Maps, Messy Data, and Misleading Correlations

BioQUEST 2012 Summer Workshop

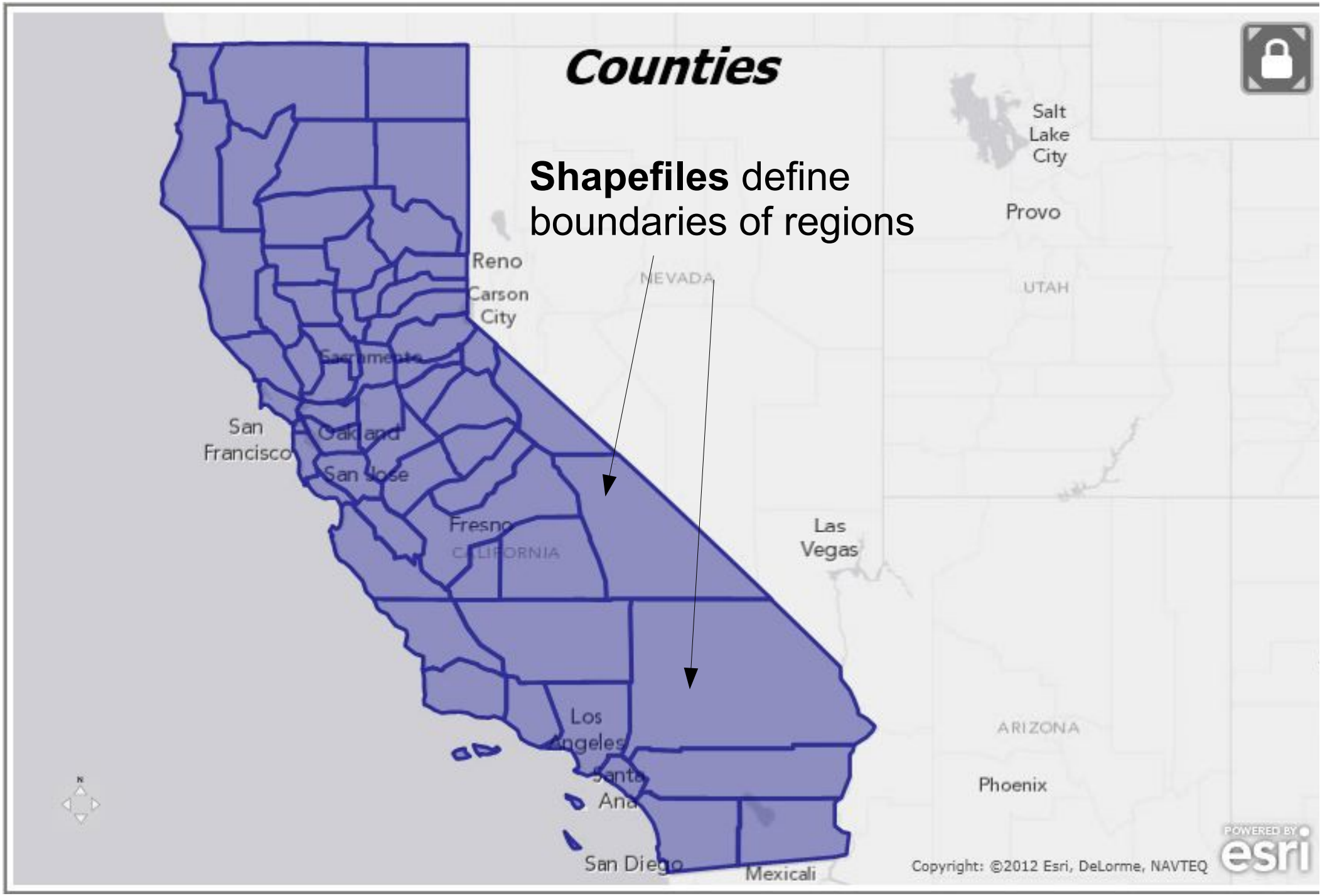
Dave Bourgaize
Jeff Lutgen

Whittier College

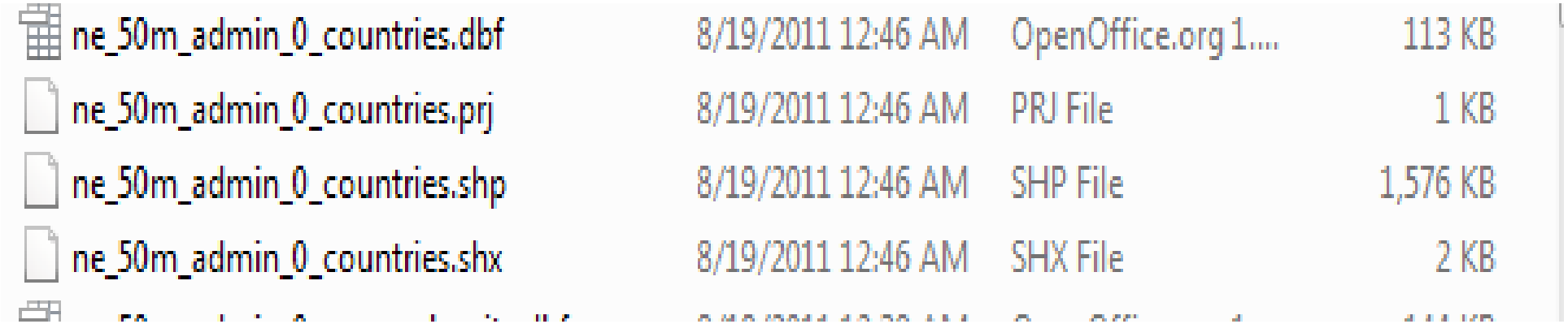
Purposes of the exercise:

1. Pose a georeferenced question that is (hopefully) interesting. We think we have an example of one that might appear to have a simple answer....
2. Find suitable data sets.
3. Manipulate data as necessary (database curation).
4. Create useful (i.e., that will help address the question) georeferenced visualizations of the data.
5. Propose hypotheses based on visual representations of data.
6. Examine and analyze data after forming hypotheses.
7. Pay attention to the reliability of data sets.


