Resources and Tools | Diversity Indices

Jerry Platt, School of Business, University of Redlands, CA

Basic Concept

This *Inland Empire Business Atlas* is populated with a wealth of data, which can be analyzed using statistics and displayed using mapping technology. Regarding the statistics, each chapter essentially presents a large rectangular block of data, with Southern California locations placed in rows and various attributes of those locations (including spatial coordinates or references) placed in the columns. Usually, interest is focused on summarizing characteristics of an individual attribute, determining whether levels of the attribute vary systematically over space, finding meaningful relations among attributes, and displaying summary results on maps.

To summarize the characteristics of a single numeric attribute, it is common to measure the *center* of the data, the *spread* of measures around that center, and the overall *shape* of the distribution of numeric values. In an introductory statistics course, it is common to represent the center and shape by the mean and standard deviation, respectively, and to simply assume that the shape resembles a "bell-shape curve". However, real-life data often has a shape that resembles some other, less symmetric, shape, and the standard deviation can be a misleading indicator of spread. The problem is compounded when the attribute of interest is expressed as multiple categories, the counts in which sum to the sample total for each location. This tutorial examines a measure that addresses the latter problem, providing an intuitive measure of dispersion that has a ready interpretation. It is known as a "diversity index", and (in at least one of its formulations) can be interpreted as the probability that any two entities selected at random are members of the same group. If all entities belong to a single group, then the probability should (and will) equal 100%; if they are distributed perfectly evenly among g groups, then the probability should approach 100%/g (and will as the sample size increases).

Examples, formulae and applications are provided. It is best to start with the fictional and highly-stylized **Toy Problem**, then proceed to the real but very **Simple Example**, and finally to read the **Applications** document. The **Formulae** are provided as a reference. Every chapter of the *Atlas* provides opportunities to apply and benefit from Diversity Indices, and there is a high likelihood they could prove valuable in your own work.

Source: Inland Empire Business Atlas, ©University of Redlands, 2009. Funded in part through a cooperative agreement with the U.S. Small Business Administration. All opinions, conclusions, or recommendations expressed are those of the author(s) and do not necessarily reflect the views of the SBA or the University of Redlands.

Toy Problem

Consider a business that invests $15,000,000 ($15m) among three sectors of the economy: $1m in Sector A, $5m in Sector B, and its remaining $9m in Sector C. How "diversified" is it?

The mean tells us that the average investment in a sector is $5m [($1m + $5m +$9m) / 3]. The standard deviation equals $2.83m. If all $15m had been invested in a single sector, the mean among these three sectors would remain at $3m (now a dubious summary measure), but the standard deviation would increase to $6.12m. This increase seems appropriate, given the increased dispersion from the (now dubious) mean, but the standard deviation stays the same whether all the funds are invested in Sector A, or in Sector B, or in Sector C -- although the shape of the distribution clearly changes. Also, the quantity 46.12m is difficult to interpret.

In this toy problem, let's assume Auditor I selects a single $1 at random from the $15m invested in this company, and traces its path; independently, Auditor II selects another $1 at random and does the same. What is the probability that the two auditors selected $1 from the same sector?

There are many (slight) variations on the computation of a diversity index, but one simple approach goes like this:

Sector A	$ 1m	→	1 * (1 – 1) = 0	
Sector B	$ 5m	→	5 * (5 – 1) = 20	
Sector C	$ 9m	→	9 * (9 – 1) = 72	Sum of Sectors = 92
Total $	$15m	==>	15 * (15 – 1) = 210	***Ratio = 92 / 210 = 0.44***

There is a ***44%*** probability that the two auditors drew from *the same* sector. If all the funds had been in one sector, the probability would be 100%; if the funds were evenly distributed among the three sectors, with $5m in each, the probability would be 29% -- as the magnitude of dollars increases (or the "m" for millions is not ignored), that lower limit approaches 33.33%.

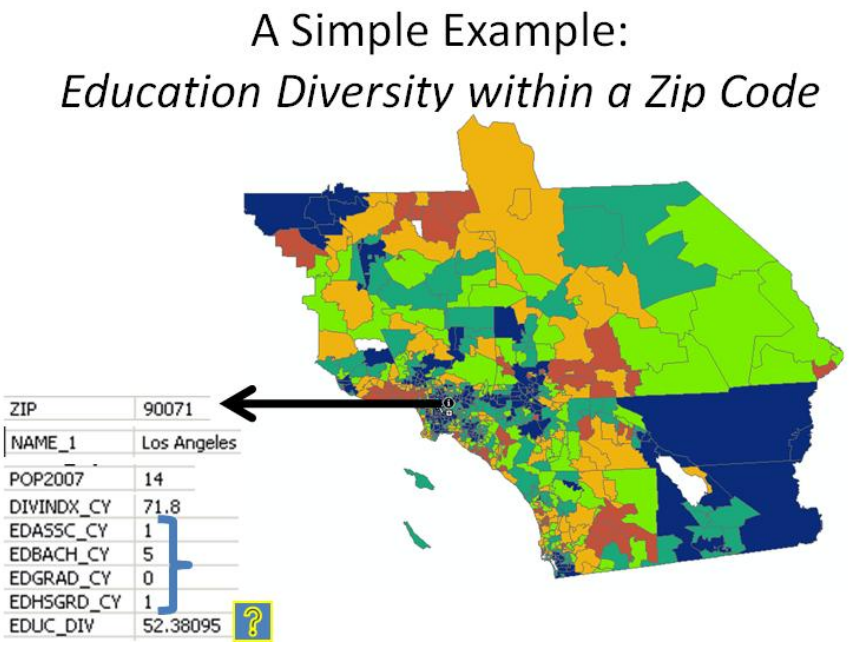
The "Diversity Index" often is computed as 1 – D, or 1 – 0.44 = 0.56, or ***56%,*** which can be interpreted as the probability that the two auditors drew from *different* sectors.

