

Bivariate Statistical Legends: Mapping Examples on Florida Covariance Data using Scatterplots from Open Source Visual Analytic Tools

Georgianna Strode, Benjamin Thornton, Victor Mesev, Nathan Johnson
Florida State University

Abstract

Maps representing two variables sometimes use a single combined, or bivariate legend to improve clarity when comparing relationships, and to avoid less convenient side-by-side legends. However, conventional bivariate legends typically omit the underlying bivariate distributions. Using simple scatterplots statistical indicators of covariance can be plotted as points directly on to the choropleth bivariate legend. This allows map readers to not only compare aggregate magnitudes between the two variables but also visualize disaggregate distributions that may represent statistical normality in the data as well as skewness, and linearity. In addition, the covariance distributions can direct class interval selection, or at least inform the reader of which classes represent data abundance or data sparsity. Bivariate statistical legends are tested using examples drawn from Florida population census data at the fine scale of block groups. Practicality is demonstrated by open source software using Data Driven Documents (D3) visual analytic software (<http://d3js.org>).

Keywords: mapping, bivariate legends, statistical covariance, data visualization

Bivariate Maps

With ever-expanding volumes of data, statistical techniques are needed to compress information, without losing clarity. Maps can compress cartographic information from two variables into one map. Such a compound, or bivariate map, attains the obvious advantage of allowing the reader to visualize two spatial distributions by comparing changes in the magnitude of one variable with respect to the other. These inter-relationships are encapsulated by the map's legend as a two-dimensional matrix, typically composed of shaded classes when applied to choropleth representation. In practical terms, bivariate mapping can aid in the exploration of the varied permutations between many spatial variables. For example in social science, the comparison of a map representing ethnic minorities with that of one representing overcrowding in housing can determine levels of poverty for targeting public spending; or maps that represent the interplay of land use zoning and real estate prices as indices for land economics. Such applications have ensured that bivariate mapping has now become an established technique (Olson, 1981; Dunn, 1989), with developments on the bivariate legend to incorporate more inventive use of color (Trumbo, 1981; Eyton, 1984; Robertson & O'Callaghan, 1986), and more intuitive symbologies (Carstensen, 1984).

