converted into yearly tables, as described below.

```
SELECT DunsNumber, Address, City, State, FipsCounty, ZipCode, Latitude, Longitude * -1, LevelCode, SIC8, FirstYear, MoveYears, LastMove, last year FROM nets
```

```
SELECT DunsNumber, MoveYear, MoveSIC, OriginAddress, OriginCity, OriginState, OriginZIP, OriginFIPSCounty, OriginLatitude, OriginLongitude * -1, OriginLevelCode FROM nets_moves
```

These two tables were then loaded into R for data processing and written out to a csv for use in geocoding.

### ***CoStar Data Processing***

The raw CoStar data consisted of a series of Excel workbooks each with a table of CoStar query results. An R script was written to read all the Excel files and merge the records into one master table. The master table of all CoStar records was written to the MySQL database as the CoStar table and also written out to a csv file for use in geocoding.

## Geocoding

Although the Nets and CoStar data have some existing geocodes (i.e. latitude and longitude attributes), the data was re-geocoded in order to build a more comprehensive spatially referenced data set. Inside the nine-county bay area, the Nets and CoStar records were geocoded to either the nine-county bay area parcel layer or the TeleAtlas street layer depending on the ability to match. Outside the bay area, all records were geocoded to the TeleAtlas street layer if possible.

### *Geocoding Layers*

The geocoding process was done using ArcGIS address locator technology. Address locators are GIS layers that are configured for geocoding. A composite address locator was created that consists of two address locators:

- 1) The parcel file with house number, street direction, street name, street type, and city (see Figure 1). The parcel file was converted to WGS 1984 lat/long projection as well so that any X/Y geocoded coordinates output fields are in lat/long.
- 2) The TeleAtlas street layer, which contained housing numbering ranges on the left and right side of the street, all the components of street addresses, city, state, and zip (see Figure 2)

The composite address locator first tries to match to the multi-field parcel address locator. If it doesn't find a reasonable match, it then tries to match to the TeleAtlas street address locator. The Composite address locator settings are shown in Figure 3.

Initially, a third locator was also used, which was a single field version of the parcel file. This locator stored the complete parcel address in one data field. However, the match rate was quite low and the performance was terrible since geocoder had to essentially look through the entire parcel data set when trying to match each record. As a result, the single field parcel locator was dropped.

### *Parcel Data Improvements*

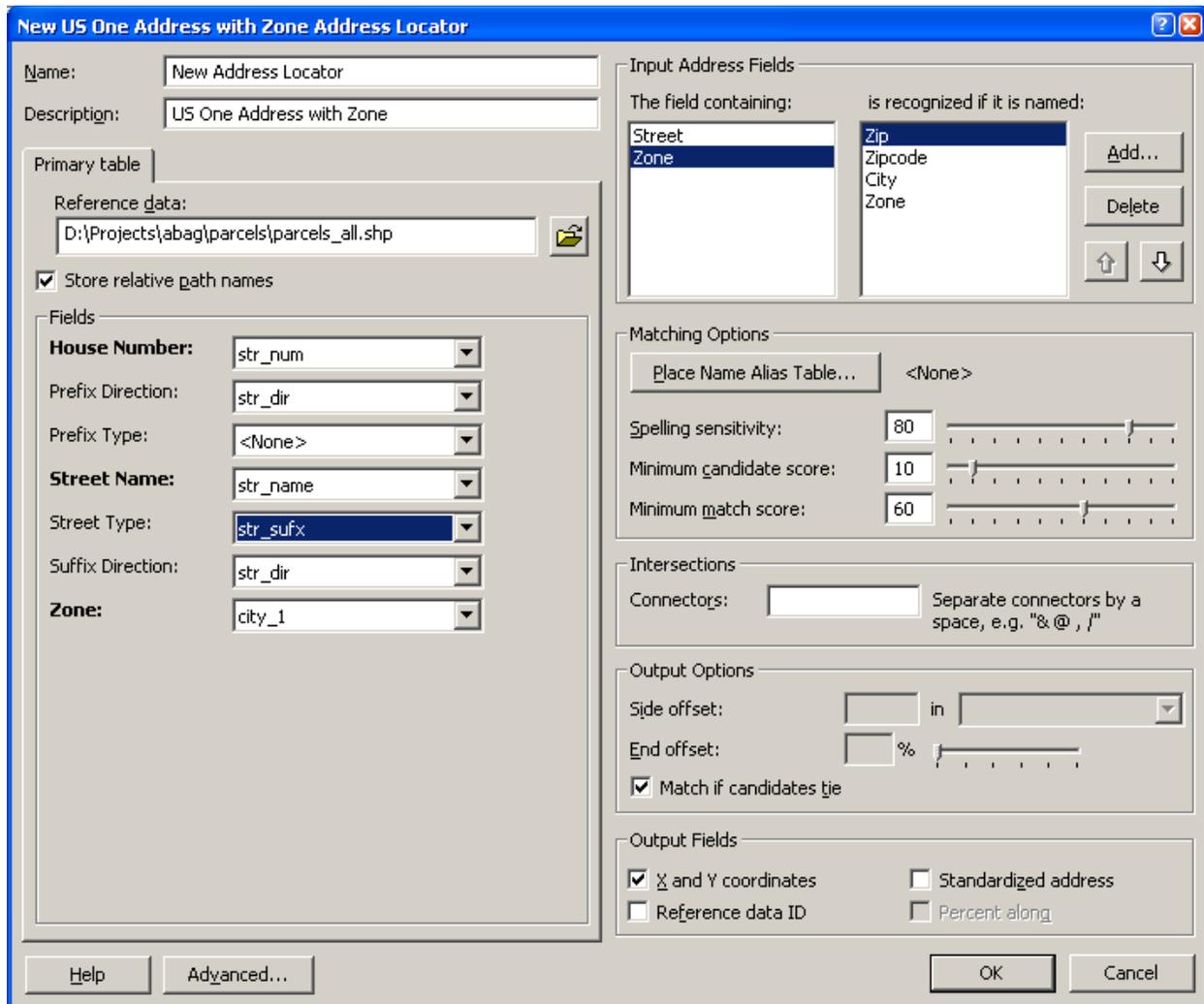
A review of the parcel data found a couple of consistent errors that reduced the parcel match rate in the geocoding step. These errors, and their fix, are described below.

- 1) The Sonoma county address information was not split into sub-components (i.e. house number, street direction, street name, street type, and city). The data was split into three fields using basic text processing functions in R: house number, street name, and street type. The complete address was split by whitespace and the house number was set to the first item, the street type was set to the last item, and the street name was set to what was in the middle of the two.
- 2) Some of the parcel city codes included " CA" and some other non-alpha characters such as "\*". These codes were removed from the parcel city field.
- 3) The parcel city value of "Unincorporated" and "Unincorp County" was replaced with "".
- 4) Some of the parcel city names were actually abbreviations instead of full city names. These city abbreviations were replaced with the actual city names in order to be consistent with the Nets and CoStar data.
- 5) Leading zeros were removed from street addresses in the parcel layer

## TeleAtlas Data Improvements

The initial TeleAtlas street layer was delivered as two shapefile – one for the bay area and one for the rest of California. These were merged into one layer inside a file geodatabase. This layer was named TANA (TeleAtlas North America) and was used for the address locator. However, it was learned through the geocoding process that the original TeleAtlas street layer outside of the bay area was corrupt. ABAG prepared a revised version of the TeleAtlas layers, this time in file geodatabase format. This revised layer was then used for geocoding.

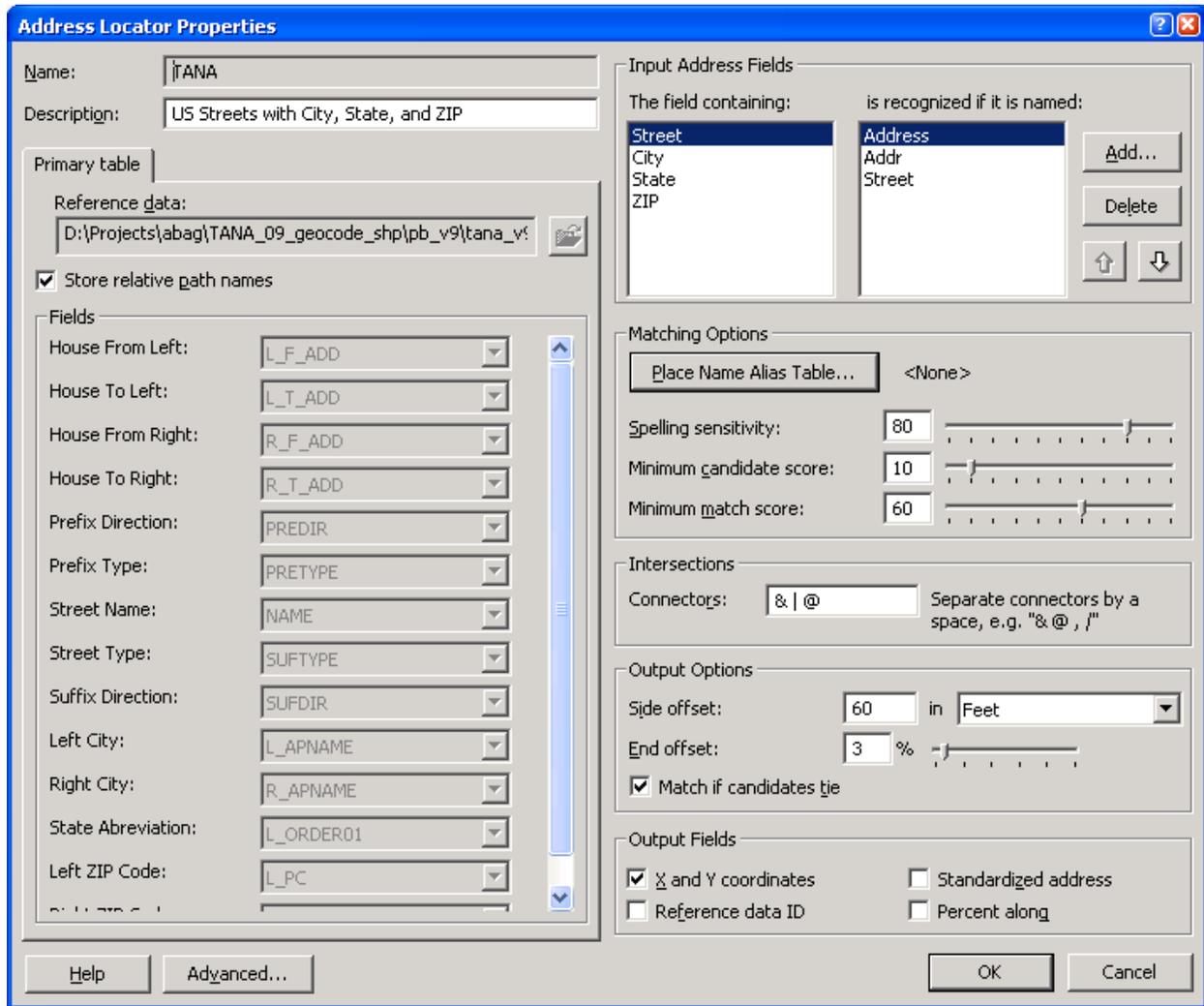
Figure 1 – Parcel Multiple Fields Address Locator



After creating the address locator, the \*.loc file was edited to add the following parameters to greatly improve the address locator performance. These parameters increase the amount of RAM that is used and the size of the caching (memory) used for pre-sorting the data for faster indexing.