Source: Inland Empire Business Atlas, ©University of Redlands, 2009. Funded in part through a cooperative agreement with the U.S. Small Business Administration. All opinions, conclusions, or recommendations expressed are those of the author(s) and do not necessarily reflect the views of the SBA or the University of Redlands.

Simple Example

A Simple Example

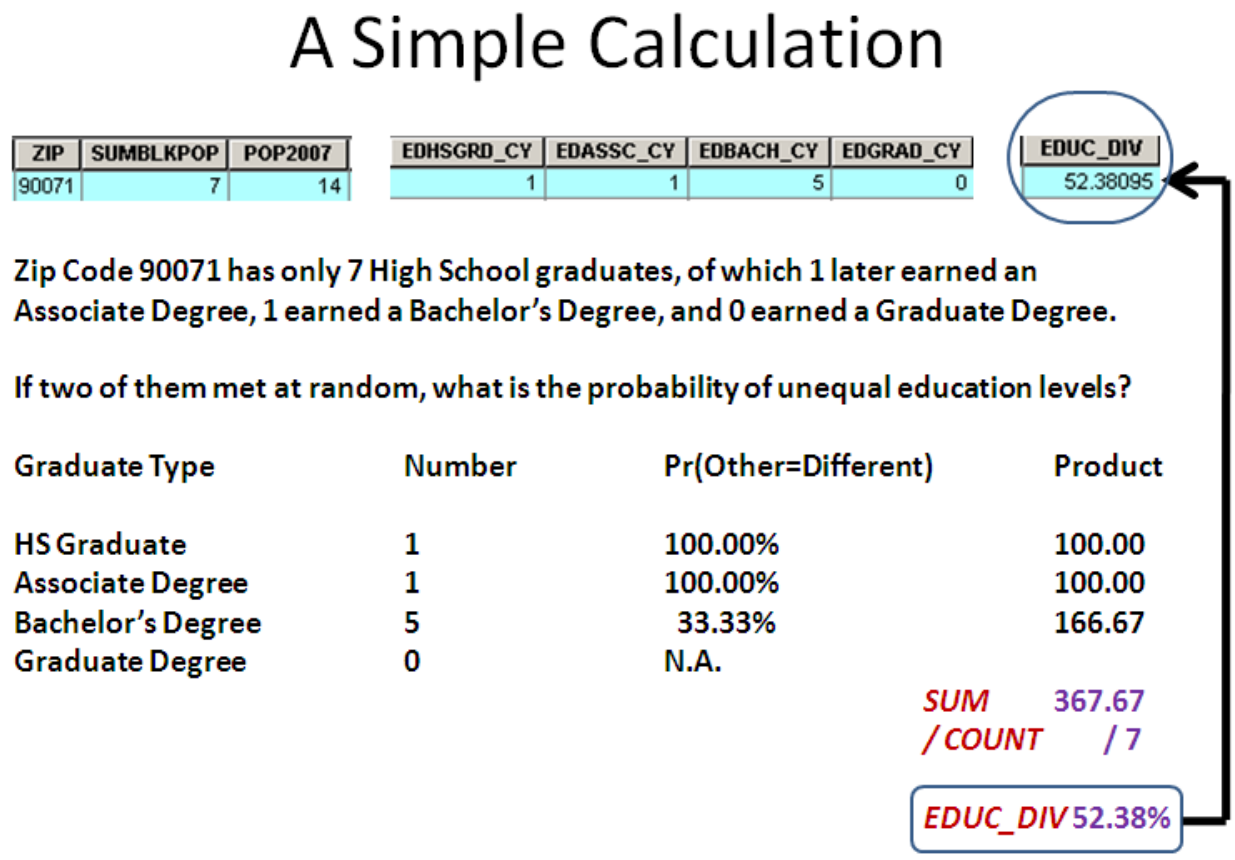
Zip Code 90071 in Los Angeles County is unusual – it has a total population of 14! Among the seven high school graduates in this county, 1 did not attend college, 1 received an Associate Degree, 5 received a Bachelor Degrees, and none received Graduate Degrees. What is the probability that, when one of the 7 meets another at random, they have the same education level? What is the probability they have different education levels? What is the "Diversity Index" for Zip Code 90071?



Using the same approach as before, there is a 47.62% they would have the same education level.

HS Grad	1	→	1 * (1 – 1) = 0	
Associate	1	→	1 * (1 – 1) = 0	
Bachelors	5	→	5 * (5 – 1) = 20	
Graduate	0	→	0 * (0 – (-1)) = 0	Sum of Sectors = 20
Total $	7	==>	7 * (7 – 1) = 42	***Ratio = 20 / 42 = 0.4762***

Therefore, there is a 52.38% probability that two HS graduates selected at random from Zip Code 90071 will have *different* levels of education. The "Diversity Index" usually is computed as 1 – D, or 1 – 0.4762 = 0.5238, or 52.38%. Following is another way to think of this calculation:



Either way it is derived, our estimate of the probability that two HS graduates selected at random from Zip Code 90071 will have *the same* levels of education = D = 47.62%.

Formulae

Alternative Measures

As indicated, there are several variations on measuring a Diversity Index. The approach used so far is known as the Simpson Index (from ecology). Primary competitors are the Shannon Index (from information theory), and the Greenberg Index (from linguistics). In the School Districts chapter of this *Atlas*, an "Ethnic Diversity Index" is reported for California school districts, based on seemingly more complicated calculations.

SHANNON $H = -\sum_{i=1}^{S} p_i \ln p_i$

SIMPSON $D = \frac{\Sigma\, n(n-1)}{N(N-1)}$

GREENBERG $DI = 1 - \Sigma(P_i)^2$

CALIFORNIA

Thus, for California schools, we have:

$$EDI(x_1,\ldots,x_7) = C_1 + C_2 \cdot d\left((x_1,x_2,\ldots,x_7),\left(\frac{1}{7},\frac{1}{7},\ldots,\frac{1}{7}\right)\right)$$

$$= C_1 + C_2 \cdot \sqrt{\left(x_1-\frac{1}{7}\right)^2 + \left(x_2-\frac{1}{7}\right)^2 + \ldots + \left(x_7-\frac{1}{7}\right)^2}$$

The good news is that choice of index tends to make relatively little difference in practice. The California school index lacks easy interpretation, but they all measure the same notion of diversity. Following is evidence drawn from the School District data, showing that (1) the California Index clearly is on a different scale than the others, (2) the congruence of the other three varies, particularly as entities become more concentrated in a single category, and (3) correlations confirm that they all measure the same basic concept.