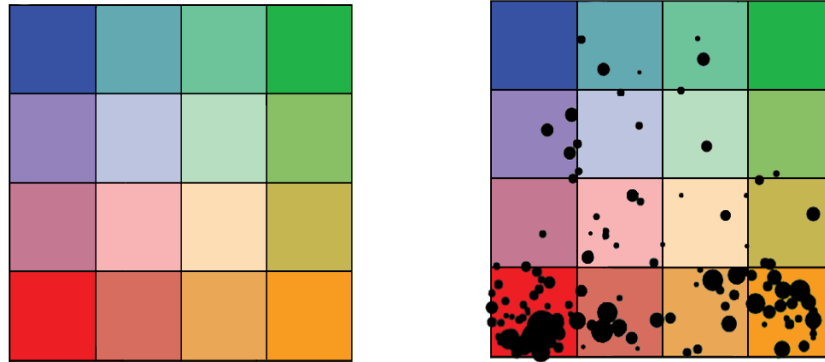


Figure 1. Traditional (left) and statistical (right) bivariate legends

More recently, bivariate legends have included measures that represent the statistical covariance of the data from the two variables being mapped (Marmelejo-Ramos, 2014). This is typically a cartographic process that plots point symbology directly on to the shaded classes of the legend. The points represent the co-occurrence of the individual events of the two variables. The intention is to allow map readers to not only compare aggregate magnitudes between the two variables but also visualize disaggregate distributions, such as statistical normality in the data, as well as evidence for skewness and linearity. In addition, covariance tendencies can help determine class interval selection, or at least inform the reader of which classes represent data abundance or data sparsity (Figure 1). Clusters or absences of data can then be mapped as ‘hot’ and ‘cold’ spots respectively (Herrmann & Pickle, 1996). Cartographically, the bivariate statistical legend has four visual progressions: low values to high values for variable X, low values to high values for variable Y, positive relationship of low X and low Y to high X and high Y (lower left corner to upper right corner of legend), and negative relationship of high Y and low X to low Y and high X (upper left to lower right of legend) (Olson, 1975). The progressions can be used to measure for covariance statistics, including any correlation or lack of significant statistical correlation, as well as evidence for similarity, a measure which relates to the tendency for variable pairs to come from similar positions (or z-scores) (Carstensen, 1986). Similarity can also be plotted on a continuous ratio scale or categorized with labels such as ‘very similar’ or ‘very dissimilar’ (Carstensen, 1984). Traditional, non-statistical bivariate legends omit all these statistical variances in the data, and give the impression that all classes have equal weighting, regardless of the underlying frequency or event counts (Kumar, 2004). In their favor, they display less information, are less likely to distract user attention (Harrower, 2003) and are less likely to dissuade the user from viewing a map when faced with unfamiliar symbolism (Robinson, 1995). But we feel these are overly cosmetic advantages, and maps drawn with statistical bivariate legends allow more experienced readers greater insights in not only the covariance relationship but how the data can be further manipulated by GIS.

Bivariate maps by GIS data visualization

Bivariate mapping is a technique that is implemented and visualized by digital graphic tools, and is part of the operations found in GIS. Some visual analytic graphics used to convey statistical information in a bivariate legend include simple bar charts and histograms where frequency counts and visual weightings are cartographically represented by the height of the bars (Dykes et al, 2010). More complex techniques, shown in Figure 2, such as treemaps determine data hierarchy using compact graphic space (Tobler, 2004) as well as placement, size, and color to convey other data characteristics (Wood & Dykes, 2006). Box plots show overall aspatial distribution, arithmetic range, and frequency counts for map classes. Ogive cumulative frequency diagrams display the statistical distribution of data without using data aggregation or class intervals (Cromley, 2006). Three-dimensional cartograms use varying heights of the geographic unit to represent the value of the statistic being mapped (Reveiu & Dârdala, 2011). Feature-expression heat maps combine a matrix table structure to facilitate visual exploration of multivariate data, and where the interior table cells are ordered by cluster analysis or use circle symbology with varying color and size to represent statistical significance and associations. Finally, sparklines have been expanded into ‘bandlines’ to convey additional information in the same legend space (Few, 2013). Needless to say, the bivariate map has immense potential to convey complex statistical relationships well beyond the capabilities of the standard choropleth technique.

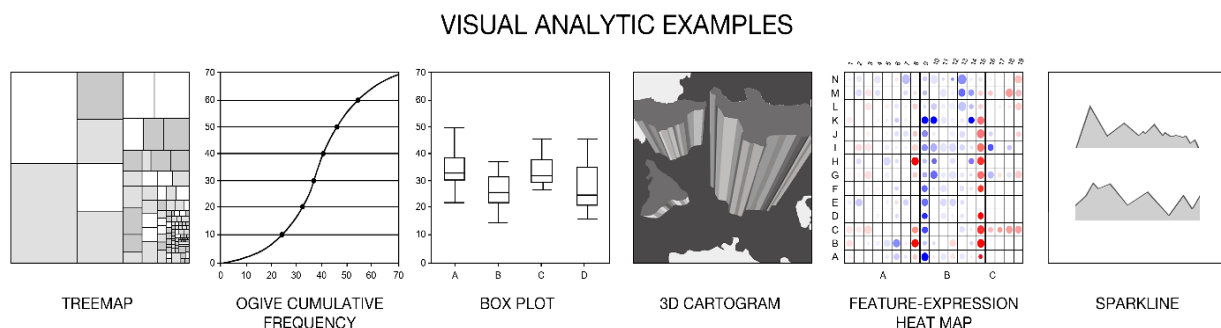


Figure 2. Examples of complex visual analytics

Scatterplots and Open Source Mapping on Florida Data

One further technique is the scatterplot, a simple way to display two-dimensional co-occurrence point clouds directly onto the bivariate legend. However, it also facilitates the calculation of positive and negative linear covariance tendencies, as well as correlation coefficients. Scatterplots display data using colors to represent classes, glyphs to represent variables, symbol size to represent value and magnitude (Jacoby, 1998), data flow, and confidence levels (Yu-Hsuan et al, 2013).