California
Salt Lake City
Provo
Reno
NEVADA
UTAH
Carson City
Sacramento
San Francisco
Oakland
San Jose
Fresno
CALIFORNIA
Las Vegas
Los Angeles
ARIZONA
Santa Ana
Phoenix
San Diego
Mexicali
POWERED BY esri


Counties
Shapefiles define boundaries of regions
Salt Lake City
Provo
Reno
Carson City
NEVADA
UTAH
San Francisco
Oakland
San Jose
Fresno
CALIFORNIA
Las Vegas
Los Angeles
Santa Ana
ARIZONA
Phoenix
San Diego
Mexicali
Copyright: ©2012 Esri, DeLorme, NAVTEQ
POWERED BY esri

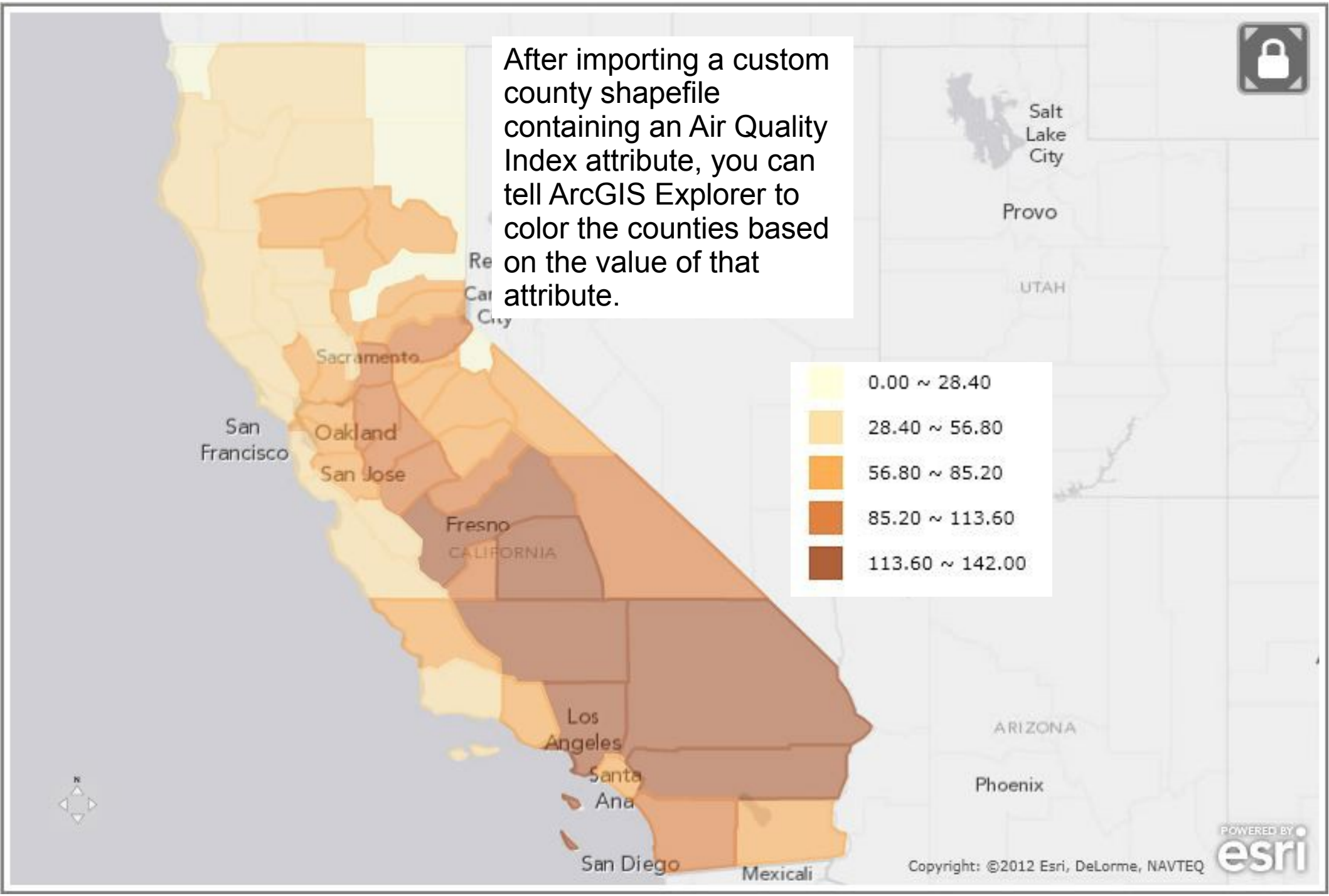
ArcGIS Explorer Online expects a shapefile to be a ZIP archive containing several files:



The .dbf file is a database file in dBASE format. It contains records of *attributes* for each shape. A typical shapefile (readily available on the internet) for U.S. counties might have a dbf file containing the population and area of each county.

That's nice, but we want to add our own custom attributes (an Air Quality Index value for each county, perhaps).

You can use OpenOffice to open dbf files and add attributes to them (but read the Wikipedia article on the dbf file format first!).


After importing a custom county shapefile containing an Air Quality Index attribute, you can tell ArcGIS Explorer to color the counties based on the value of that attribute.
0.00 ~ 28.40
28.40 ~ 56.80
56.80 ~ 85.20
85.20 ~ 113.60
113.60 ~ 142.00
Salt Lake City
Provo
UTAH
Sacramento
San Francisco
Oakland
San Jose
Fresno
CALIFORNIA
Los Angeles
Santa Ana
San Diego
Mexicali
ARIZONA
Phoenix
Copyright: ©2012 Esri, DeLorme, NAVTEQ
POWERED BY esri