SANTA ANA UNIFIED
IRVINE UNIFIED
NEWPORT-MESA UNI

Field	Value
SHANNONDI	0.18259
SIMPSONDI	0.14688
GREENBERGD	0.14687
CALIDI	8

REDLANDS UNIFIED

Field	Value
SHANNONDI	0.70834
SIMPSONDI	0.70855
GREENBERGD	0.70851
CALIDI	53

	SHANNON	SIMPSON	GREENBERG	CALIFORNIA
SHANNON	1.0000			
SIMPSON	0.9804	1.0000		
GREENBERG	0.9813	0.9999	1.0000	
CALIFORNIA	0.9575	0.9749	0.9763	1.0000

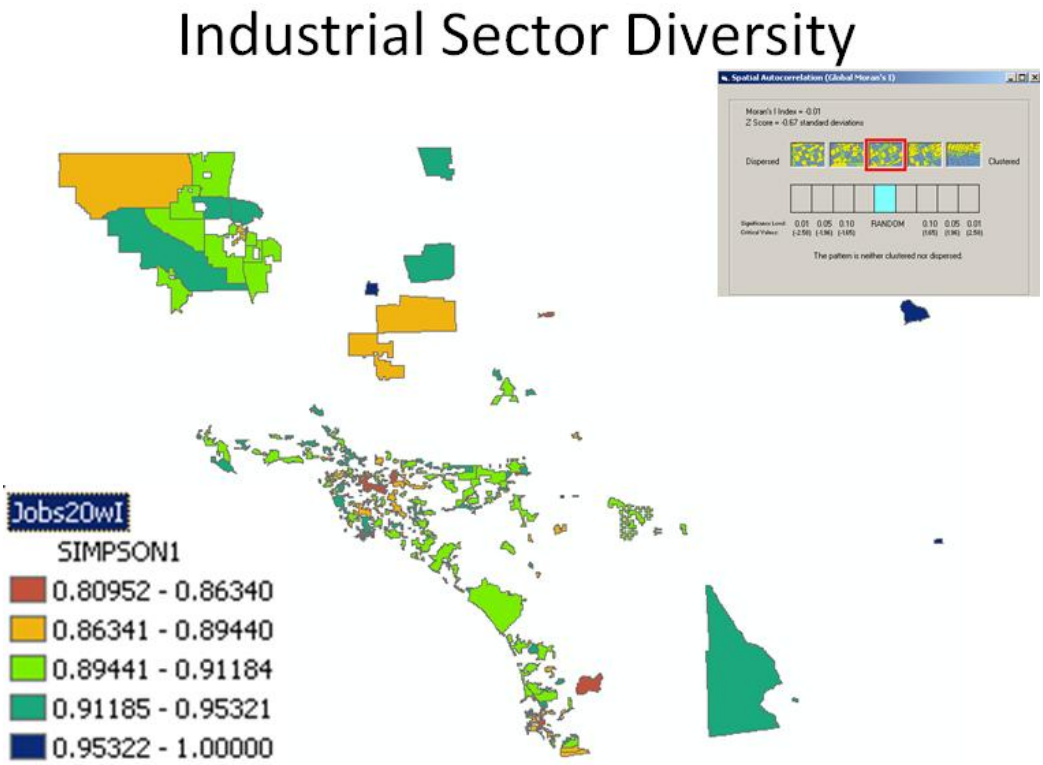
Source: Inland Empire Business Atlas, ©University of Redlands, 2009. Funded in part through a cooperative agreement with the U.S. Small Business Administration. All opinions, conclusions, or recommendations expressed are those of the author(s) and do not necessarily reflect the views of the SBA or the University of Redlands.

Applications

Every chapter of the *Atlas* provides opportunities to apply and benefit from Diversity Indices. In what follows three applications are briefly considered.

Industrial Sector Diversity

During the study year, there were 512,408 records of business establishments in the 8 counties that comprise Southern California. Each organization is associated with one of 20 2-digit NAICS industry classification codes. The distribution of jobs among these 20 sectors of the economy is of interest, and likely is different for any two geographic locations. A Location Quotient computes the relative intensity of employment in a given sector to the corresponding relative intensity in a base geography, such as among all employment in Southern California. While it does provide a basis for identifying the center (mode) among the computed Location Quotients, a Diversity Index is needed to address the relative concentration or dispersion of employment within a given job center, or to compare job dispersions among two or more job centers. The example below maps the Job Center Diversity Index among the 20 industry sectors for each of the 205 Job Centers, and the inserted chart indicates that the Diversity Index at a Job Center is not influenced by the level of Diversity Index at nearby Job Centers. Note that with 20 categories of jobs, the Diversity Index = 1.00 (or 100%) if all workers at the Job Center are employed in the same Industry Sector, and the lower limit of 0.20 (or 20%) is approached as employees are spread more evenly across all 20 Industry Sectors.