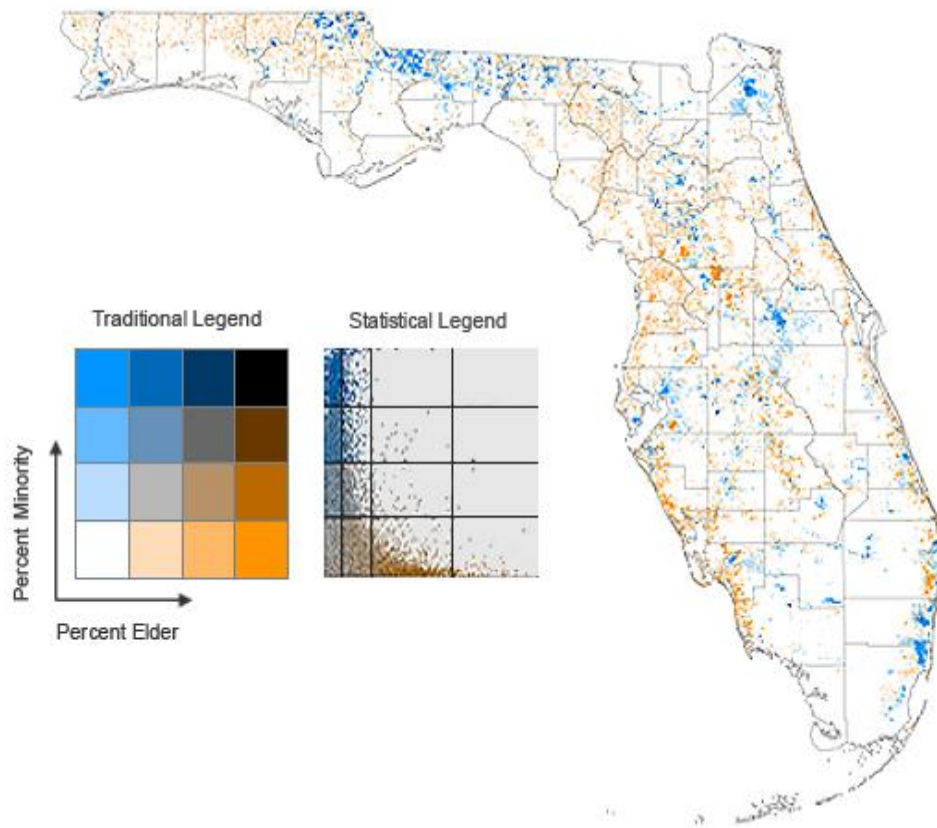


Figure 3. Traditional and statistical bivariate mapping in Florida (elderly population and ethnic minorities) at a 1-km gridded scale

Figure 3 illustrates the use of a scatterplot to enhance the legend of a bivariate map. When compared to a traditional legend, the scatterplot allows the user to visualize, in this case, the statistical distribution of the elderly population with respect to ethnic minorities in US state of Florida from the Population Census. The immediate visual impact of the scatterplot lacks positive covariance, where an increase in elderly would correlate to an increase in ethnicity minorities. The legend is an example of output generated by the web-based GIS software DIY ('do-it-yourself') Florida Census Data Scatterplot Maker (<http://freac.fsu.edu/scatterplots/>). It is hosted by the software outreach enterprise, Florida Resources and Environmental Analysis Center (FREAC) at the Florida State University, and uses open source web-based programming tools from D3 (Data-Driven Documents) (<http://d3js.org/>). D3 is a Java script library (with standard web elements, HTML, SVG, and CSS) for customizing graphic data visualizations, and includes tools to plot simple bar charts, line graphs, boxplots, and co-occurrence matrices, as well as more complex treemaps and circle graphs. All can be static or interactive, and all can embed graphics within web pages or manipulated by databases by WebGIS (Viau, 2012; Bostock 2015). The Florida Census Data Scatterplot Maker allows anyone access to examine possible relationships between two census variables, such as race, age, education,