RuntimeMemoryLimit = 2048000000  
 BatchPresortInputs = Zone  
 BatchPresortCacheSize = 100000

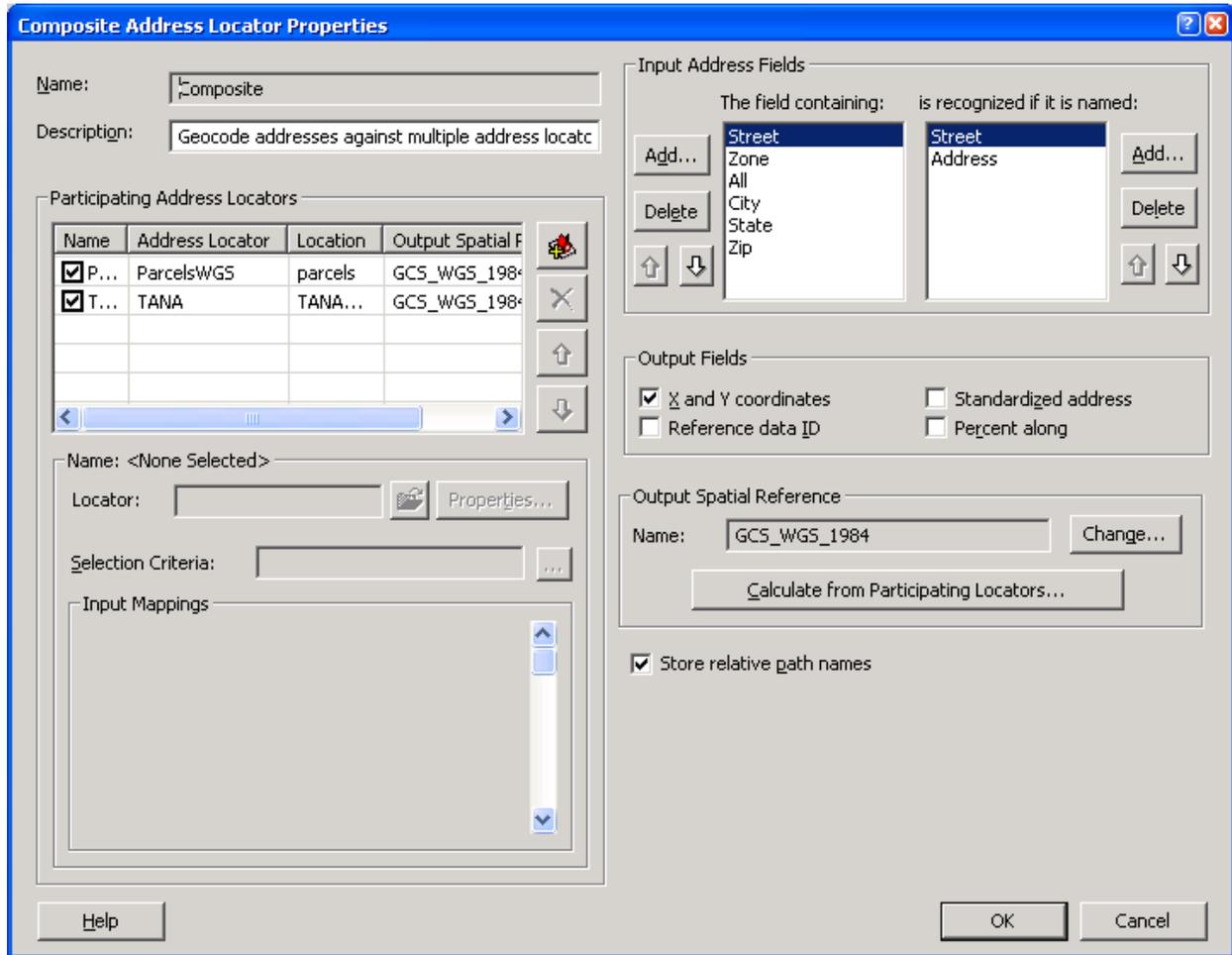
**Figure 2 – TeleAtlas Street Layer Address Locator**



After creating the address locator, the \*.loc file was edited to add the following parameters to greatly improve the address locator performance. These parameters increase the amount of RAM that is used and the size of the caching (memory) used for pre-sorting the data for faster indexing.

RuntimeMemoryLimit = 2048000000  
 BatchPresortInputs = State  
 BatchPresortInputs = City  
 BatchPresortInputs = ZIP  
 BatchPresortCacheSize = 100000

**Figure 3 - Composite Address Locator**



After creating the address locator, the \*.loc file was edited to add the following parameter to greatly improve the address locator performance. The parameter increases the amount of RAM available for geocoding.

RuntimeMemoryLimit = 2048000000

### ***Geocoding Nets and CoStar***

The geocoding step was split into nine separate processes in order to take advantage of the eight CPUs on the geocoding machine. Before creating the Nets year specific tables, an R script split the main Nets file into seven files each with about 750k records. These tables were then geocoded with the composite address locator. The other two geocoding processes were to geocode the Nets\_Moves data set and the entire CoStar data set. The eight Nets geocoding processes were run simultaneously and completed within one working day. The entire CoStar data was geocoded on a separate day using the Building\_Address field as the street address. The match rates are shown in **Error! Reference source not found.** below.

**Table 1 – Geocoding Match Rates**

File	Matched	Tied	Unmatched	Speed	Match Rate (Matched+Tied)
Nets 1	620028	55056	74916	407k/hr	90%
Nets 2	621839	53807	74354	430k/hr	90%
Nets 3	533051	48606	68343	435k/hr	89%
Nets 4	606476	56377	87147	379k/hr	89%
Nets 5	565020	42334	142646	368k/hr	81%
Nets 6	576511	43711	129778	344k/hr	83%
Nets 7	622515	49109	143432	353k/hr	82%
Nets_Moves	438683	43372	105236	358k/hr	82%
CoStar	98563	21017	2573	2040k/hr	98%

When the geocoding completed, the following fields were added to the Nets and CoStar data sets:

- 1) Loc\_name – matched address locator type (parcels or TANA (TeleAtlas))
- 2) Status – match status (M=match, T=tied (matched multiple candidates), U=unmatched)
- 3) Score – match quality (0-100 [best])
- 4) X – x coordinate (longitude)
- 5) Y – y coordinate (latitude)
- 6) Match\_addr – cleaned up address
- 7) Side – side of street matched if TANA

### ***Spatial Joins***

After the geocoding step, the nine Nets and CoStar geocoding layers were spatial joined to the parcel and TAZ layers in order to assign parcel IDs and TAZ numbers to the Nets and CoStar records based on the spatial location. This was especially important for records which matched to the TANA layer since these records were not assigned a parcel ID, but were offset from the street based on the locator offset settings (currently 20 feet). The spatial join process created two new fields in the Nets and CoStar data sets:

- 1) sj\_puid – spatial join parcel ID
- 2) sj\_taz – spatial join TAZ number

### **Creation of Yearly Nets Tables**

The geocoded Nets and Nets\_Moves tables were then read into R and processed to create yearly snapshots of establishments tables. The result of this processing step was twenty year tables, each with the establishments operating in the year. The data processing steps were as follows:

- 1) Join ABAG SICGroup codes to the Nets and Nets\_Moves tables by SIC Code
- 2) Calculate a number of years the establishment exists based off of the lastYear and FirstYear field.

- 3) Expand the Nets table so there is a record for each year an establishment exists and create a primary key from the DunsNumber (establishment ID) and year.
- 4) Create a primary key in the Nets\_Moves table from the DunsNumber (establishment ID) and move year.
- 5) Match the primary keys between the two tables and update the Nets records based on the Nets\_Moves records. Since the Nets table has the current (more recent) attributes of the establishment, the setting of the historic attributes was done in reverse, i.e. the previous location was set to be the origin of the record in the Nets\_Moves table for the move year. For example, an establishment exists from 1995 to 2005 and moved in 2001. The expanded Nets years table then has the current Nets table location from 2002 to 2005 and the origin location in the Nets\_Moves table from 1995 to 2001. The SIC code, address, city, state, zip code, FIPS county code, latitude, longitude, levelcode (Nets geocoding quality code), SICGroup, sj\_puid, and sj\_taz attributes were all set based on the Nets\_Moves data.
- 6) The expanded Nets table was written out to a series of CSV files with the following naming convention – nets\_year\_<year>.csv. These yearly tables were then read into the MySQL database for later analysis.

## Nets Data Summaries

Using the Nets year tables, a number of data summaries and maps were produced to review the data sets and the geocoding efforts for quality and completeness across multiple dimensions.

### ***Quality of Match***

The count of quality of match of each record in the dataset was determined by analyzing the loc\_name field. The loc\_name field indicates the layer that matched with Nets data. The sum of records matching each type of layer in year 1989 was calculated using the following query:

```
SELECT loc_name,year,count('loc_name') as num FROM nets_geocode_spatialjoin_1989 n group by loc_name;
```

The share of match rate by locator type is similar in each year, as shown in the year result tables below. The match rates are pretty consistent across years.

**Table 2 - Nets Records by Geocode Quality for Year 1989**

<b>Loc_name</b>	<b>year</b>	<b>number of establishments (num)</b>	<b>percent</b>
<no match>	1989	197996	16%
ParcelsWGS	1989	259074	21%
TANA	1989	772491	63%

**Table 3 - Nets Records by Geocode Quality for Year 1994**

<b>Loc_name</b>	<b>year</b>	<b>number of establishments (num)</b>	<b>percent</b>
<no match>	1994	249789	15%
ParcelsWGS	1994	351863	21%
TANA	1994	1086152	64%

**Table 4 - Nets Records by Geocode Quality for Year 1999**

<b>Loc_name</b>	<b>year</b>	<b>number of establishments (num)</b>	<b>percent</b>
<no match>	1999	245911	14%
ParcelsWGS	1999	371712	21%
TANA	1999	1141130	65%

**Table 5 - Nets Records by Geocode Quality for Year 2004**

<b>Loc_name</b>	<b>year</b>	<b>number of establishments (num)</b>	<b>percent</b>
<no match>	2004	312734	13%
ParcelsWGS	2004	482832	20%
TANA	2004	1578164	67%

**Table 6 - Nets Records by Geocode Quality for Year 2009**

<b>Loc_name</b>	<b>year</b>	<b>number of establishments (num)</b>	<b>percent</b>
<no match>	2009	359706	12%
ParcelsWGS	2009	583850	20%
TANA	2009	1950137	68%

The total number of Nets records matching either the Parcel layer or TeleAtlas layer was determined by creating a temporary table named “quality\_match\_all”, which stored the results of the quality of match query for all years. The following query, which is abbreviated for ease-of-display, was executed:

```
CREATE TABLE quality_match_all
SELECT loc_name,year,count('loc_name')as num FROM nets_geocode_spatialjoin_1989 n group by
loc_name
UNION
SELECT loc_name,year,count('loc_name')as num FROM nets_geocode_spatialjoin_1990 n group by
loc_name
UNION
SELECT loc_name,year,count('loc_name')as num FROM nets_geocode_spatialjoin_1991 n group by
loc_name
UNION
....
SELECT loc_name,year,count('loc_name')as num FROM nets_geocode_spatialjoin_2009 n group by
loc_name;
```

After creating the “quality\_match\_all” table, the quality of match for the entire Nets data was obtained from the following query:

```
SELECT loc_name,sum(num) FROM quality_match_all q group by loc_name;
```

**Table 7 – Total Nets Records by Geocode Quality**

<b>Loc_name</b>	<b>number of establishments (num)</b>	<b>percent</b>
<no match>	5786819	13%
Parcels	8635318	21%
TANA	27439258	66%

The number of 1989 establishments by bay area county that were geocoded to the Parcel layer, TeleAtlas layer or were not geocoded was obtained with the following query:

```
SELECT fipscounty,loc_name,year,count('loc_name')as num FROM nets_geocode_spatialjoin_1989 n
WHERE FIPSCounty IN (6001, 6013, 6041, 6055, 6075, 6081, 6085, 6095, 6097) group by fipscounty,
loc_name,;
```

As shown in the table below, the results show a somewhat significant difference in the parcel match rate by county. The results are similar for each year as well. The CoStar data, which is summarized later in this document, does not have the same magnitude of differences in parcel match rate by county. This suggests the quality of Nets data varies by county.