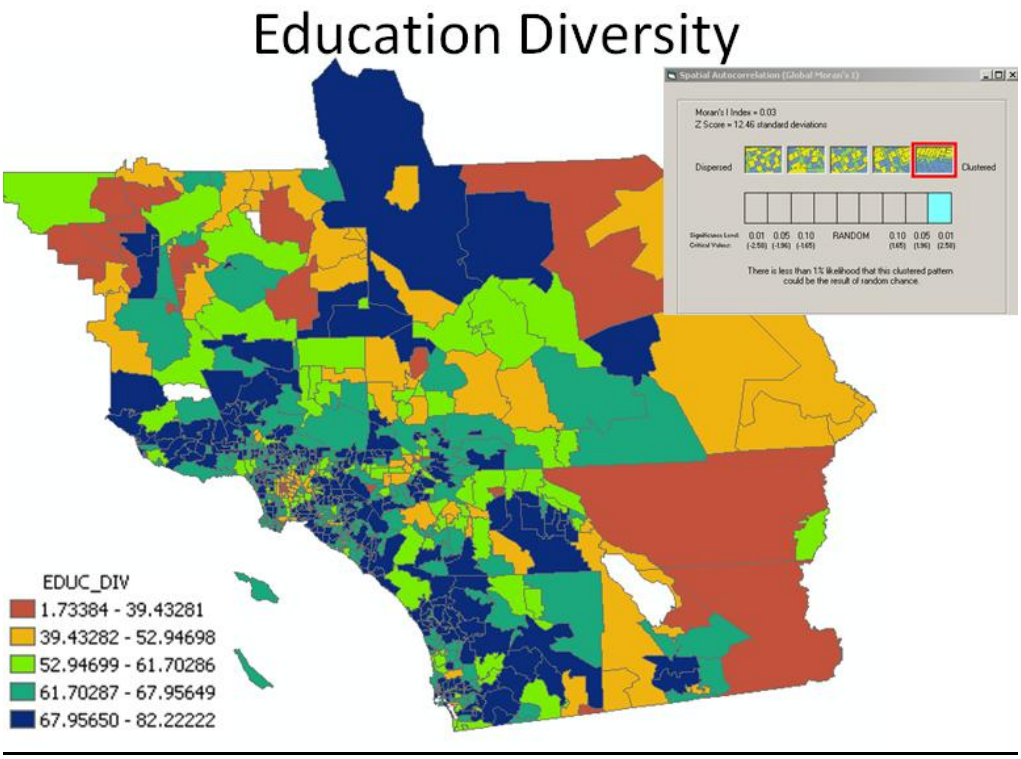


Demographic Diversity

This section applies the Diversity Index measure at the Zip Code level to classifications of education attainment, and then to reported racial and ethnic identification.

Education Diversity

As shown previously, the Census provides counts of population by level of education attainment. The map below displays that information at the Zip Code level. It appears that Zip Codes along the urban coastline have a relatively greater degree of Educational Diversity.

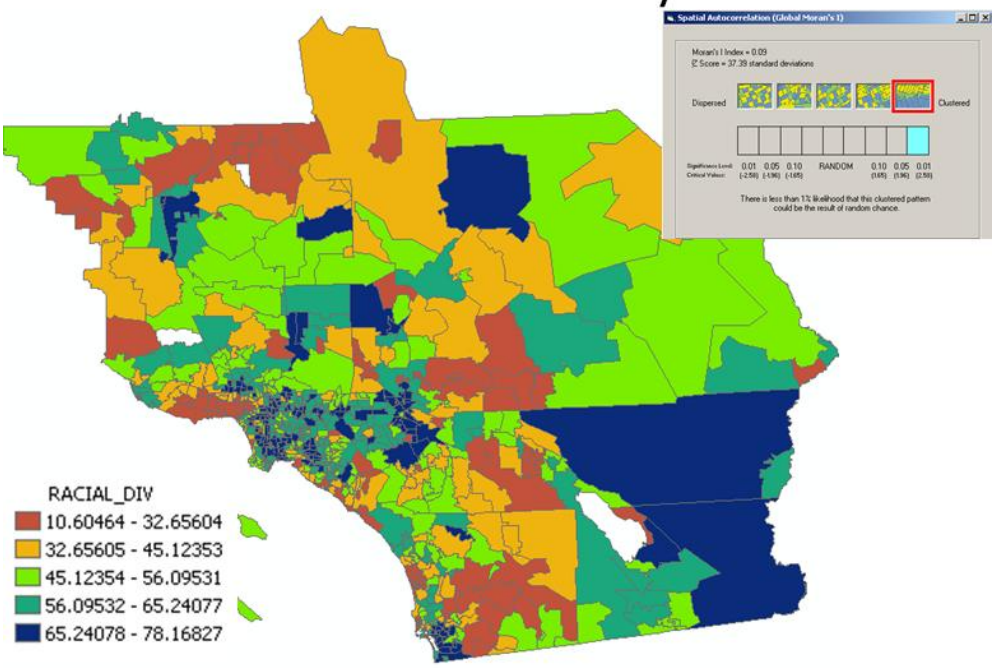


Racial and/or Ethnic Diversity

From Census Data, it is a non-trivial task to identify race or ethnicity. The survey asks respondents "which race or races, cultural groups, or national origins they identify with most closely". This is particularly problematic for identifying a "Hispanic" category, since it is a cultural choice. The Census reports 7 Race Categories: White, African-American, Native American, Asian, Pacific Islander, Other, and Two or More of Preceding. It separately reports a count of those who identify with "Hispanic Origin". The map below collapses Asian with Pacific Islander, and map a Racial Diversity Index for the resulting 6 racial classifications. As confirmed by the inset, there is a strong tendency for "birds of a feather to flock together".