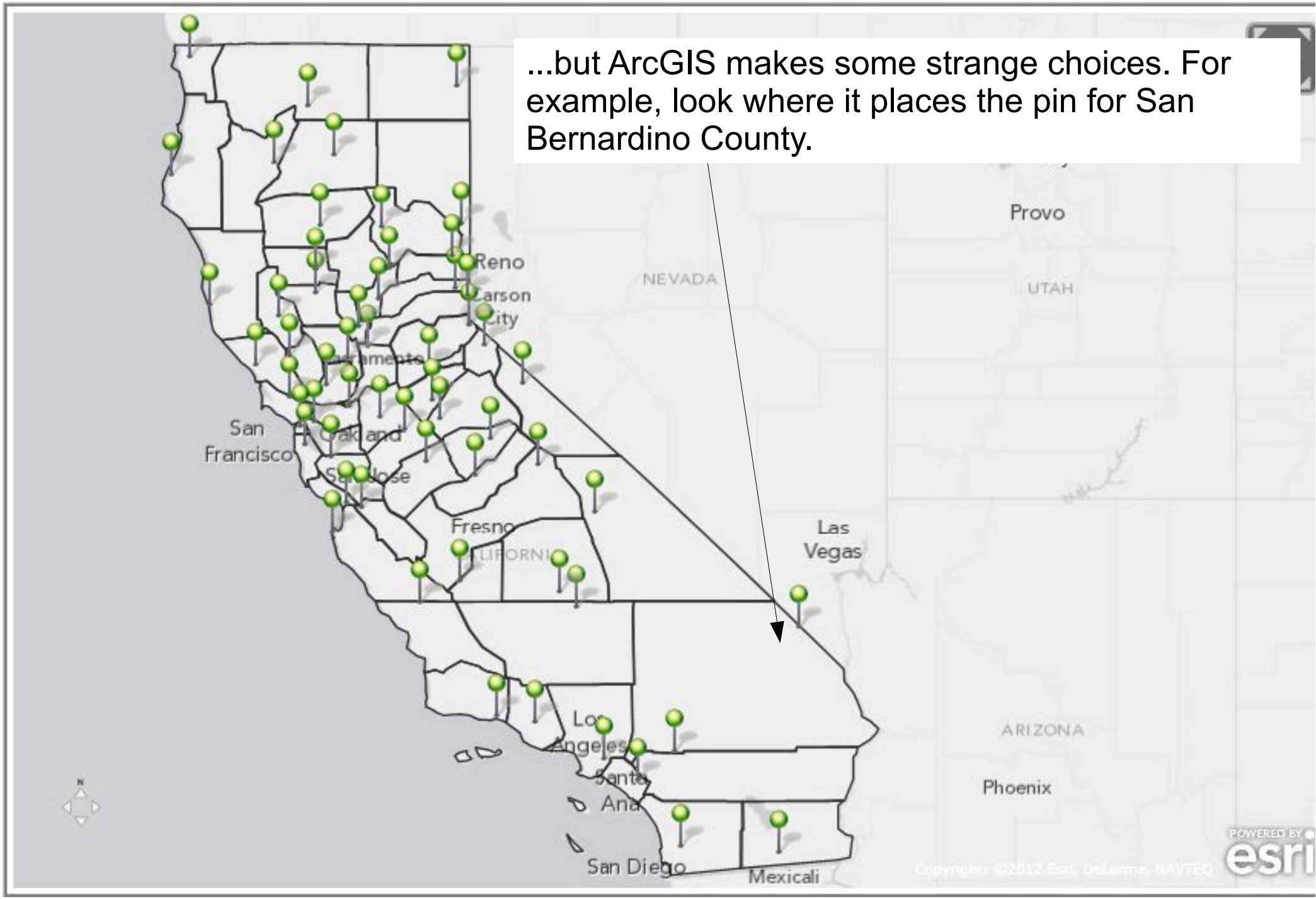
Association/State/County	Total Population[2]	Pediatric Asthma[3]	Adult Asthma[4]	Chronic Bronchitis[5]	Emphysema[6]	Lung Cancer[7,8]
California continued						
Glenn	28,122	522	1,566	883	399	14
Humboldt	134,623	1,795	8,305	4,634	2,011	68
Imperial	174,528	3,390	9,415	5,139	2,104	89
Inyo	18,546	259	1,155	683	348	9
Kern	839,631	16,855	44,572	24,047	9,398	426
Kings	152,982	2,822	8,348	4,423	1,619	79
Lake	64,665	907	4,023	2,359	1,171	33
Lassen	34,895	417	2,185	1,179	456	18
Los Angeles	9,818,605	159,352	566,147	308,756	125,611	4,947
Madera	150,865	2,842	8,270	4,562	1,933	76
Marin	252,409	3,464	15,773	9,179	4,449	127

American Lung Association data is available only in PDF reports, not as plain CSV text. Grrrrrrr. Can cut and paste into spreadsheet or text document, but some tedious manual reformatting is unavoidable.

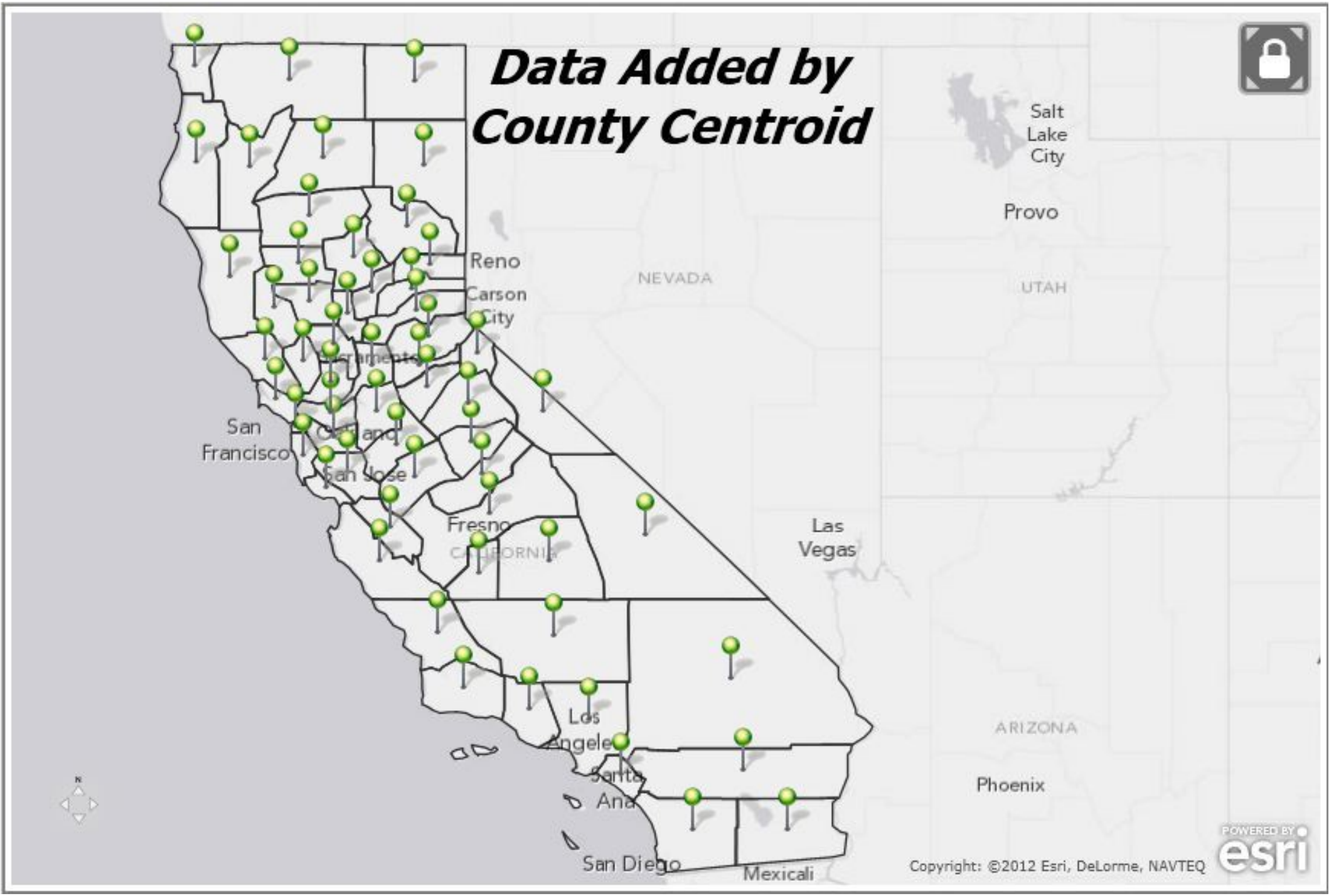
Notice that the disease incidence data seems to be expressed as raw counts, but the population of each county is also given, so it's easy enough to compute incidence rates per 100,000.