**Table 8 - Nets Records Geocode Quality by County for Year 1989**

County Name	County (fipscounty)	Loc_name	year	number of establishments (num)	percent
Alameda	6001	<no match>	1989	3104	6%
Alameda	6001	ParcelsWGS	1989	47950	89%
Alameda	6001	TANA	1989	3059	6%
Contra Costa	6013	<no match>	1989	2739	8%
Contra Costa	6013	ParcelsWGS	1989	27698	85%
Contra Costa	6013	TANA	1989	2333	7%
Marin	6041	<no match>	1989	1953	12%
Marin	6041	ParcelsWGS	1989	12957	77%
Marin	6041	TANA	1989	1873	11%
Napa	6055	<no match>	1989	429	8%
Napa	6055	ParcelsWGS	1989	4803	87%
Napa	6055	TANA	1989	268	5%
San Francisco	6075	<no match>	1989	2290	5%
San Francisco	6075	ParcelsWGS	1989	40668	92%
San Francisco	6075	TANA	1989	1419	3%
San Mateo	6081	<no match>	1989	2191	7%
San Mateo	6081	ParcelsWGS	1989	27754	86%
San Mateo	6081	TANA	1989	2265	7%
Santa Clara	6085	<no match>	1989	3850	6%
Santa Clara	6085	ParcelsWGS	1989	51587	85%
Santa Clara	6085	TANA	1989	5376	9%
Solano	6095	<no match>	1989	661	7%
Solano	6095	ParcelsWGS	1989	8365	91%
Solano	6095	TANA	1989	208	2%
Sonoma	6097	<no match>	1989	1703	9%
Sonoma	6097	ParcelsWGS	1989	15758	79%
Sonoma	6097	TANA	1989	2375	12%

The number of establishments by SICgroup in each bay area county in year 1989 that were geocoded was obtained with the following query:

```
SELECT fipscounty,year,sicgroup,count(sicgroup)as num FROM nets_geocode_spatialjoin_1989 n where fipscounty IN (6001, 6013, 6041, 6055, 6075, 6081, 6085, 6095, 6097) and loc_name ="PARCELSWGS" or loc_name = "TANA " group by fipscounty,sicgroup;
```

**Table 9 - Establishments in 1989 by SICGroup and County (incomplete list)**

County (fipscounty)	year	sicgroup	number of establishments (num)
6001	1989	ag	338
6001	1989	art_rec	554
6001	1989	constr	3778
6001	1989	eat	2316
6001	1989	ed_high	60
6001	1989	ed_k12	523
6001	1989	ed_oth	297
6001	1989	fire	3840
6001	1989	gov	114
6001	1989	health	2941
6001	1989	hotel	232
6001	1989	info	2142
6001	1989	lease	194
6001	1989	logis	5300
6001	1989	man_bio	30
6001	1989	man_hvy	1677
6001	1989	man_lgt	1590
6001	1989	man_tech	662
6001	1989	natres	31
6001	1989	prof	2546

The number of establishments by SICGroup in the nine county region that were geocoded for year 1989 was obtained with the following query:

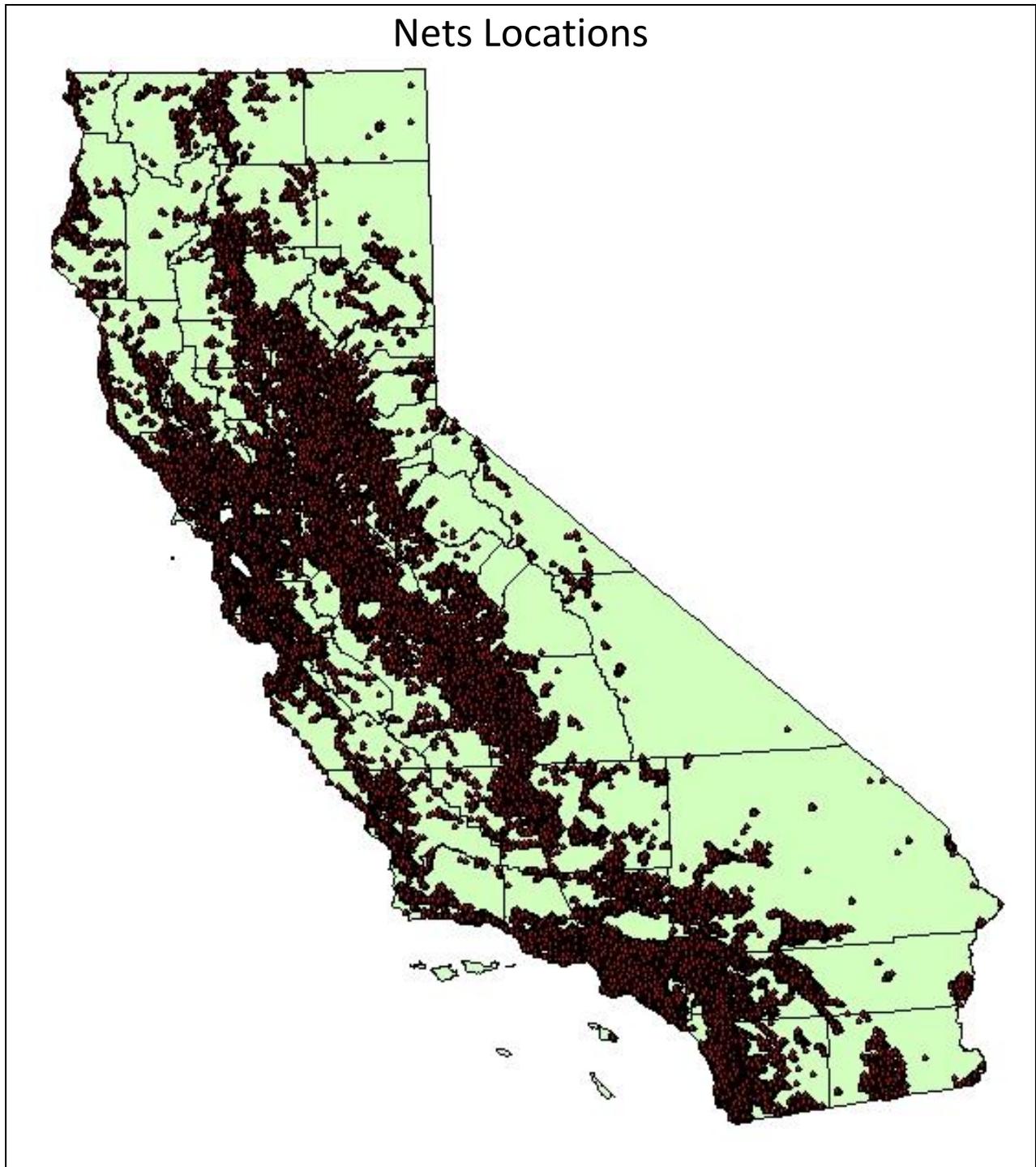
```
SELECT year,sicgroup,count(sicgroup)as num FROM nets_geocode_spatialjoin_1989 n where fipscounty IN (6001, 6013, 6041, 6055, 6075, 6081, 6085, 6095, 6097) and loc_name ="PARCELSWGS" or loc_name = "TANA " group by sicgroup;
```

**Table 10 - Establishments in 1989 by SICGroup (incomplete list)**

year	sicgroup	number of establishments (num)
1989	ag	20797
1989	art_rec	12176
1989	constr	84788
1989	eat	40511
1989	ed_high	662
1989	ed_k12	9908
1989	ed_oth	4842

### ***Net Geocode Locations Map***

The location of all the geocoded Nets establishments is shown in the map below. The map shows a good distribution of results across the state.



The number of establishments by SICGroup by year is shown in the tables that follow. The results show that the number of establishments increases every five years for just about every SICGroup. There are a few exceptions though, such as metal mining and colleges and universities.

**Table 11 - Agriculture Production Establishments**

year	sicgroup	number of establishments
1989	Agriculture Production	21174
1994	Agriculture Production	24727
1999	Agriculture Production	27866
2004	Agriculture Production	33351
2009	Agriculture Production	36291

**Table 12 - Motion Picture Theater Establishments**

year	sicgroup	number of establishments
1989	Motion Picture Theater	12504
1994	Motion Picture Theater	20713
1999	Motion Picture Theater	26355
2004	Motion Picture Theater	39190
2009	Motion Picture Theater	45973

**Table 13 - Construction Establishments**

year	sicgroup	number of establishments
1989	Construction	86049
1994	Construction	109387
1999	Construction	106092
2004	Construction	140945
2009	Construction	171457

**Table 14 - Eating and Drinking Place Establishments**

year	sicgroup	number of establishments
1989	Eating and Drinking Places	41523
1994	Eating and Drinking Places	50132
1999	Eating and Drinking Places	56576
2004	Eating and Drinking Places	65034
2009	Eating and Drinking Places	61025

**Table 15 - Colleges and Universities Establishments**

year	sicgroup	number of establishments
1989	Colleges and Universities	678
1994	Colleges and Universities	966
1999	Colleges and Universities	1264
2004	Colleges and Universities	1567
2009	Colleges and Universities	1459

**Table 16 - Elementary and Secondary School Establishments**

year	sicgroup	number of establishments
1989	Elementary and Secondary Schools	10163
1994	Elementary and Secondary Schools	11181
1999	Elementary and Secondary Schools	11711
2004	Elementary and Secondary Schools	12019
2009	Elementary and Secondary Schools	11273

**Table 17 - Library Establishments**

year	sicgroup	number of establishments
1989	Libraries	4966
1994	Libraries	7630
1999	Libraries	8957
2004	Libraries	12063
2009	Libraries	14500

**Table 18 - Depository Establishments**

year	sicgroup	number of establishments
1989	Depository Institutions	90809
1994	Depository Institutions	128624
1999	Depository Institutions	136258
2004	Depository Institutions	204703
2009	Depository Institutions	251924

**Table 19 - Executive, Legislative & General Government Establishments**

year	sicgroup	number of establishments
1989	Executive, Legislative & General Government	3057
1994	Executive, Legislative & General Government	4906
1999	Executive, Legislative & General Government	7335
2004	Executive, Legislative & General Government	10494
2009	Executive, Legislative & General Government	10566

**Table 20 - Health Service Establishments**

year	sicgroup	number of establishments
1989	Health Services	58670
1994	Health Services	89219
1999	Health Services	88131
2004	Health Services	116027
2009	Health Services	134314

**Table 21 - Hotels, Camps, and Other Lodging Places**

year	sicgroup	number of establishments
1989	Hotels, Camps, and Other Lodging Places	7187
1994	Hotels, Camps, and Other Lodging Places	8568
1999	Hotels, Camps, and Other Lodging Places	9242
2004	Hotels, Camps, and Other Lodging Places	10607
2009	Hotels, Camps, and Other Lodging Places	12811

**Table 22 - Information Services Establishments**

year	sicgroup	number of establishments
1989	Information Services	44495
1994	Information Services	70140
1999	Information Services	97314
2004	Information Services	127156
2009	Information Services	130362

**Table 23 - Misc. Equipment Rental & Leasing Establishments**

year	sicgroup	number of establishments
1989	Misc. Equipment Rental & Leasing	5126
1994	Misc. Equipment Rental & Leasing	5832
1999	Misc. Equipment Rental & Leasing	5847
2004	Misc. Equipment Rental & Leasing	7696
2009	Misc. Equipment Rental & Leasing	8889

**Table 24 - Logistical Service Establishments**

year	sicgroup	number of establishments
1989	Logistical Services	98085
1994	Logistica l Services	130913
1999	Logistical Services	133047
2004	Logistical Services	158367
2009	Logistical Services	178657

**Table 25 - Drug Establishments**

year	sicgroup	number of establishments
1989	Drugs	525
1994	Drugs	663
1999	Drugs	971
2004	Drugs	1071
2009	Drugs	1272

**Table 26 Saw mill and Planing mill Establishments**

year	sicgroup	number of establishments
1989	Sawmills and Planing Mills	32191
1994	Sawmills and Planing Mills	34923
1999	Sawmills and Planing Mills	35866
2004	Sawmills and Planing Mills	38865
2009	Sawmills and Planing Mills	40006

**Table 27 - Food and Kindred Products**

<b>year</b>	<b>sicgroup</b>	<b>number of establishments</b>
1989	Food and Kindred Products	32947
1994	Food and Kindred Products	40409
1999	Food and Kindred Products	42187
2004	Food and Kindred Products	48981
2009	Food and Kindred Products	49060

**Table 28 - Computer and Office Equipment Establishments**

<b>year</b>	<b>sicgroup</b>	<b>number of establishments</b>
1989	Computer and Office Equipment	11197
1994	Computer and Office Equipment	12992
1999	Computer and Office Equipment	14676
2004	Computer and Office Equipment	15236
2009	Computer and Office Equipment	15591

**Table 29 - Metal Mining Establishments**

<b>year</b>	<b>sicgroup</b>	<b>number of establishments</b>
1989	Metal Mining	1622
1994	Metal Mining	1611
1999	Metal Mining	1582
2004	Metal Mining	1842
2009	Metal Mining	1750

**Table 30 - Professional & Managerial Service Establishments**

<b>year</b>	<b>sicgroup</b>	<b>number of establishments</b>
1989	Professional & Managerial Services	50349
1994	Professional & Managerial Services	83168
1999	Professional & Managerial Services	82866
2004	Professional & Managerial Services	89058
2009	Professional & Managerial Services	103314

**Table 31 - Lumber and Other Building Material Establishments**

<b>year</b>	<b>sicgroup</b>	<b>number of establishments</b>
1989	Lumber and Other Building Materials	110834
1994	Lumber and Other Building Materials	147259
1999	Lumber and Other Building Materials	148816
2004	Lumber and Other Building Materials	182672
2009	Lumber and Other Building Materials	189159

**Table 32 - Regular Retail Establishments**

<b>year</b>	<b>sicgroup</b>	<b>number of establishments</b>
1989	Regular Retail	56550
1994	Regular Retail	76239
1999	Regular Retail	72734
2004	Regular Retail	94023
2009	Regular Retail	102301

**Table 33 - Business Service Establishments**

<b>year</b>	<b>sicgroup</b>	<b>number of establishments</b>
1989	Business Services	83917
1994	Business Services	137831
1999	Business Services	149254
2004	Business Services	309453
2009	Business Services	563170

**Table 34 - Personal Services Establishments**

<b>year</b>	<b>sicgroup</b>	<b>number of establishments</b>
1989	Personal Services	133015
1994	Personal Services	187313
1999	Personal Services	185259
2004	Personal Services	257002
2009	Personal Services	298936

**Table 35 - Social Service Establishments**

<b>year</b>	<b>sicgroup</b>	<b>number of establishments</b>
1989	Social Services	17564
1994	Social Services	28450
1999	Social Services	34275
2004	Social Services	45500
2009	Social Services	49682

**Table 36 - Transportation Establishments**

<b>year</b>	<b>sicgroup</b>	<b>number of establishments</b>
1989	Transportation	11967
1994	Transportation	18003
1999	Transportation	20545
2004	Transportation	27869
2009	Transportation	36965

**Table 37 - Unclassified Establishments**

<b>year</b>	<b>sicgroup</b>	<b>number of establishments</b>
1989	unclassified	302
1994	unclassified	546
1999	unclassified	1411
2004	unclassified	2696
2009	unclassified	4984

**Table 38 - Utilities Establishments**

<b>year</b>	<b>sicgroup</b>	<b>number of establishments</b>
1989	utilities	4099
1994	utilities	5670
1999	utilities	6405
2004	utilities	7509
2009	utilities	8296

## CoStar Data Summaries

A number of data summaries and maps were produced to review the CoStar data sets and the geocoding efforts for quality and completeness across multiple dimensions.