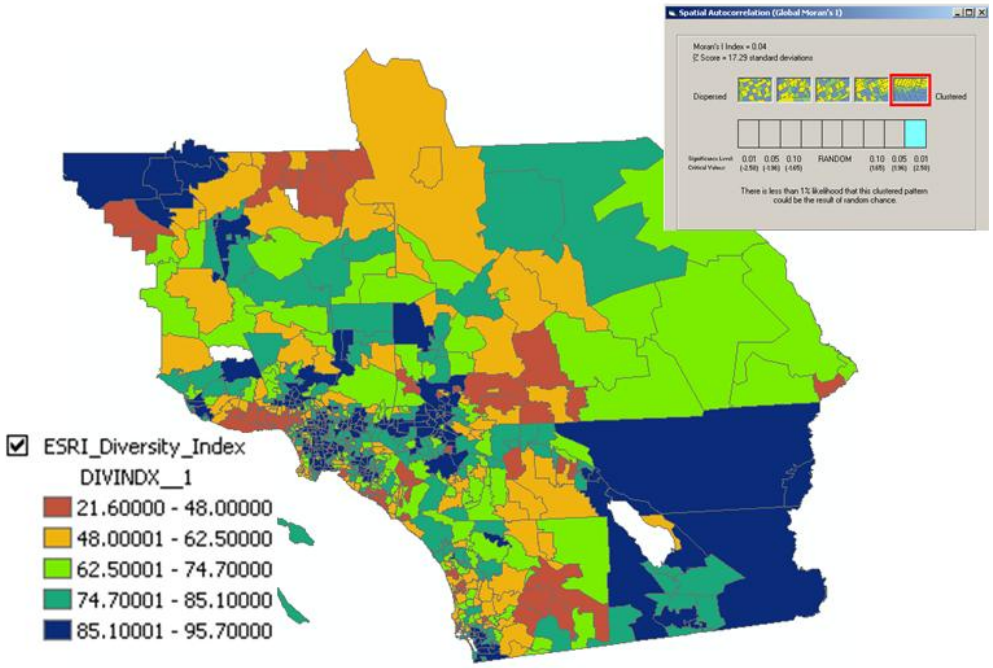
Source: Inland Empire Business Atlas, ©University of Redlands, 2009. Funded in part through a cooperative agreement with the U.S. Small Business Administration. All opinions, conclusions, or recommendations expressed are those of the author(s) and do not necessarily reflect the views of the SBA or the University of Redlands.

Racial Diversity



ESRI, Inc. reports a "Diversity Index" in its data that adjusts the information above to also reflect ethnicity. As shown below, it yields similar results.

ESRI Racial and Ethnic Diversity Index



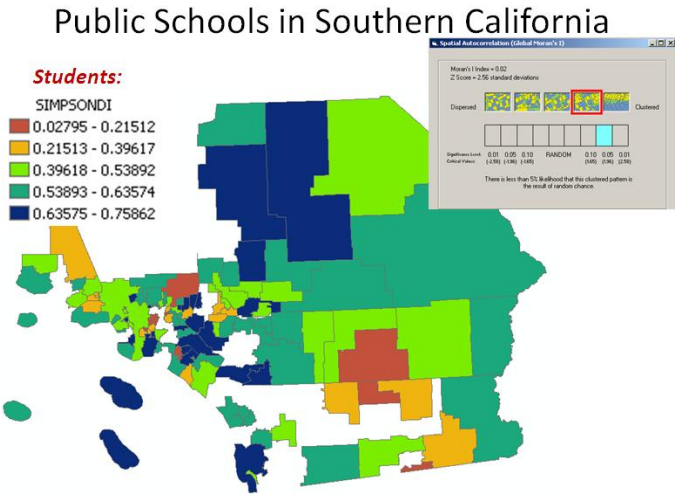
Source: Inland Empire Business Atlas, ©University of Redlands, 2009. Funded in part through a cooperative agreement with the U.S. Small Business Administration. All opinions, conclusions, or recommendations expressed are those of the author(s) and do not necessarily reflect the views of the SBA or the University of Redlands.

School Diversity in Southern California:

Finally, let's apply the Diversity Index concept to better understand patterns in public school districts in Southern California. This data is taken from http://www.ed-data.k12.ca.us/, which counts both *students* and *teachers* by reported ethnicity, using 8 categories: American Indian, Asian, Pacific Islander, Filipino, Hispanic, African American, White and "Multiple / No Response".

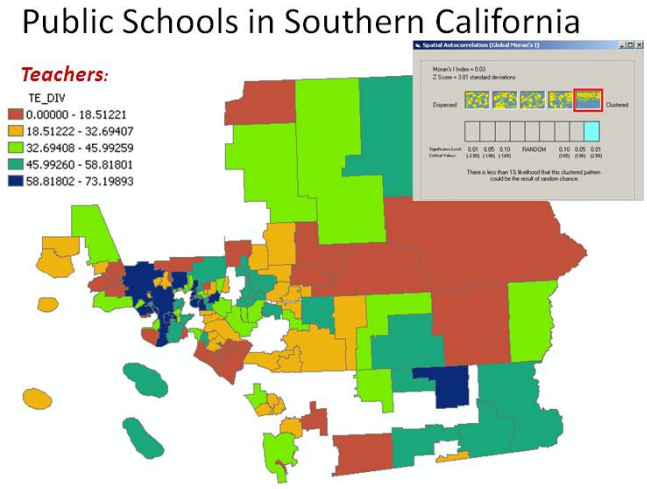
Student Diversity

Among student populations, there are many with high diversity, though less so for rural districts to the east, and but mild evidence of districts having similar levels of diversity to nearby districts.



Teacher Diversity

However, the teacher population map reveals a different story:

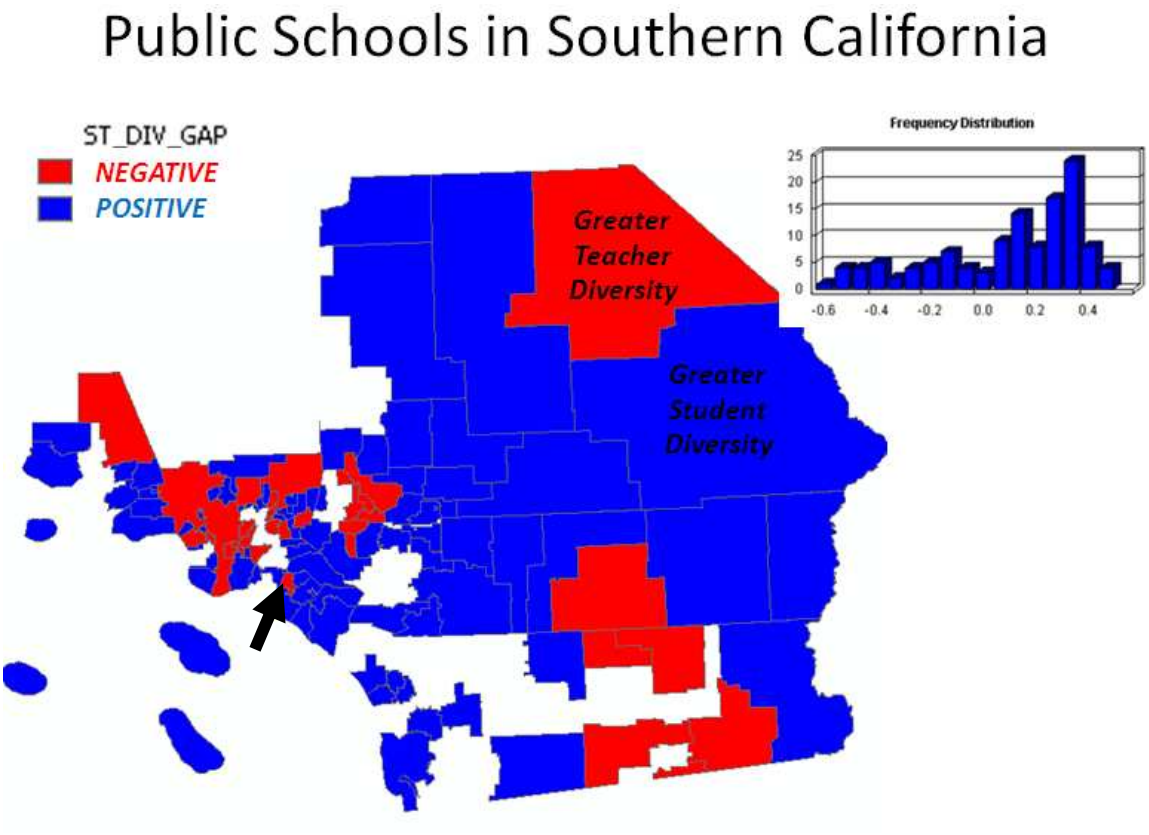


Source: Inland Empire Business Atlas, ©University of Redlands, 2009. Funded in part through a cooperative agreement with the U.S. Small Business Administration. All opinions, conclusions, or recommendations expressed are those of the author(s) and do not necessarily reflect the views of the SBA or the University of Redlands.

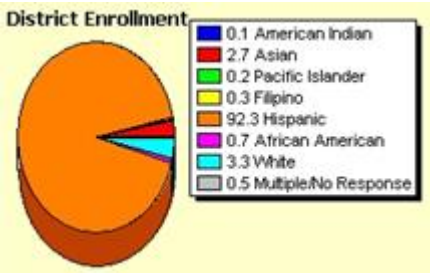
Overall, there appears to be less strong evidence of ethnic diversity, and relatively more strong evidence of spatial association, such that the Diversity Index is similar to its neighbors.

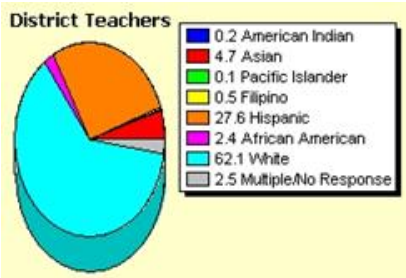
The Gap

Superimposing the teacher layer on the student layer reveals discrepancies in the relative distribution of ethnicities. At a basic level, consider the following display:



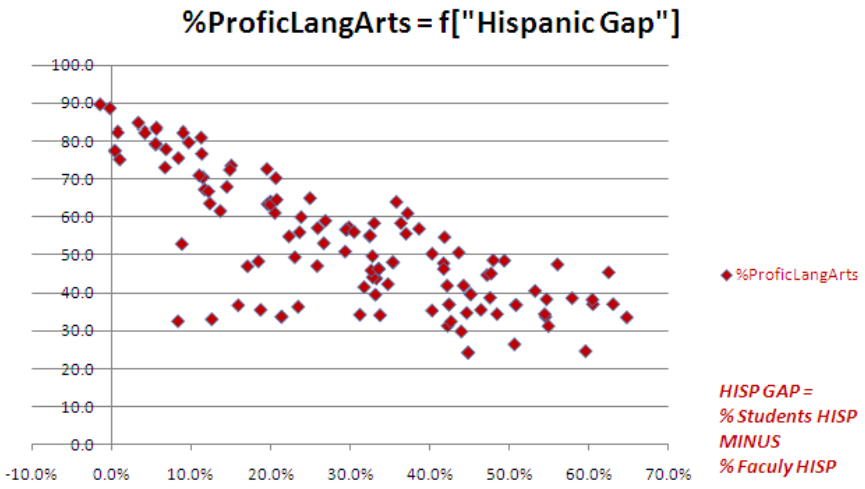
The Student Diversity Index exceeds the corresponding Teacher Diversity Index in 103 of the 124 School Districts. Does this "Diversity Gap" make a difference? Possibly so. Consider just one School District: Santa Ana Unified is a red cell, at the arrow. It has Greater Teacher Diversity, but only because the student population is heavily concentrated in a single ethnic group. Look at the "Gap":





Source: Inland Empire Business Atlas, ©University of Redlands, 2009. Funded in part through a cooperative agreement with the U.S. Small Business Administration. All opinions, conclusions, or recommendations expressed are those of the author(s) and do not necessarily reflect the views of the SBA or the University of Redlands.

Only 33.7% of students in Santa Ana School District are deemed *proficient* in Language Arts. Might that result be different if the Teacher Diversity Profile more closely resemble that of Students? Following is evidence that does not discourage such thinking. In districts with no Gap, the average proficiency is about 75%; For every 10% increase in the Gap, the proficiency drops about 7.5%.