ArcGIS Explorer knows about county names, so to map the county asthma rates, you can import a CSV file like this one:

	A	H	I
1	County	pediatric asth	adult asthma
2	Alameda	1496.089112	5939.662484
3	Alpine	1446.808511	6297.87234
4	Amador	1113.123835	6594.733664
5	Butte	1392.272727	6086.818182
6	Calaveras	1301.066304	6415.375839
7	Colusa	1984.219618	5392.408609
8	Contra Costa	1647.339196	5830.652272
9	Del Norte	1422.579518	6085.284865
10	El Dorado	1508.356438	6106.882877
11	Fresno	1978.505024	5343.758396
12	Glenn	1856.197994	5568.593983

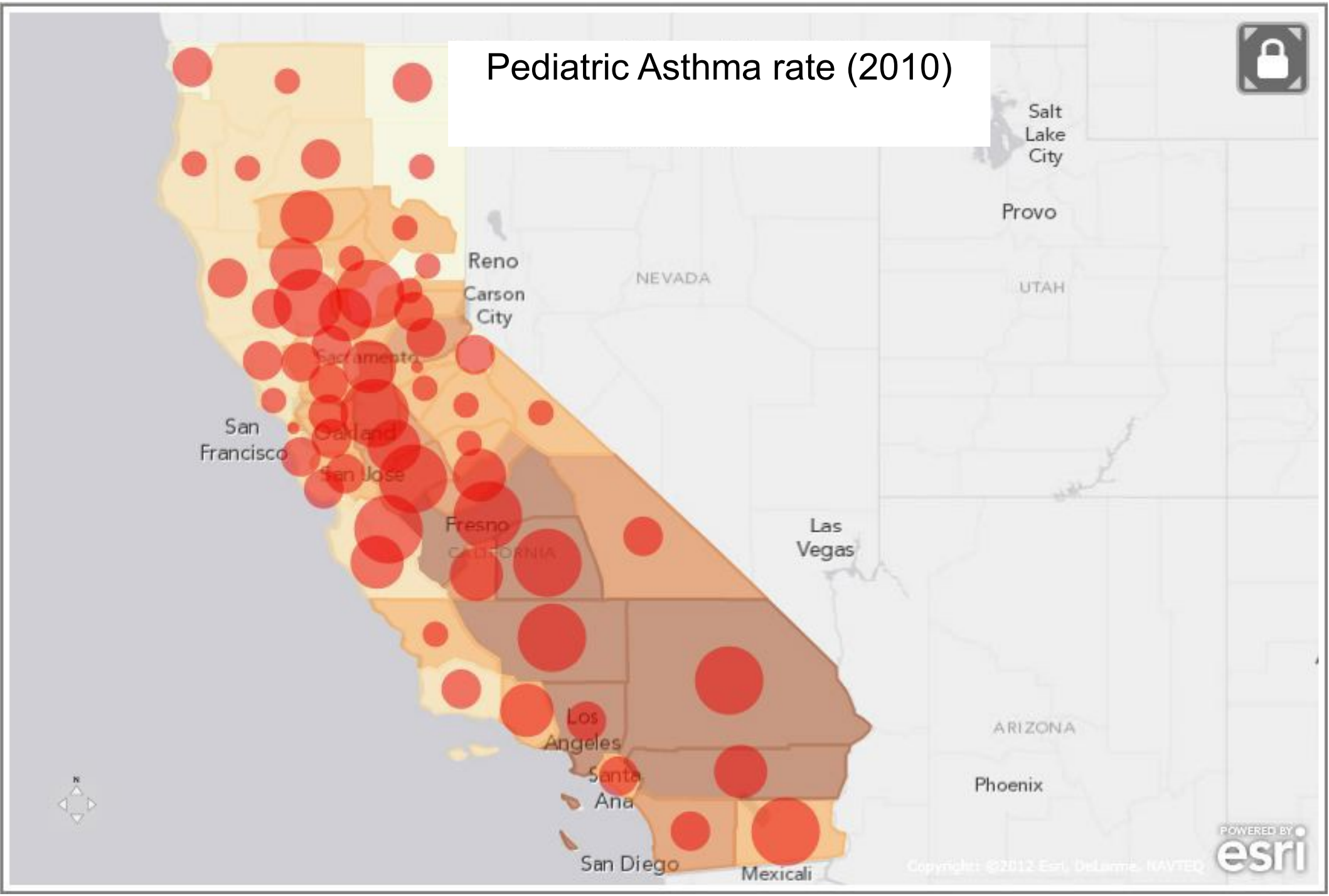

...but ArcGIS makes some strange choices. For example, look where it places the pin for San Bernardino County.
Provo
NEVADA
UTAH
Reno
Carson City
San Francisco
Oakland
San Jose
Fresno
Las Vegas
Los Angeles
Santa Ana
San Diego
Mexicali
ARIZONA
Phoenix
POWERED BY esri