### *Quality of Match*

A summary of the quality of match was obtained using the following query:

```
SELECT loc_name,count('loc_name')as num FROM costar_geocode_spatialjoin c group by loc_name;
```

**Table 39 – CoStar Records by Geocode Quality**

Loc_name	number of establishments (num)	percent
<no match>	2573	2%
Parcels	113247	93%
TANA	6333	5%

As shown in the table below, the parcel match rate by county varies, but not as much as for the Nets data. This suggests that the CoStar data has more reliable geographic attributes.

**Table 40 - CoStar Records Geocode Quality by County**

County Name	County (fipscounty)	Loc_name	number of establishments (num)	percent
Alameda	6001	<no match>	744	2%
Alameda	6001	ParcelsWGS	29158	94%
Alameda	6001	TANA	1307	4%
Contra Costa	6013	<no match>	318	3%
Contra Costa	6013	ParcelsWGS	10656	88%
Contra Costa	6013	TANA	1030	9%
Marin	6041	<no match>	158	4%
Marin	6041	ParcelsWGS	3580	88%
Marin	6041	TANA	304	8%
Napa	6055	<no match>	111	5%
Napa	6055	ParcelsWGS	2269	92%
Napa	6055	TANA	79	3%
San Francisco	6075	<no match>	282	1%
San Francisco	6075	ParcelsWGS	20560	98%
San Francisco	6075	TANA	224	1%
San Mateo	6081	<no match>	148	1%
San Mateo	6081	ParcelsWGS	12595	95%

San Mateo	6081	TANA	487	4%
Santa Clara	6085	<no match>	455	2%
Santa Clara	6085	ParcelsWGS	25886	91%
Santa Clara	6085	TANA	2080	7%
Sonoma	6097	<no match>	357	4%
Sonoma	6097	ParcelsWGS	8543	88%
Sonoma	6097	TANA	822	8%

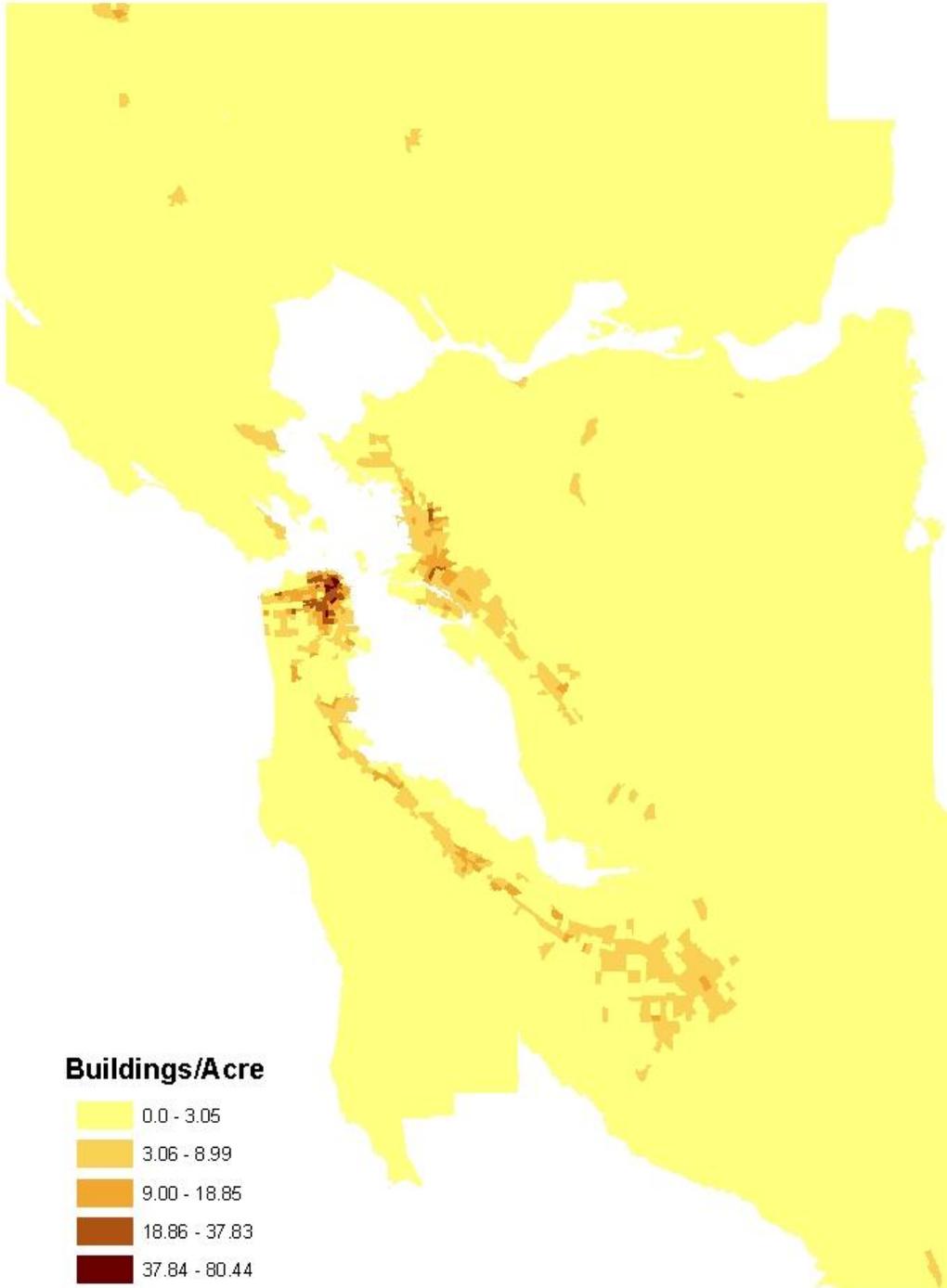
The following query was executed to count the number of buildings by TAZ. The number of buildings by TAZ, normalized by TAZ acres, is shown in the map that follows. The map nicely illustrates the dense urban areas of the region.

*SELECT sj\_taz,count(building\_a)as num FROM costar\_geocode\_spatialjoin c group by sj\_taz;*

**Table 41 - Number of Buildings by TAZ (incomplete list)**

TAZ (sj_taz)	number of buildings (num)
0 (no match)	2647
1	38
2	83
3	55
4	44
5	180
6	161
7	158
8	158
9	230
10	196
11	192
12	87
13	66
14	26
15	22
16	147
17	173
18	219
19	204
20	368

## Number of Buildings



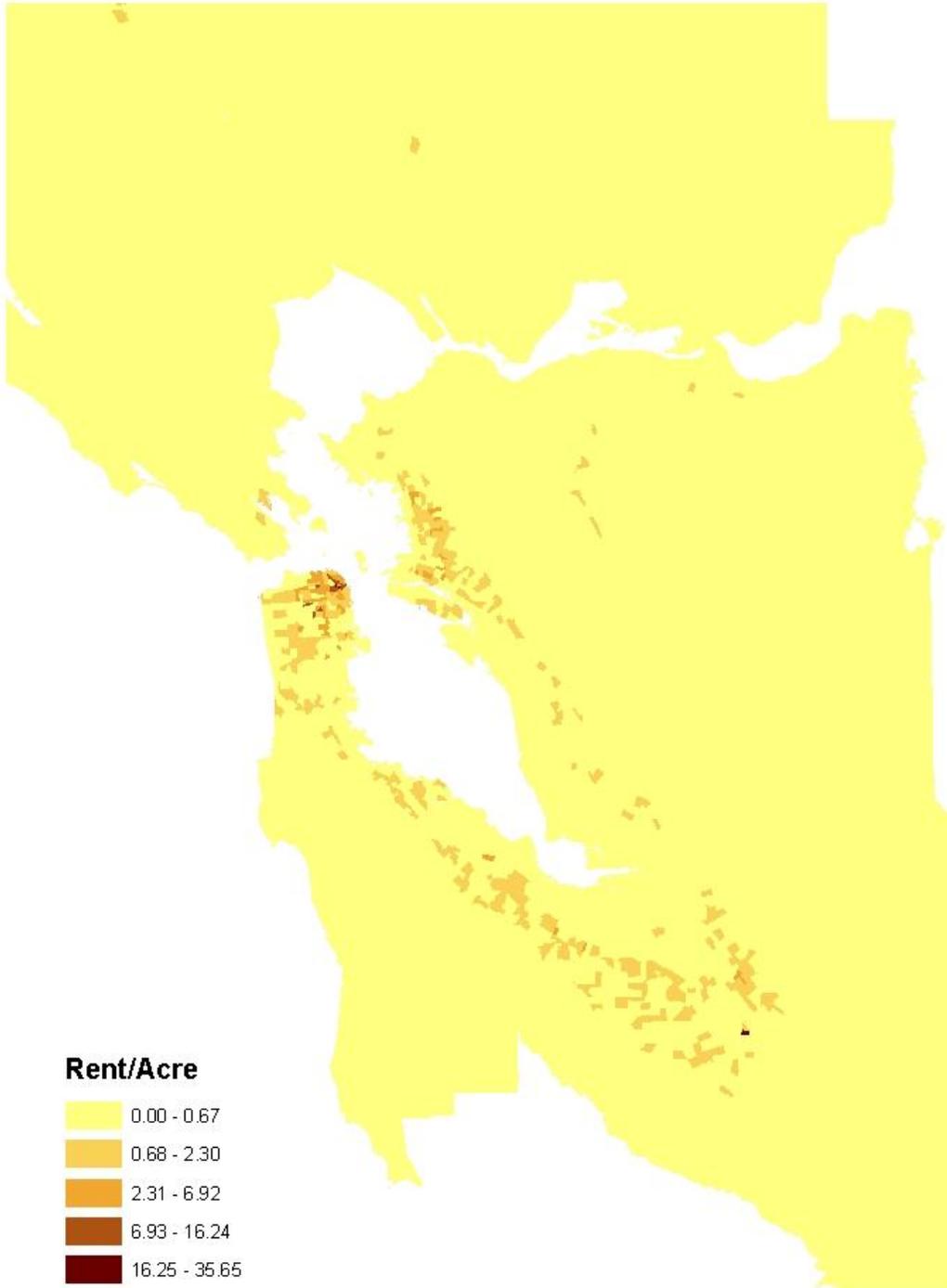
In addition to number of buildings by TAZ, average rents by TAZ were obtained by running the following query. A map of the results, again normalized by TAZ acres, shows a similar pattern to the previous map.