transportation by Florida county. The results are pushed by D3 code where multiple scatterplots are displayed to the user and can be incorporated into bivariate maps. Examples are shown in Figure 4 where relationships of the census variables, ethnic minorities and population classed as black are plotted for three Florida counties. Volusia County has a more linear pattern with a pocket of non-black minorities; Miami-Dade County shows fewer black population minorities; and Palm Beach County, in addition to being geographically located between the other two counties, has minority data patterns that are between those of the other counties.

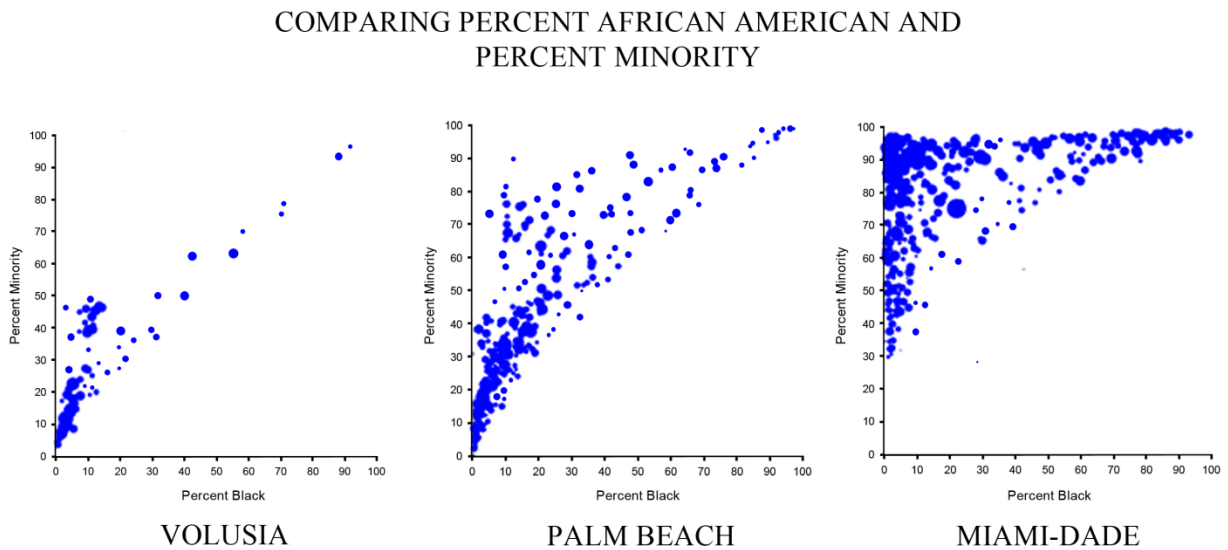


Figure 4. Scatterplots showing percentages of ethnic minority and black at census block group for three Florida counties.

Other than trends in the bivariate distribution scatterplots also facilitate the calculation of statistical metrics. Figure 5 demonstrates the median housing rent by population groups, where the white population appears to have a positive covariance, the black population a negative covariance, and the Hispanic population more of a random distribution. If a strong trend is suspected statistical significance can be calculated with a simple linear regression of ordinary least squares and with one explanatory variable Figure 6 reports a lack of significance and no correlation between the census variables housing rent and travel to work under 30 minutes. Nevertheless, the lack of correlation can still be accommodated within the bivariate legend as classes of 'very dissimilar' association.