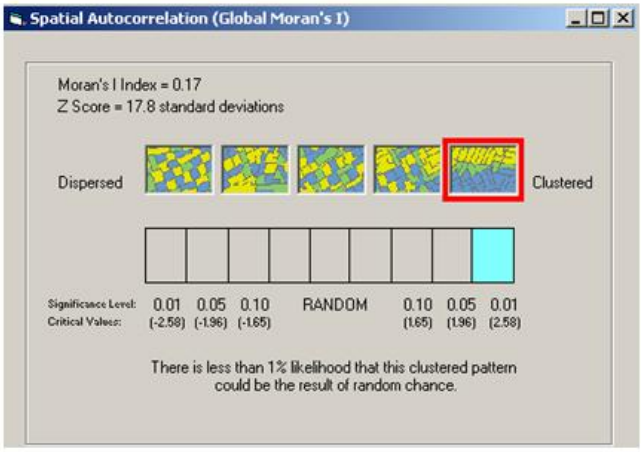


So, let's conclude by looking more carefully at this notion. A simple linear regression can explain variations in *Proficiency in Language Arts* as a linear function of variation in the *Student-Teacher Gap*. For the School District data. Standard (OLS) explains about 1/3 of the total variability – but also indicates strong spatial association that needs to be accounted for in the final analysis.

OLS

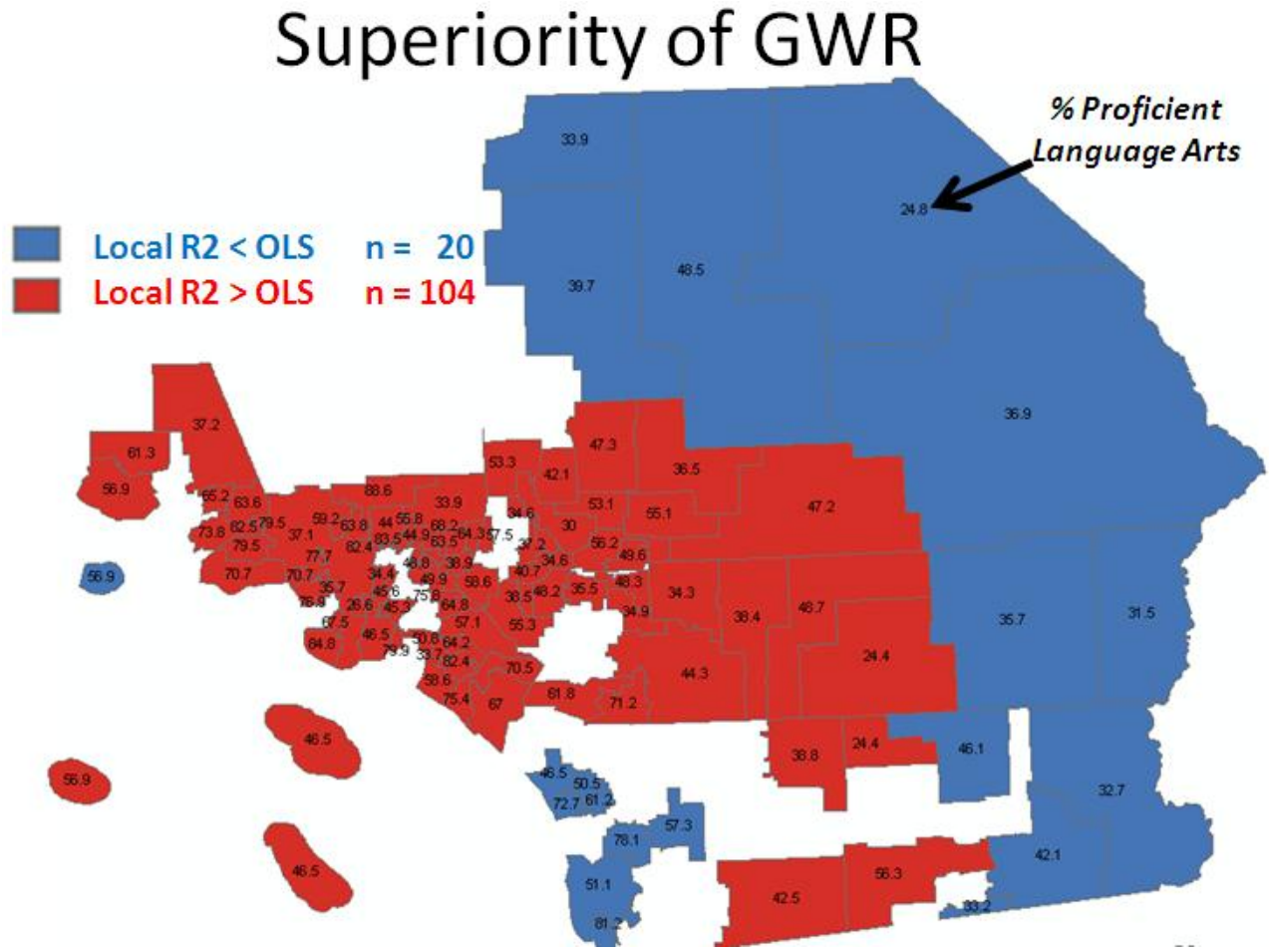
Variable	Coef	StdError	t_Stat	Prob	Robust_SE	Robust_t	Robust_Pr
Intercept	48.3191	1.3696	35.2795	0.0000	1.0638	45.4201	0.0000
ST_DIV_GAP	34.1741	4.3793	7.8035	0.0000	3.5692	9.5747	0.0000

Diag_Name	Diag_Value
AIC	992.5478
R2	0.3348
AdjR2	0.3293
F-Stat	60.8951
F-Prob	0.0000
Wald	91.6758
Wald-Prob	0.0000
K(BP)	9.9323
K(BP)-Prob	0.0016
JB	1.2392
JB-Prob	0.5381
Sigma2	184.1021



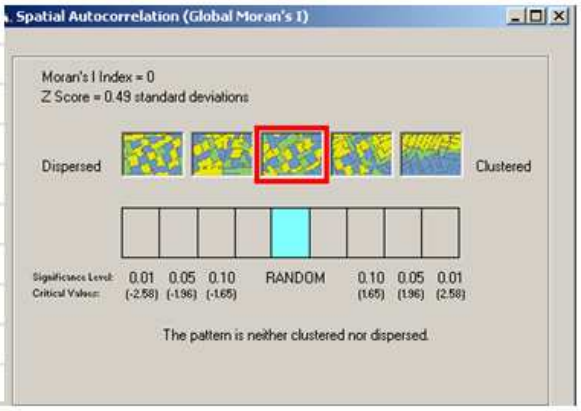
Source: Inland Empire Business Atlas, ©University of Redlands, 2009. Funded in part through a cooperative agreement with the U.S. Small Business Administration. All opinions, conclusions, or recommendations expressed are those of the author(s) and do not necessarily reflect the views of the SBA or the University of Redlands.