*SELECT sj\_taz,average\_we FROM costar\_geocode\_spatialjoin c where average\_we>0 group by sj\_taz;*

**Table 42 - TAZ by Average Rent (incomplete list)**

<b>TAZ (sj_taz)</b>	<b>average weighted rent (average_we)</b>
0 (no match)	7.45
1	22
2	20
3	30
4	45
5	75
6	21
7	50.38
8	16
9	24
10	15
11	10.8
12	12
13	21.5
14	35
15	28.78
16	12
17	15
18	24.36
19	12
20	13.2
21	18
22	30.54
23	29.76
24	35
25	32.04
26	38.32
29	25.98
30	18
31	18.24

## Average Weighted Rent



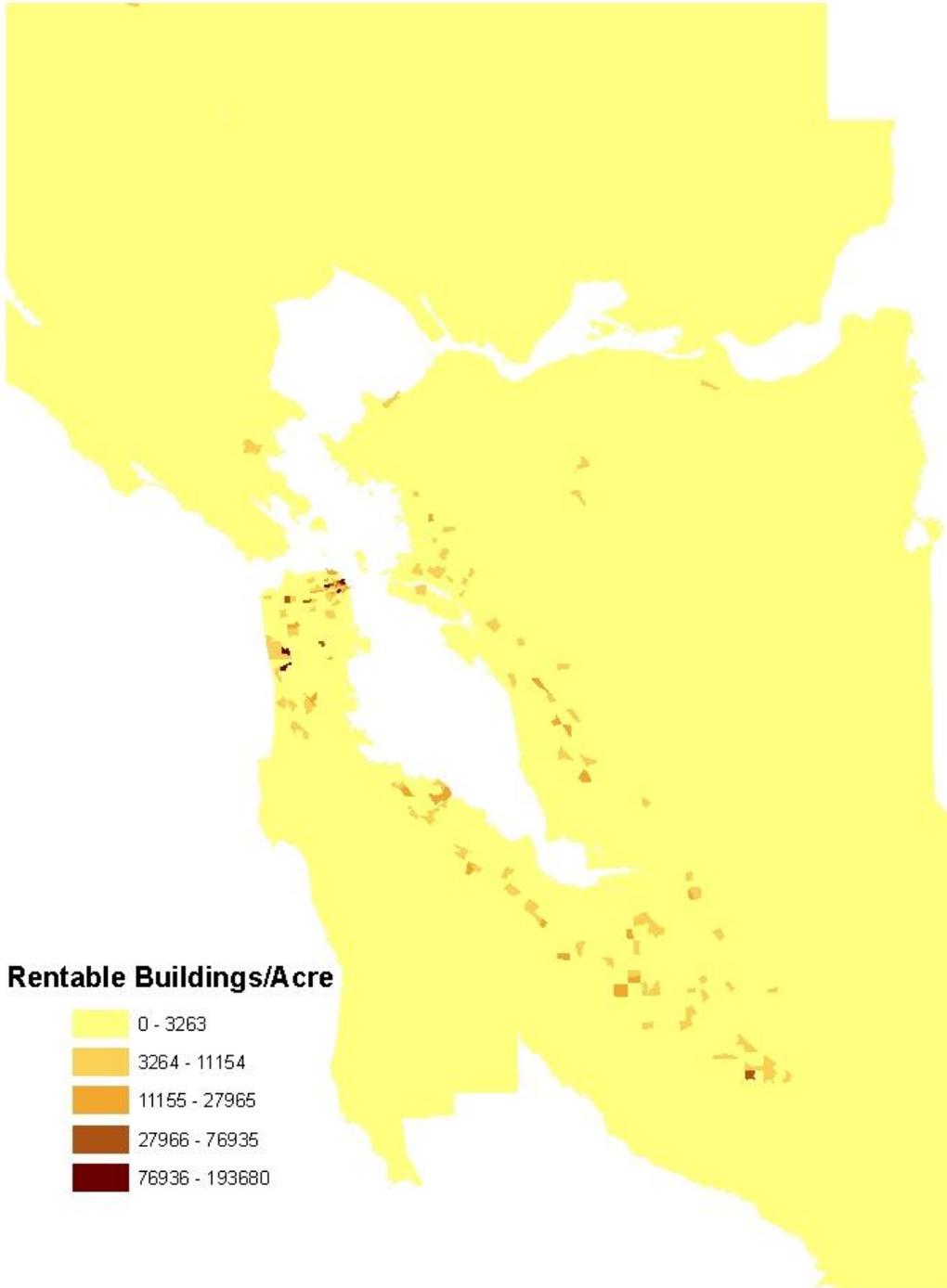
Total square feet by TAZ was obtained by the following query. Total square feet by TAZ normalized by TAZ acres is shown in the map below. The spatial pattern of total square feet is similar to the previous measures but not quite as pronounced.

*SELECT sj\_taz,rentable\_b FROM costar\_geocode\_spatialjoin c group by sj\_taz;*

**Table 43 - TAZ by Total Square Feet (incomplete list)**

<b>TAZ (sj_taz)</b>	<b>Rentable Buildings (rentable_b)</b>
0 (no match)	12771
1	105840
2	37895
3	49634
4	7000
5	16590
6	22062
7	3420
8	125088
9	56320
10	45945
11	66700
12	490000
13	14260
14	11000
15	31400
16	320256
17	24000
18	10000
19	105864
20	6000
21	15106
22	863441
23	4809
24	281581
25	36700
26	12256
27	36642
28	212000
29	38088
30	11609

# Total Square Feet



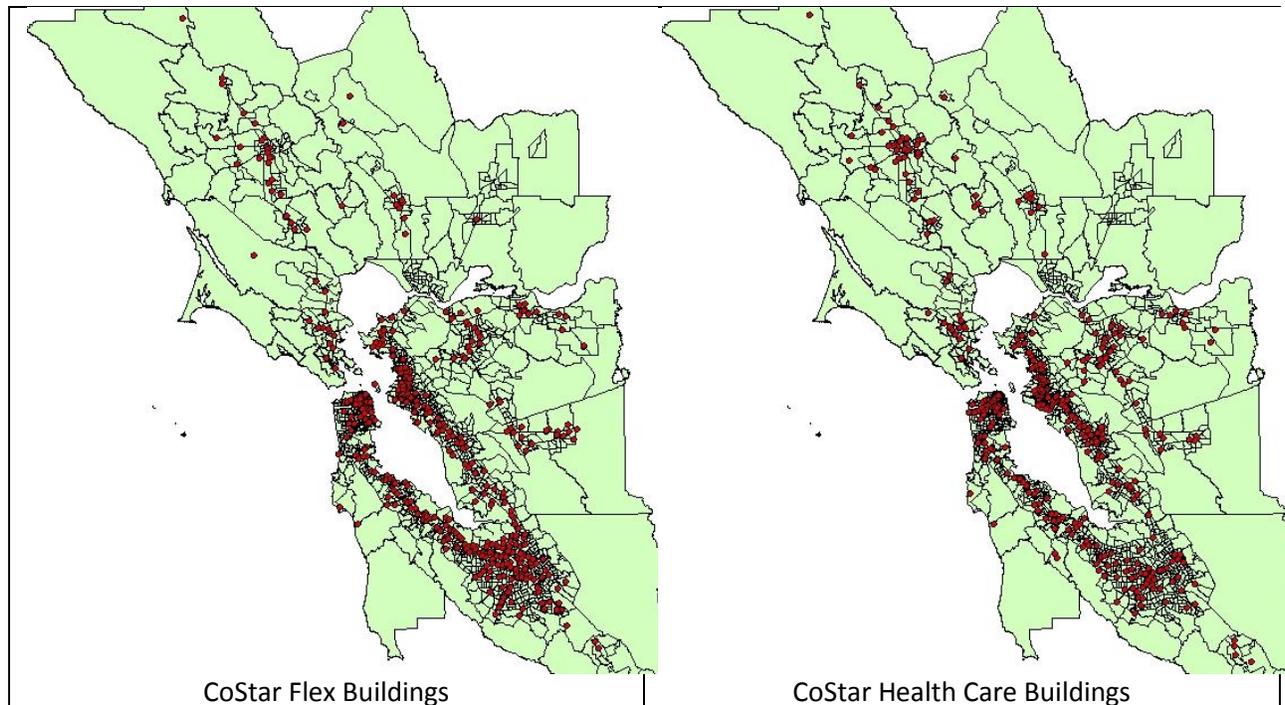
## Maps of Buildings by Building Type

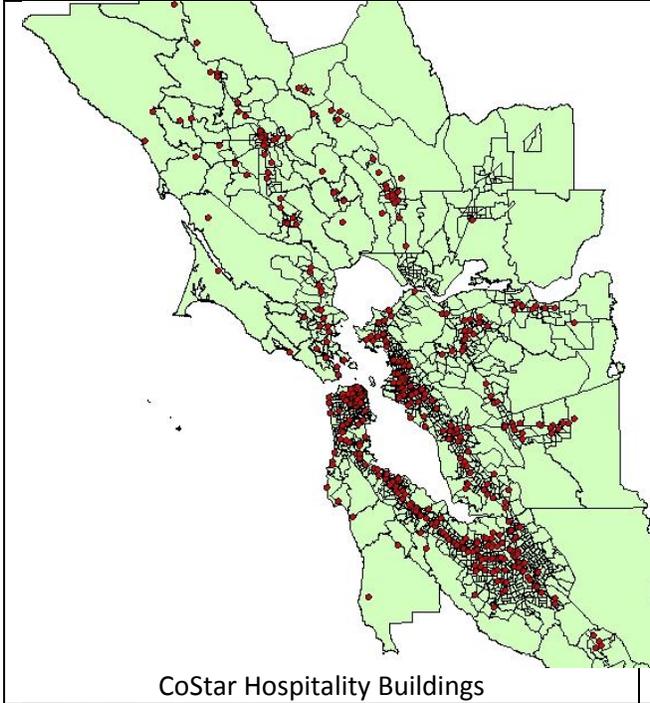
The following queries were executed to get CoStar buildings by building type in order to generate the maps that follow:

```
SELECT * FROM costar_geocode_spatialjoin c where propertytype="Flex";  
SELECT * FROM costar_geocode_spatialjoin c where propertytype="Health Care"  
SELECT * FROM costar_geocode_spatialjoin c where propertytype="Hospitality";  
SELECT * FROM costar_geocode_spatialjoin r c where propertytype="Industrial";  
SELECT * FROM costar_geocode_spatialjoin c where propertytype="Land";  
SELECT * FROM costar_geocode_spatialjoin c where propertytype="Multi-Family";  
SELECT * FROM costar_geocode_spatialjoin c where propertytype="Office";  
SELECT * FROM costar_geocode_spatialjoin c where propertytype="Retail";  
SELECT * FROM costar_geocode_spatialjoin c where propertytype="Specialty";  
SELECT * FROM costar_geocode_spatialjoin c where propertytype="Sports and Entertainment";
```

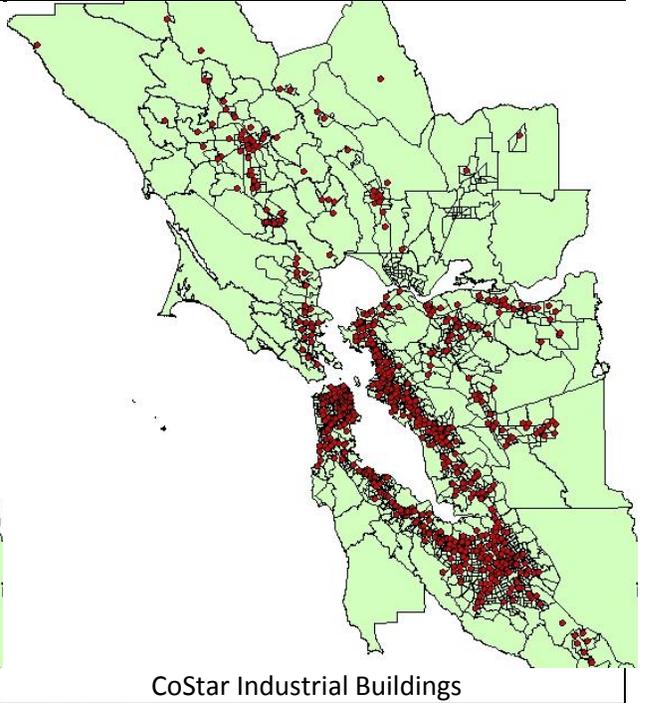
Note that the CoStar property type maps also include sub classifications of each property type. For example, property type "Flex" includes Flex, Flex (community center), Flex (neighborhood center) and Flex (strip center).

The maps show a good spatial distribution of buildings by build type.