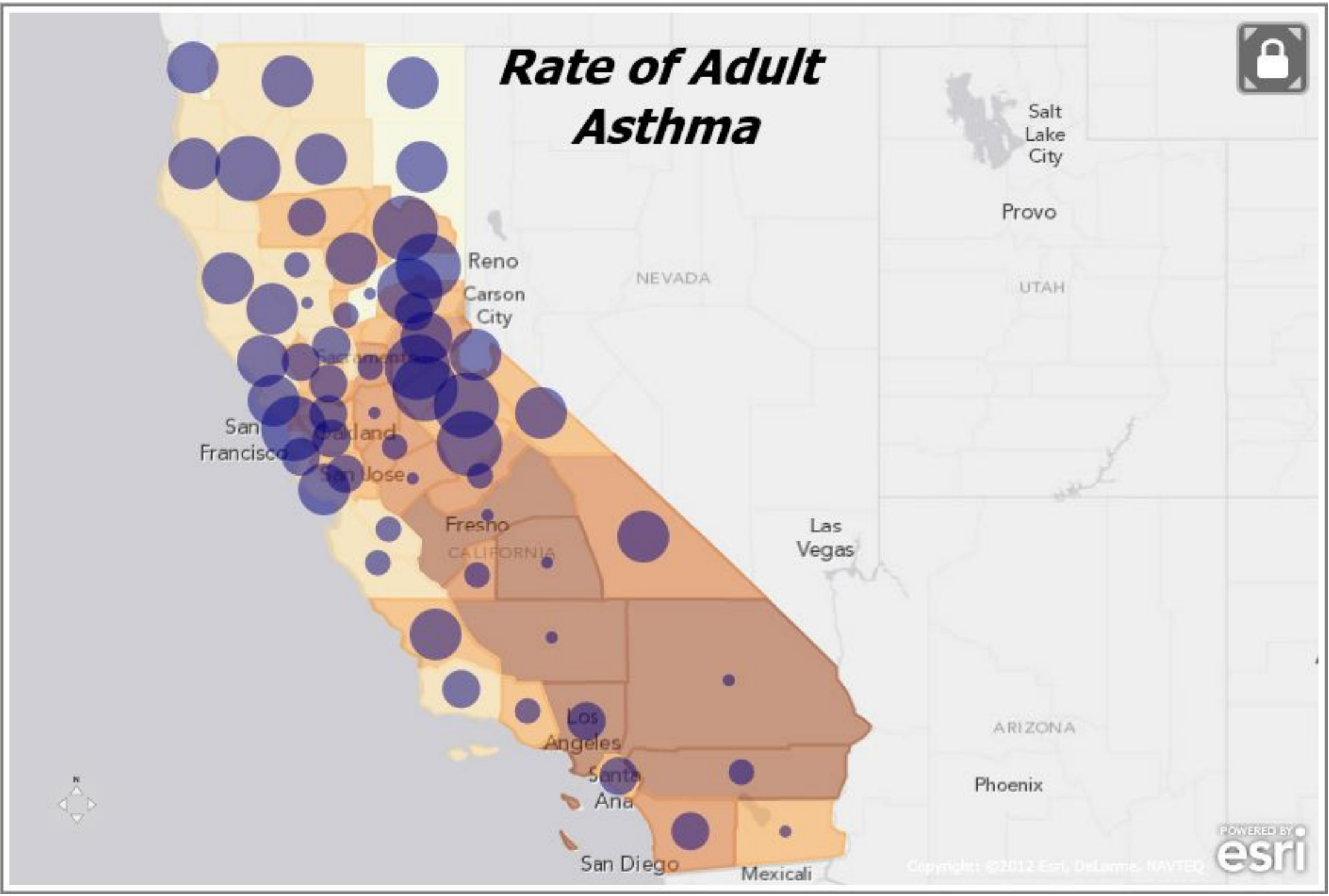
Apparently we must help ArcGIS by telling it the longitude and latitude of the center of each county. Luckily, data on U.S. county centroids is readily available (from census.gov, for example). After adding columns for latitude and longitude to our CSV file and reimporting, we get a much more pleasing map:

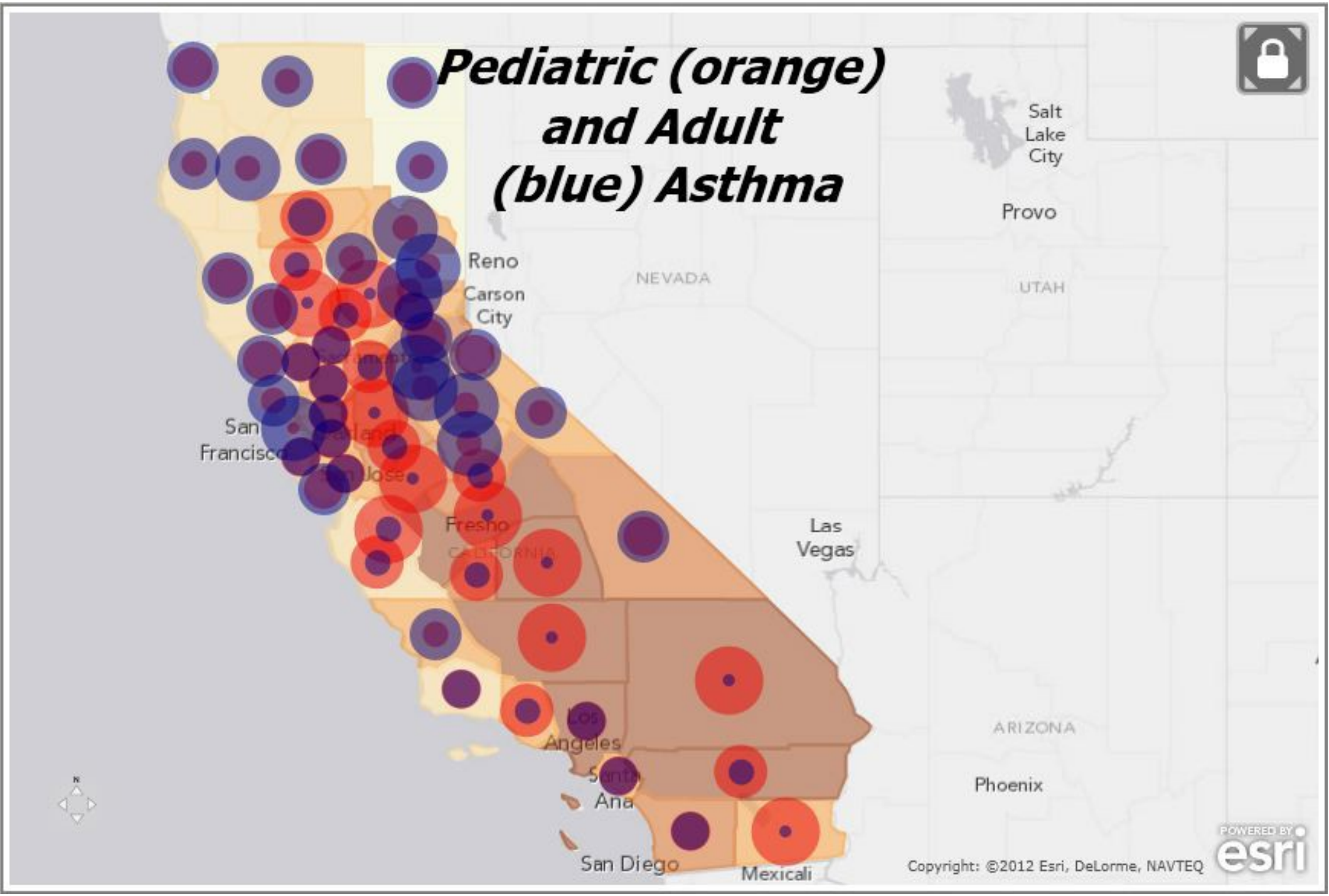

Data Added by
County Centroid
Salt
Lake
City
Provo
Reno
NEVADA
UTAH
Carson
City
San
Francisco
Fresno
Las
Vegas
Los
Angele
Santa
Ana
ARIZONA
Phoenix
San Diego
Mexicali
Copyright: ©2012 Esri, DeLorme, NAVTEQ
POWERED BY
esri

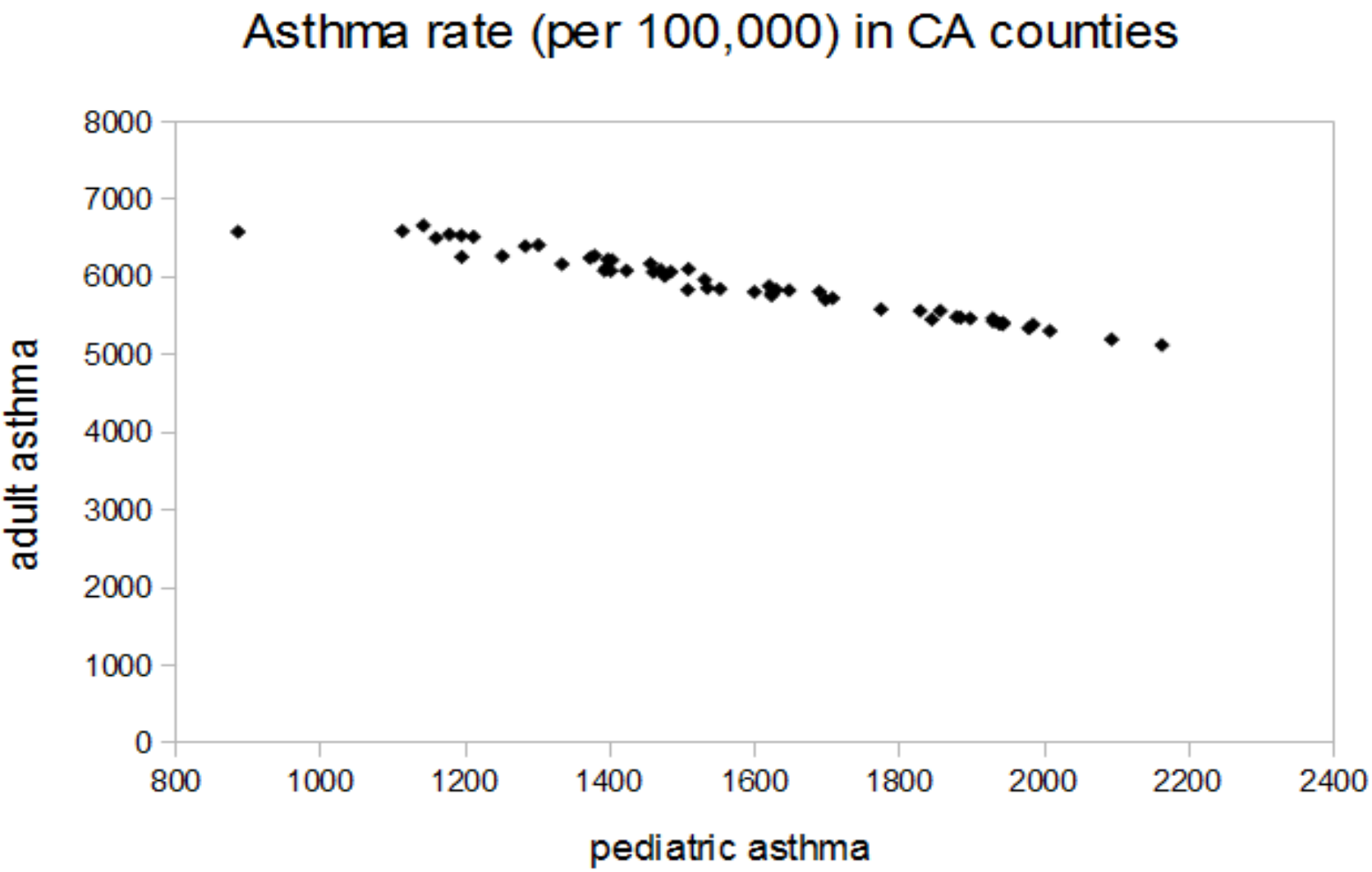
With our custom shapefile and centroid files in place, it is straightforward to add map layers for any data set for California counties by adding columns to the centroid CSV file.