Geographically Weighted Regression (GWR) is a means of explicitly capturing spatial associations inherent in the data, and exploiting such patterns to produce locally-varying statistical relationships. In this example, GWR removes all spatial association, and in the process produces localized fits of the model that are superior to the overall model in most instances (104 of the 124 School Districts). The percent of Proficiency now "explained" by the Gap has increased from 33% to 65%. Note that the fit has improved in almost all of the urban School Districts, to the west. The numbers in each cell correspond to the actual Proficiency in Language Arts reported for that School District.

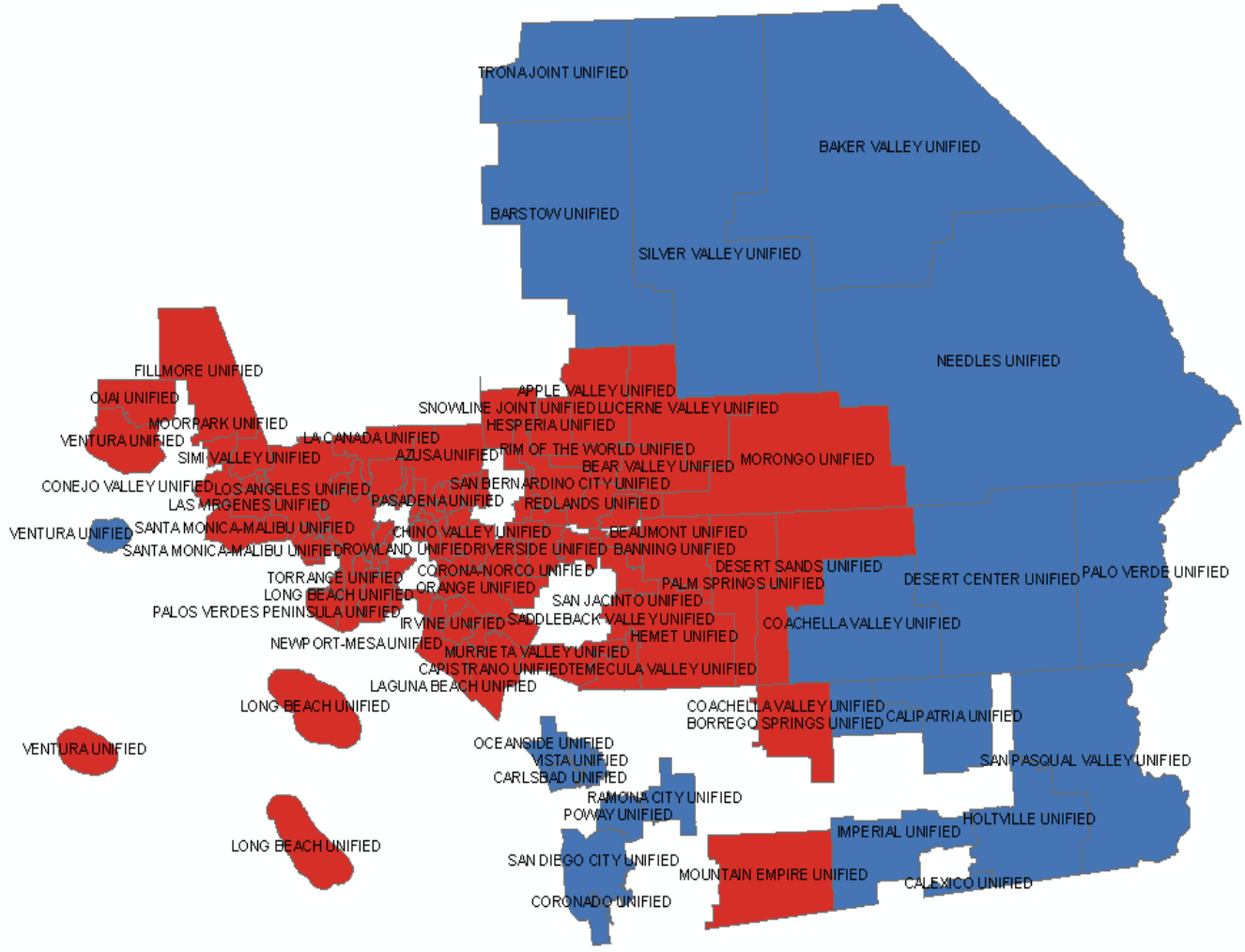


Source: Inland Empire Business Atlas, ©University of Redlands, 2009. Funded in part through a cooperative agreement with the U.S. Small Business Administration. All opinions, conclusions, or recommendations expressed are those of the author(s) and do not necessarily reflect the views of the SBA or the University of Redlands.

NAME	VALUE	DESC
Neighbours	42.00	
ResidualSquares	9787.52	
EffectiveNumber	21.72	
Sigma	9.83	
AICc	927.29	
R2	0.71	
R2Adjusted	0.65	
Dependent Field	0.00	PCTProfici
Explanatory Field	1.00	ST_DIV_GAP



Finally, for easy reference, the School Districts are labeled in this last map:



Source: Inland Empire Business Atlas, ©University of Redlands, 2009. Funded in part through a cooperative agreement with the U.S. Small Business Administration. All opinions, conclusions, or recommendations expressed are those of the author(s) and do not necessarily reflect the views of the SBA or the University of Redlands.