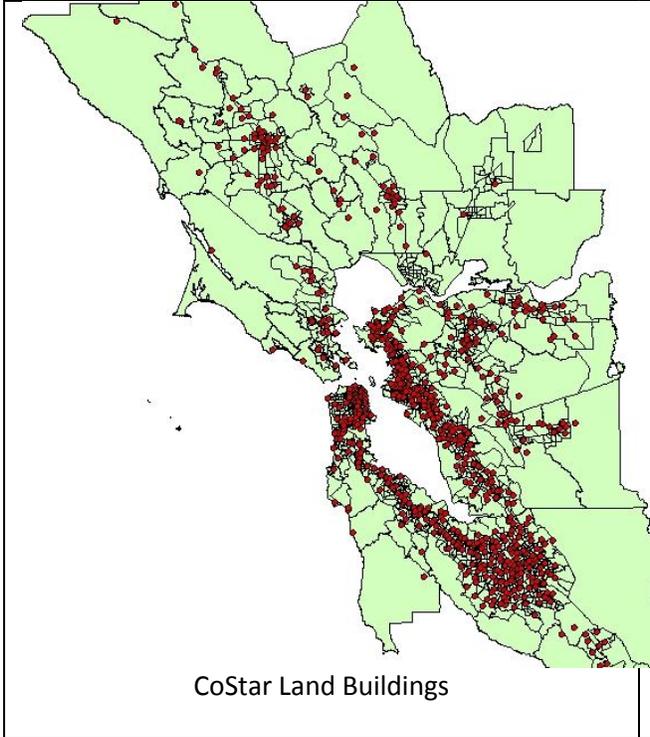




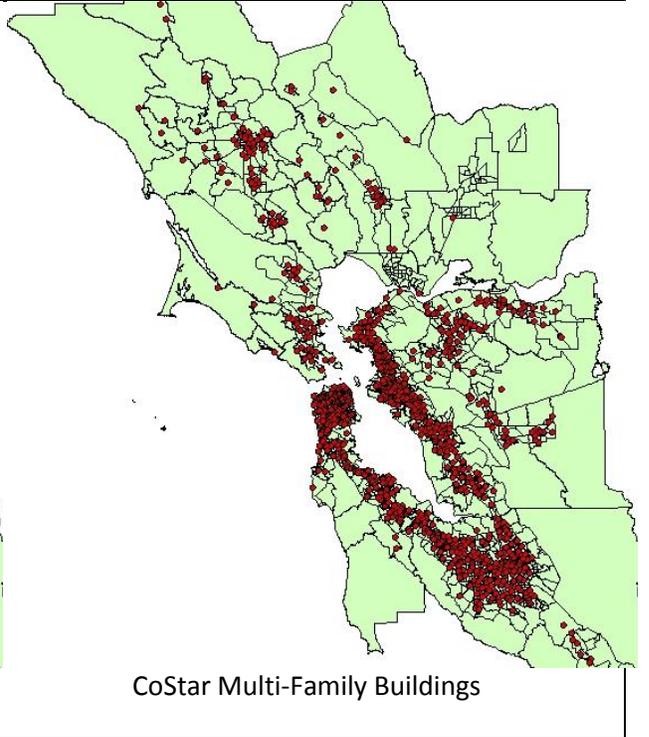
CoStar Hospitality Buildings



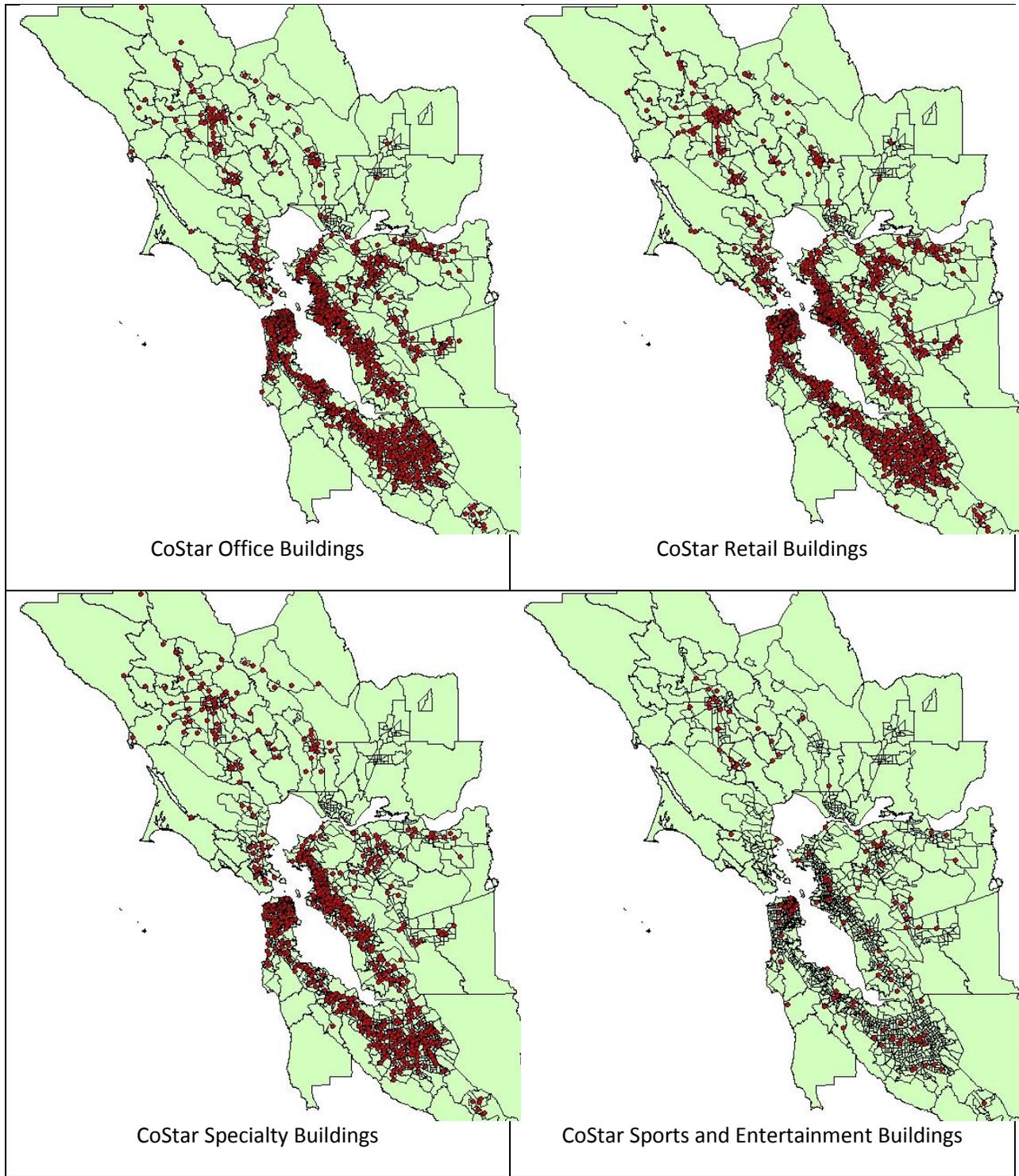
CoStar Industrial Buildings



CoStar Land Buildings

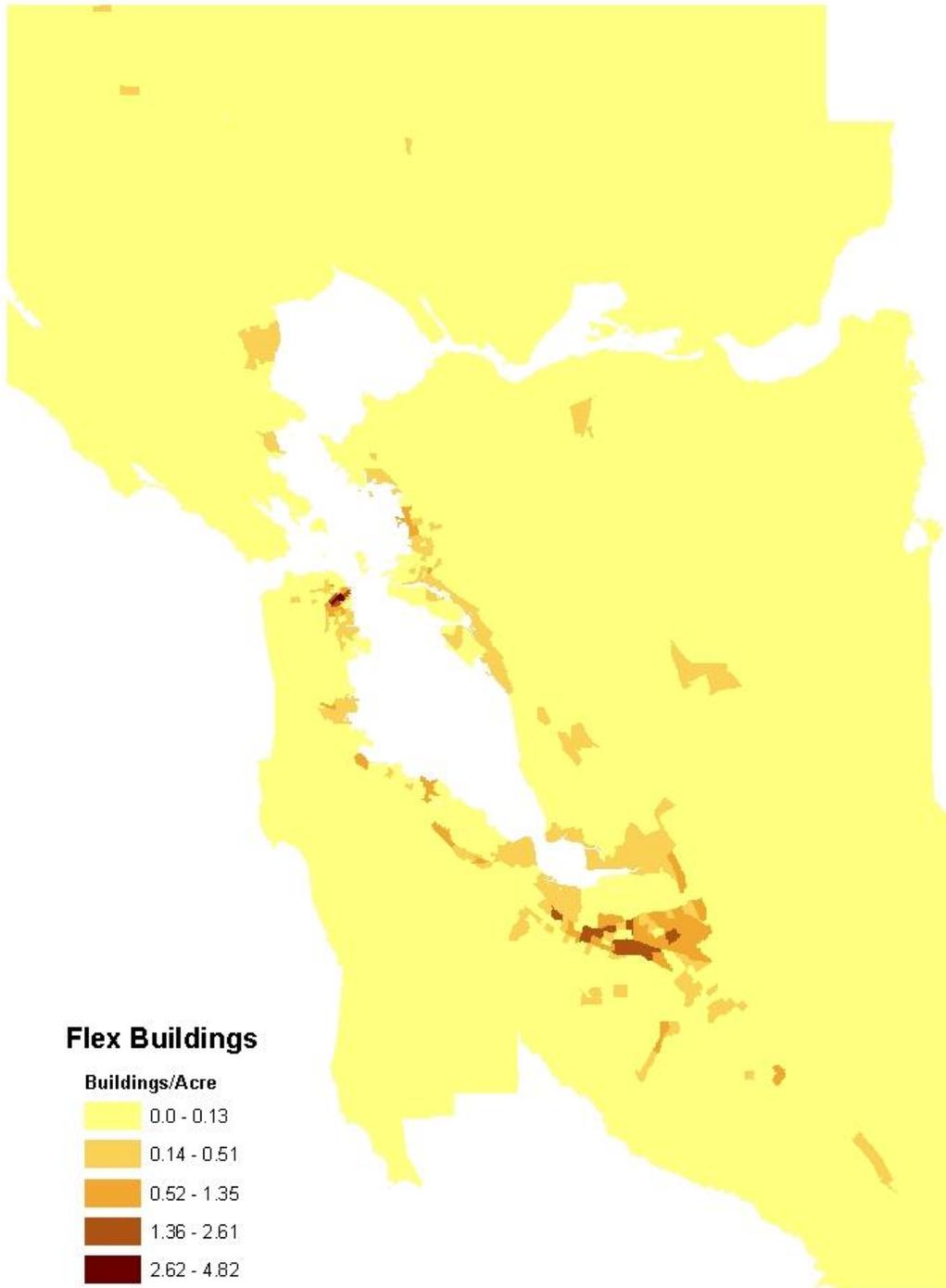


CoStar Multi-Family Buildings

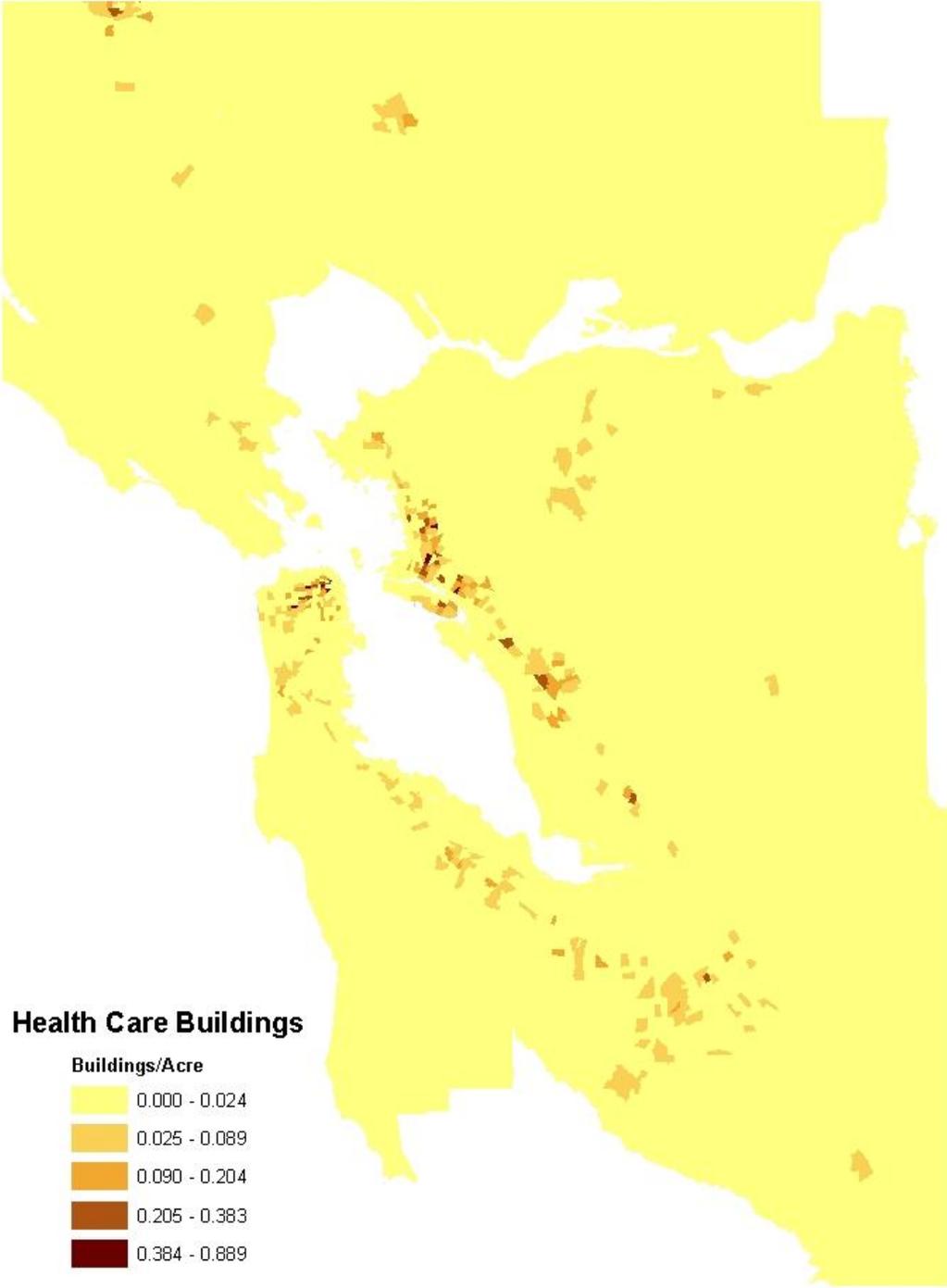


The CoStar point features were spatial joined to the TAZ layer and a map produced for each building type that summarizes the count of buildings by TAZ, normalized by TAZ acres. The spatial distribution of buildings by building type looks reasonable, with different building types showing greater concentrations in certain areas as opposed to others.