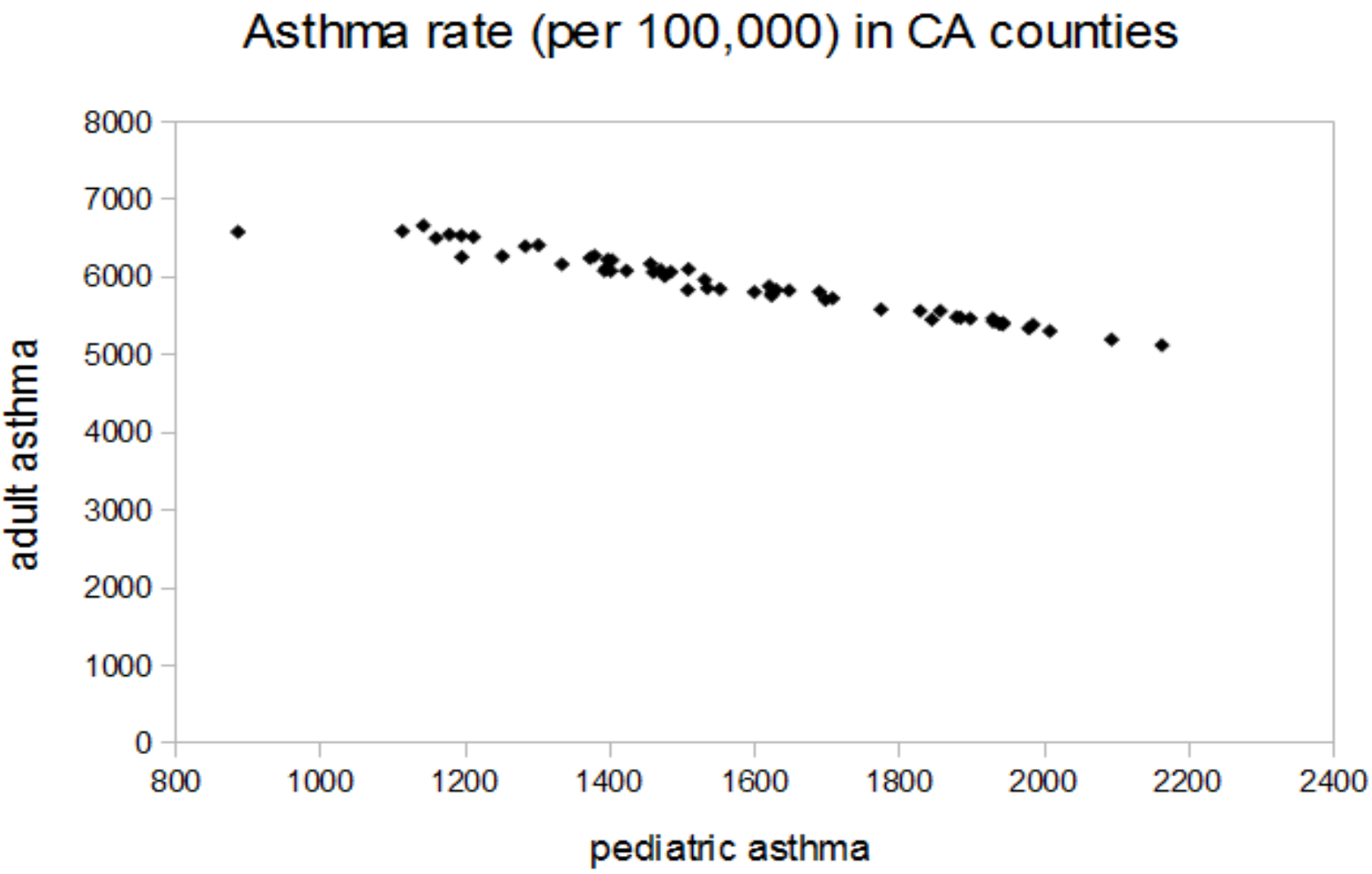
Let's map our 2010 pediatric and adult asthma rate data from the American Lung Association on top of the AQI data, first separately, then together:


Pediatric Asthma rate (2010)
Salt Lake City
Provo
Reno
NEVADA
Carson City
UTAH
San Francisco
Oakland
San Jose
Fresno
Las Vegas
Los Angeles
Santa Ana
ARIZONA
Phoenix
San Diego
Mexicali
POWERED BY esri


Rate of Adult Asthma
Salt Lake City
Provo
Reno
NEVADA
UTAH
Carson City
San Francisco
Oakland
San Jose
Fresno
CALIFORNIA
Las Vegas
Los Angeles
Santa Ana
ARIZONA
Phoenix
San Diego
Mexicali
POWERED BY esri


Pediatric (orange) and Adult (blue) Asthma
Salt Lake City
Provo
Reno
NEVADA
Carson City
UTAH
San Francisco
Fresno
Las Vegas
Los Angeles
ARIZONA
Santa Ana
Phoenix
San Diego
Mexicali
Copyright: ©2012 Esri, DeLorme, NAVTEQ
POWERED BY esri


Asthma rate (per 100,000) in CA counties
adult asthma
8000
7000
6000
5000
4000
3000
2000
1000
0
800
1000
1200
1400
1600
1800
2000
2200
2400
pediatric asthma


Asthma rate (per 100,000) in CA counties
adult asthma
8000
7000
6000
5000
4000
3000
2000
1000
0
800
1000
1200
1400
1600
1800
2000
2200
2400
pediatric asthma

$R^2 = 0.96$ (!)

Suspiciously high correlations?

	pediatric asth	adult asthma	chronic bron	emphysema	lung cancer r
pediatric asth	1				
adult asthma	-0.98	1			
chronic bron	-0.90	0.97	1		
emphysema	-0.77	0.87	0.97	1	
lung cancer r	-0.21	0.24	0.25	0.26	1

Hmmmm. Back to the data source (American Lung Association) to read the fine print...

Association/State/County	Total Population[2]	Pediatric Asthma[3]	Adult Asthma[4]	Chronic Bronchitis[5]	Emphysema[6]	Lung Cancer[7,8]
California continued						
Glenn	28,122	522	1,566	883	399	14
Humboldt	134,623	1,795	8,305	4,634	2,011	68
Imperial	174,528	3,390	9,415	5,139	2,104	89
Inyo	18,546	259	1,155	683	348	9
Kern	839,631	16,855	44,572	24,047	9,398	426
Kings	152,982	2,822	8,348	4,423	1,619	79
Lake	64,665	907	4,023	2,359	1,171	33
Lassen	34,895	417	2,185	1,179	456	18
Los Angeles	9,818,605	159,352	566,147	308,756	125,611	4,947
Madera	150,865	2,842	8,270	4,562	1,933	76
Marin	252,409	3,464	15,773	9,179	4,449	127

"[County] prevalence of adult asthma is estimated **by applying age-specific state prevalence rates** from the 2010 BRFSS **to age-specific county-level resident populations** obtained from the U.S. Census Bureau web site."

Uh-oh.