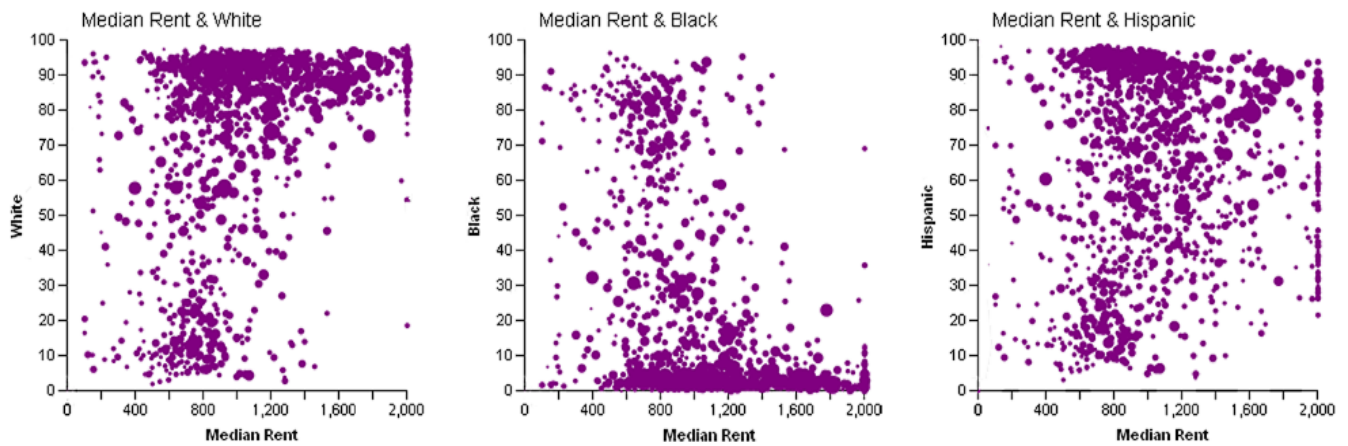
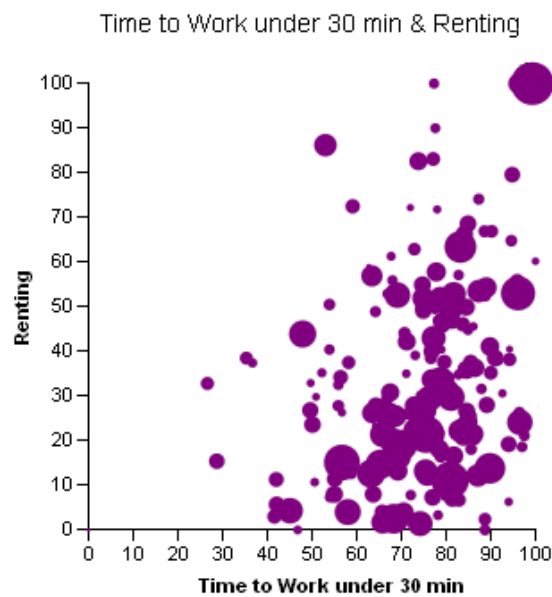


Figure 5. Scatterplots showing median rent and population group



Slope: 0.35
 Intercept: 9.09
 R^2 : 0.06
 No correlation

Figure 6. Scatterplot of no correlation between renting and travel to work under 30 minutes

Further statistical indices are illustrated by Figures 7, 8, 9, and 10 measuring median values, trend lines, compactness, and outliers respectively. Median values allow bivariate legends to be apportioned into four classes, with each class containing one quarter of the total bivariate distribution; trend lines indicate the direction and gradient of the linear tendency;

compactness determine variance; and variability, and outliers identify vital anomalies for further investigation.