# CoStar Flex Buildings



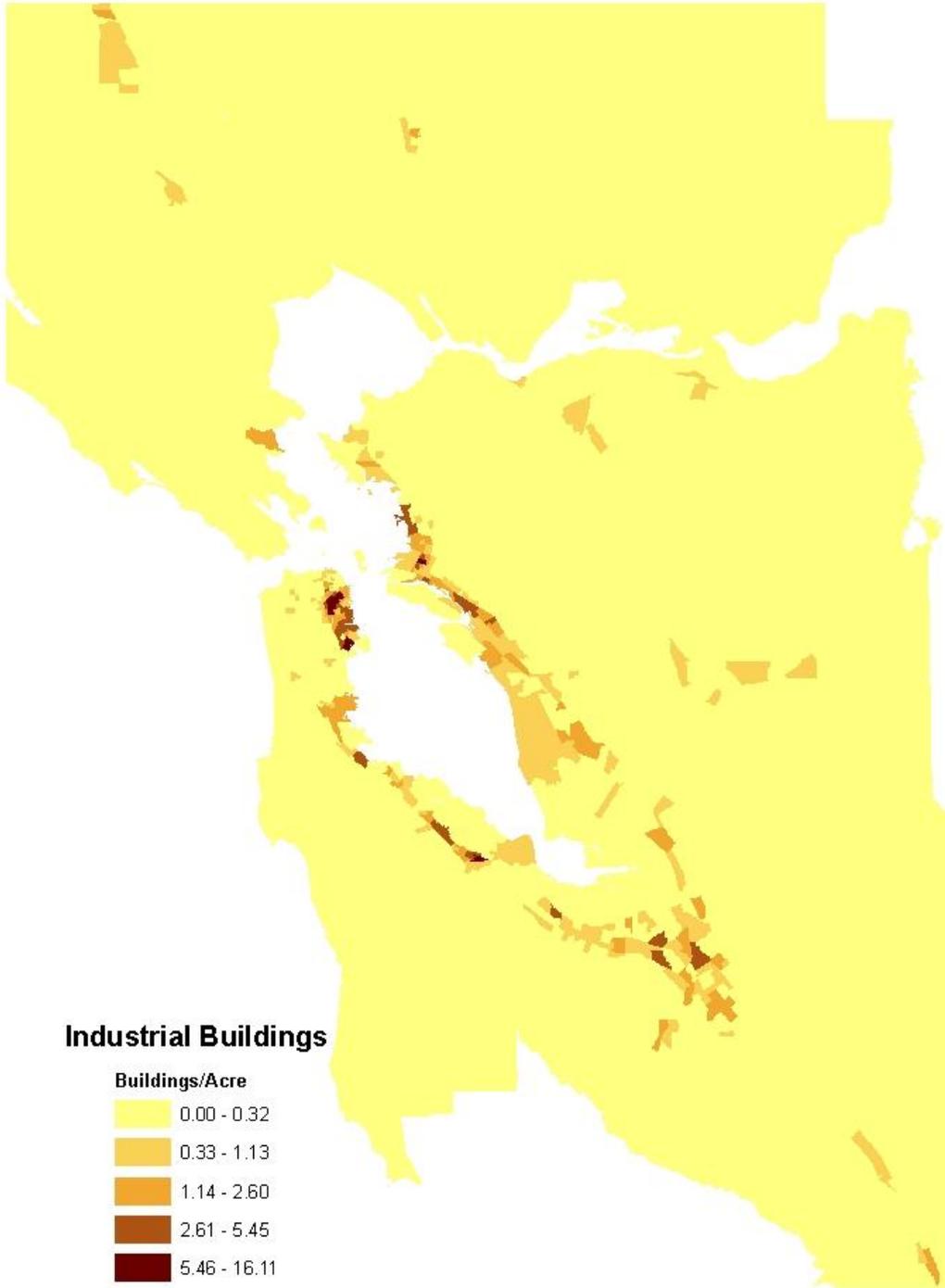
# CoStar Health Care Buildings



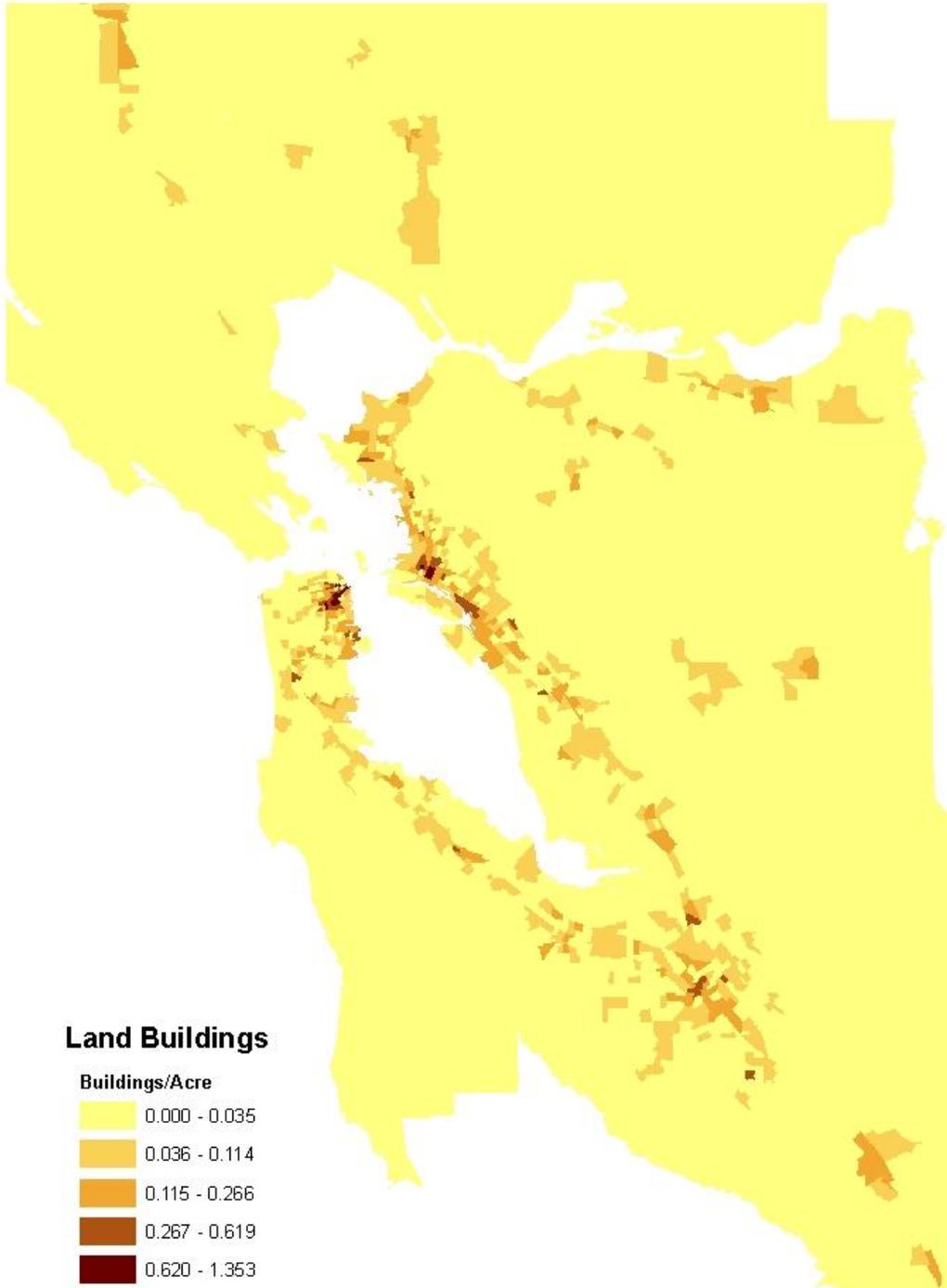
## CoStar Hospitality Buildings



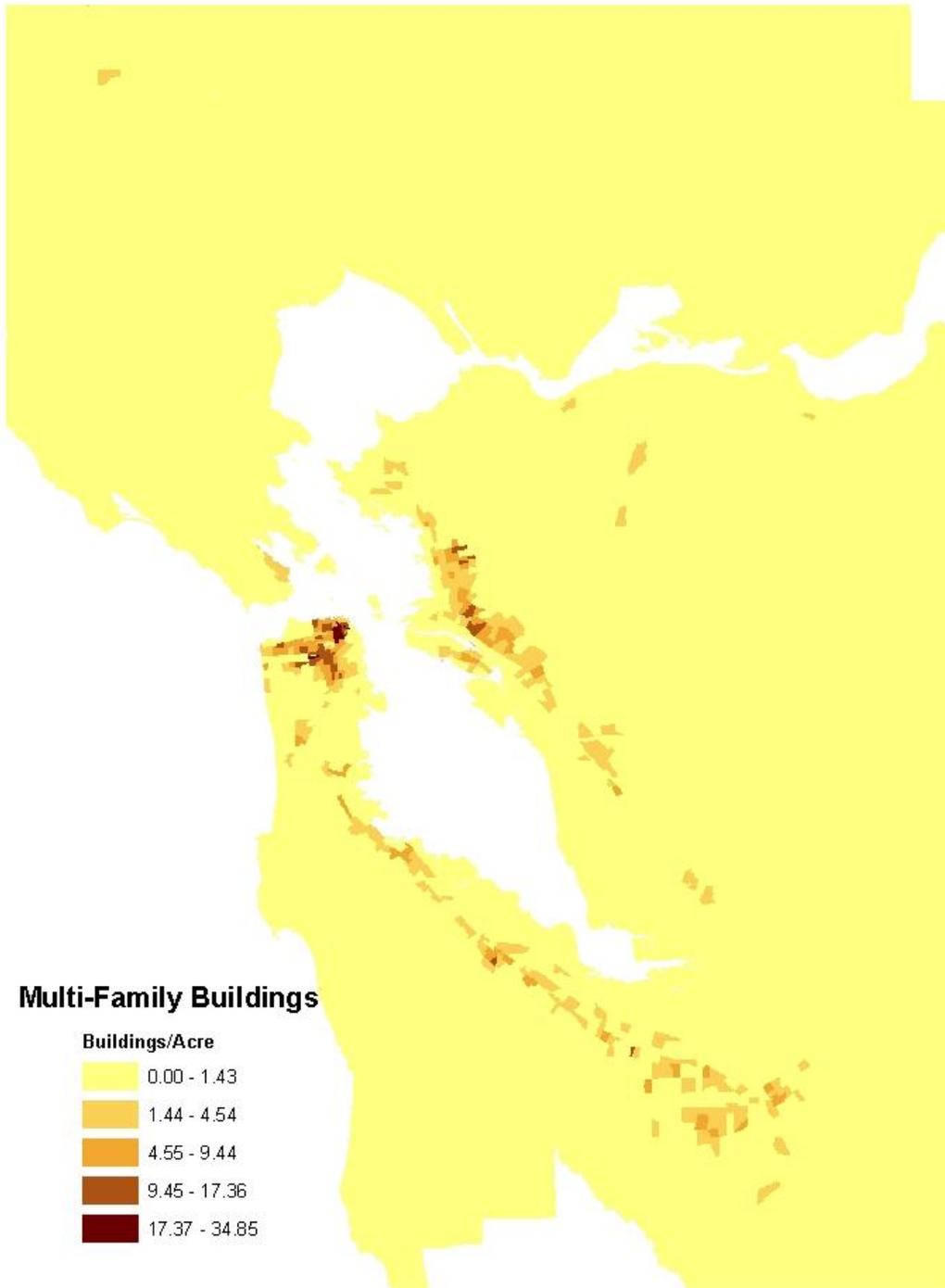
# CoStar Industrial Buildings



# CoStar Land Buildings



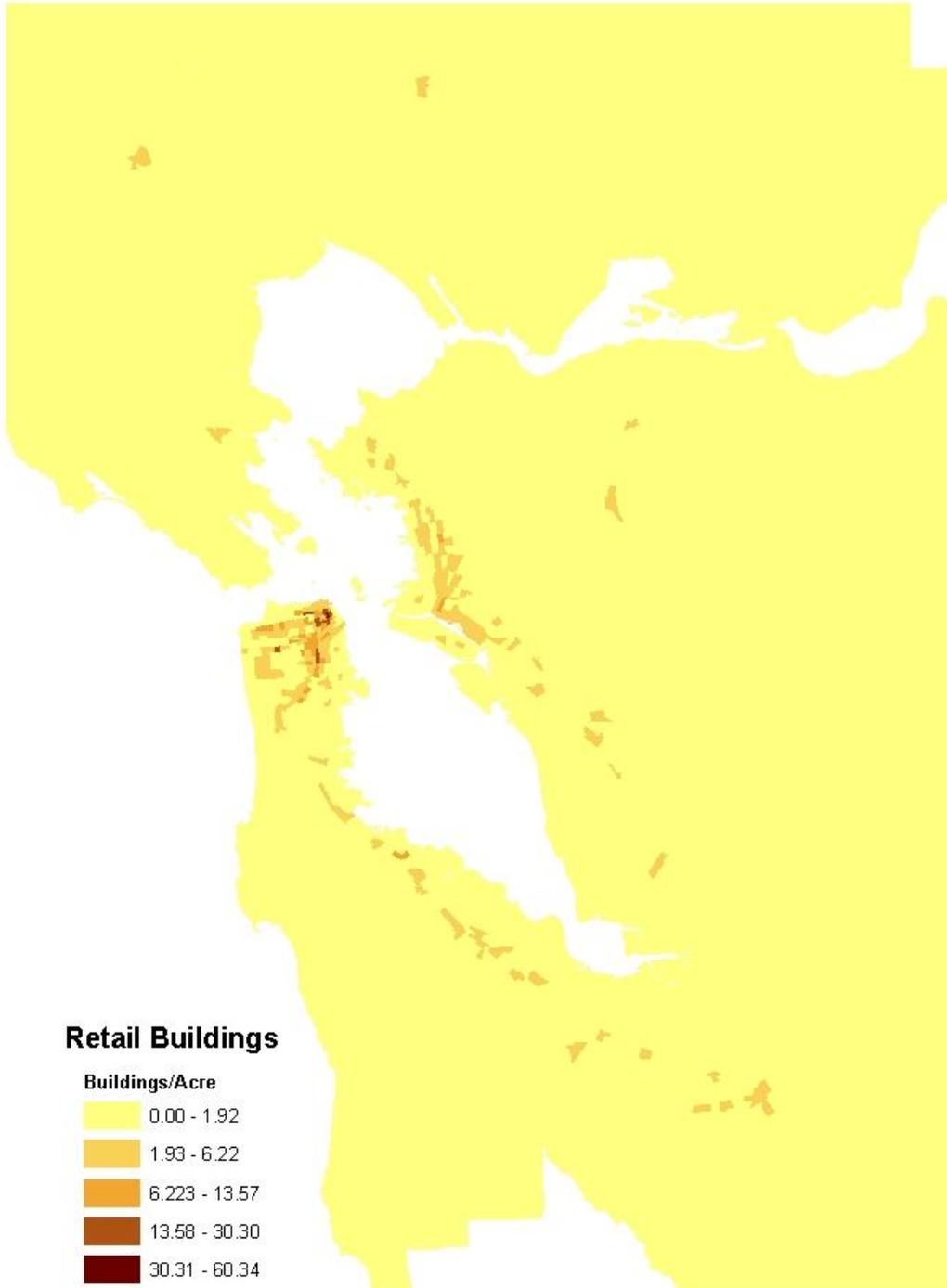
## CoStar Multi-Family Buildings



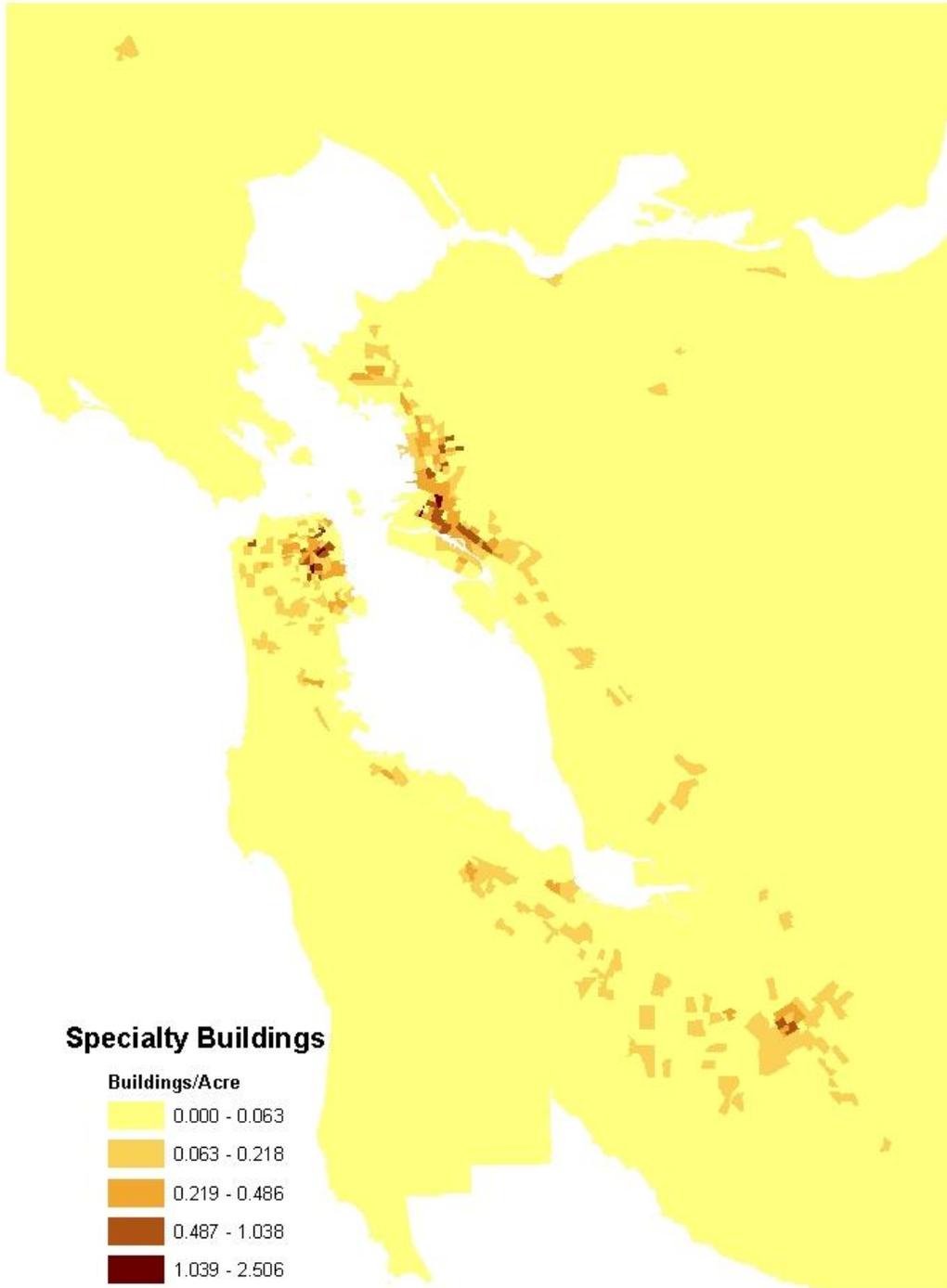
# CoStar Office Buildings



# CoStar Retail Buildings



# CoStar Specialty Buildings



## CoStar Sports and Entertainment Buildings



## Parcel Data Summaries

The final series of summaries is related to the parcel data set. A table named “nets\_all\_sj\_puid” was created to store the entire distinct parcel IDs data set that matched with Nets establishments in all years. The query used to create this table is below.

```
CREATE TABLE nets_all_sj_puid
SELECT sj_puid FROM nets_geocode_spatialjoin_1989
WHERE sj_puid > 0
UNION DISTINCT
SELECT sj_puid FROM nets_geocode_spatialjoin_1990
WHERE sj_puid > 0
.....
SELECT sj_puid FROM nets_geocode_spatialjoin_2009
WHERE sj_puid > 0
```

The distinct parcel IDs in the CoStar table were then compared with the distinct parcel IDs in the Nets table to determine the parcel IDs that matched in both the Nets and CoStar data sets. The table below shows that about 75% of the parcels with CoStar data also have Nets data. The following queries were used to generate the summary table below:

```
SELECT count(*) FROM parcels_ba

select count(distinct sj_puid) from costar_geocode_spatialjoin;

SELECT count(distinct c.sj_puid) FROM costar_geocode_spatialjoin as c WHERE c.sj_puid

SELECT count(distinct n.sj_puid) FROM nets_all_sj_puid as n WHERE nets_all_sj_puid.sj_puid

SELECT count(distinct c.sj_puid)
FROM costar_geocode_spatialjoin as c, nets_all_sj_puid
WHERE c.sj_puid=nets_all_sj_puid.sj_puid
```

**Table 44 – Parcel Match Rate**

Data Set	Records	Percent of Parcels
Parcels	2079235	100%
Parcels w/ Nets Records	436561	21%
Parcels w/ CoStar Records	97003	5%
Parcels w/ Nets and CoStar Records	73385	4%