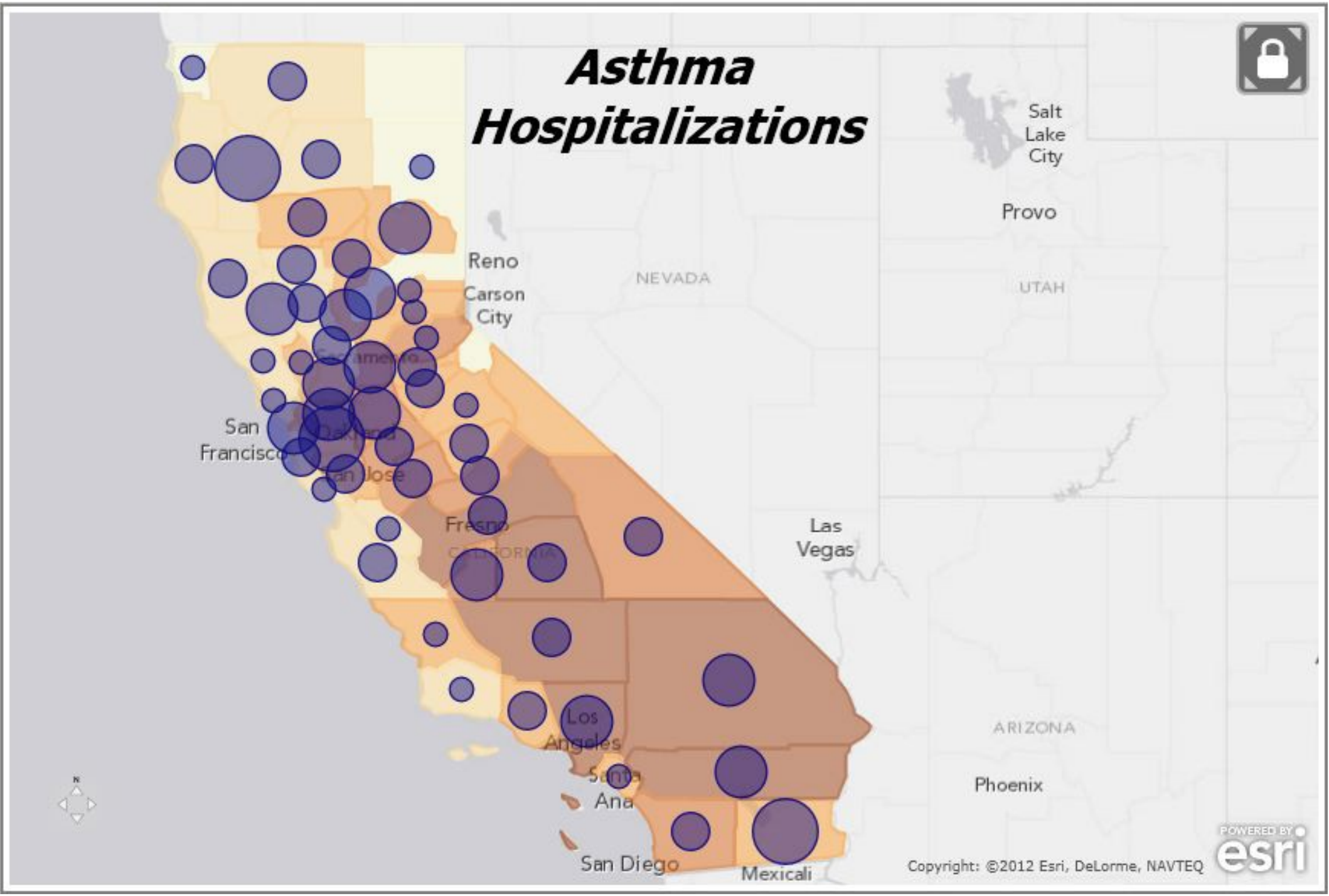
We need to get some real data. Eventually we find a report from the California Department of Public Health, Environmental Health Investigations Branch (ehib.org). Another PDF. Grrrrr.

Table 1: Age-Adjusted Asthma Hospitalization Rates* in California by County and Race/Ethnicity for All Ages, 1998-2000.

	Total		Non-Hispanic White		Black		Hispanic		Asian/ Pacific Islander	
County	Annual Rate	95% CI	Annual Rate	95% CI	Annual Rate	95% CI	Annual Rate	95% CI	Annual Rate	95% CI
CALIFORNIA	**11.11**	**(11.00 - 11.22)**	**9.48**	**(9.34 - 9.63)**	**33.01**	**(32.26 - 33.77)**	**10.25**	**(10.04 - 10.47)**	**7.82**	**(7.54 - 8.11)**
Alameda	17.92	(17.24 - 18.62)	11.06	(10.25 - 11.90)	43.07	(40.50 - 45.71)	11.66	(10.30 - 13.11)	10.97	(9.77 - 12.25)
Alpine	**	** **	**	** **	**	** **	**	** **	**	** **
Amador	8.79	(5.17 - 13.36)	9.16	(5.05 - 14.48)	**	** **	**	** **	**	** **
Butte	10.50	(9.13 - 11.97)	10.61	(9.09 - 12.23)	**	** **	9.26	(3.96 - 16.79)	**	** **
Calaveras	9.78	(6.71 - 13.44)	10.36	(7.01 - 14.35)	**	** **	**	** **	**	** **
Colusa	11.07	(6.14 - 17.45)	16.05	(8.23 - 26.45)	**	** **	9.90	(3.21 - 20.27)	**	** **
Contra Costa	12.85	(12.13 - 13.60)	9.47	(8.70 - 10.28)	37.88	(33.78 - 42.22)	10.94	(9.08 - 12.98)	9.79	(7.84 - 11.96)
Del Norte	6.94	(3.83 - 10.97)	8.07	(4.25 - 13.12)	**	** **	**	** **	**	** **
El Dorado	6.43	(5.23 - 7.76)	6.45	(5.16 - 7.87)	**	** **	**	** **	**	** **
Fresno	11.19	(10.45 - 11.94)	9.37	(8.39 - 10.40)	28.69	(22.46 - 45.45)	11.12	(9.86 - 12.45)	4.66	(3.16 - 6.45)

The data in the California EHBI report appear to be more realistic:

"Hospitalization data … was obtained from the California Office of Statewide Health Planning and Development. These computerized records included all hospital discharges in California, except from federal facilities. This database contains demographic information on each patient discharge, including age, sex, race, and zip code of residence. All discharges with asthma as the primary diagnosis were selected, based on the ninth revision of the International Classification of Diseases (ICD-9), code 493."


Asthma Hospitalizations
Salt Lake City
Provo
Reno
NEVADA
Carson City
UTAH
San Francisco
Fresno
Las Vegas
Los Angeles
ARIZONA
Santa Ana
Phoenix
San Diego
Mexicali
Copyright: ©2012 Esri, DeLorme, NAVTEQ
POWERED BY esri

Asthma Rate vs AQI for CA counties