## Conclusions

Overall the geocoding results look pretty good. The Nets data was matched successfully to either to the Parcel or TeleAtlas layer 85 percent of the time, whereas the CoStar data was matched 98 percent of the time. Initially the match rate was much better for TeleAtlas than for Parcels, but after some significant improvements to the Parcel layer, the Parcel match rate improved significantly. The Nets data is not quite as geographic accurate as the CoStar data, as evidenced by its lower match rate. The parcel match rates vary by county in both the Nets and CoStar data sets, which suggests the quality of the parcel data varies by county. All of the maps produced for both Nets and CoStar show reasonable results, especially the CoStar TAZ maps. Finally, the join of the two data sets to the parcel layer resulted in about 75 percent of the parcels with CoStar data also having Nets data. As with any geocoding effort, additional time could be spent investigating the unmatched records for systematic data discrepancies that could be fixed in order to improve the match rate. Based on the data cleaning that was done to date, it is recommended to continue reviewing the parcel layer since it is a key data set in this process and since it required significant revisions.

## Files

The following files are included with this memo:

- 1) R Data Processing Scripts
  - a. abag\_data\_processing.R – script for data processing /cleaning
  - b. createFinalCoStarFile.R – script to create final CoStar table
  - c. createNetsYearTables.R – script to create Nets yearly tables
  - d. ReadJoinWriteFiles.R – script to merge the CoStar Excel files into one table
- 2) Final Nets and Costar Data
  - a. nets\_year\_<year>.csv – Nets yearly tables with geocoding and spatial join results
  - b. costar.csv – CoStar table with geocoding and spatial join results
- 3) Geocoding Layers
  - a. Parcels.gdb – parcel geodatabase
  - b. tana\_v9\_ba\_g.gdb – TeleAtlas geodatabase
- 4) Address Locators
  - a. Composite.loc, Composite.loc.xml – composite locator
  - b. TANA.lox, TANA.loc, TANA.loc.xml – TeleAtlas locator
  - c. ParcelsWGS.lox, ParcelsWGS.loc, ParcelsWGS.loc.xml - Parcel locator