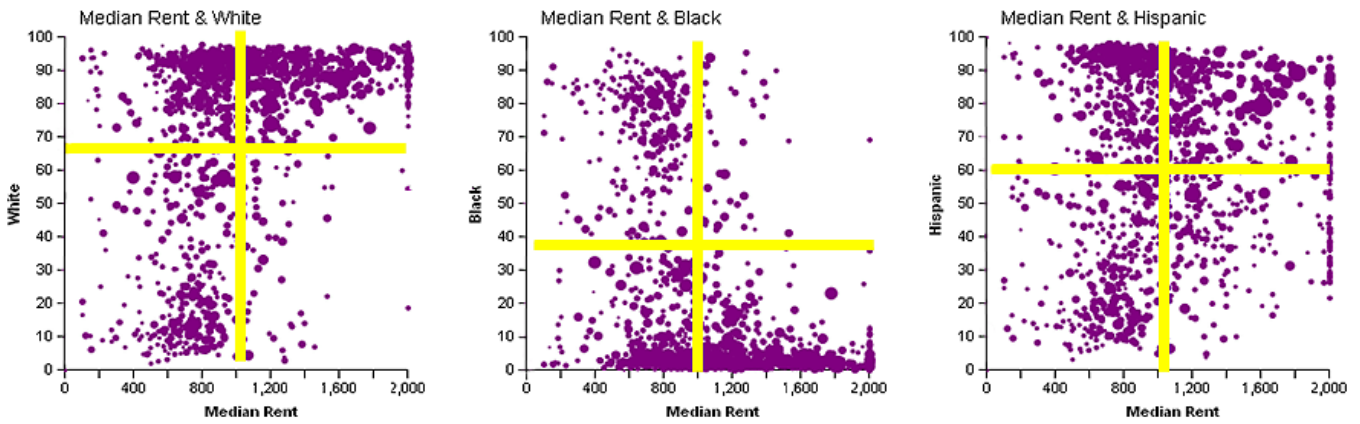


Figure 7. Scatterplots with median values separating the total bivariate distribution into four quadrants

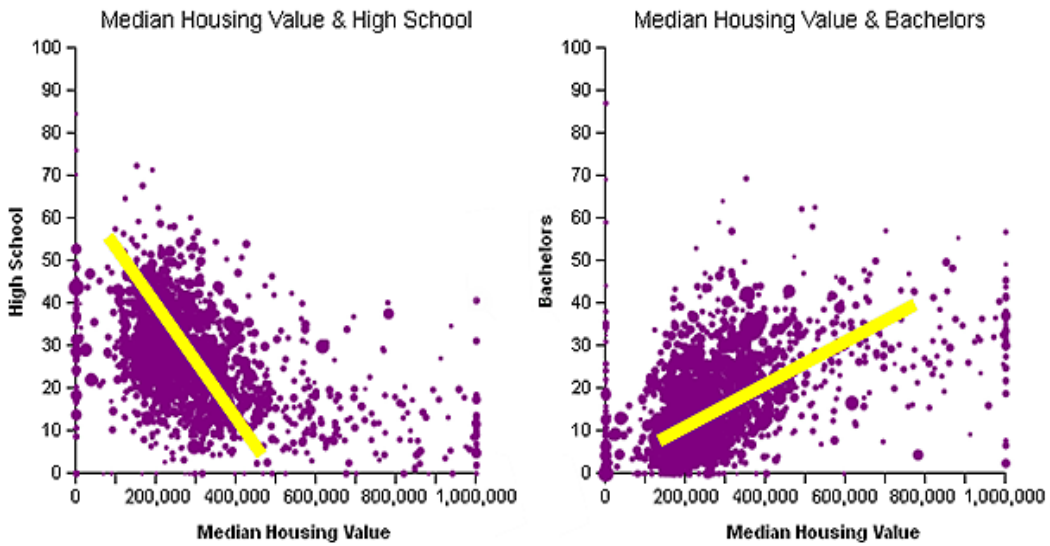


Figure 8. Trend lines showing direction and gradient

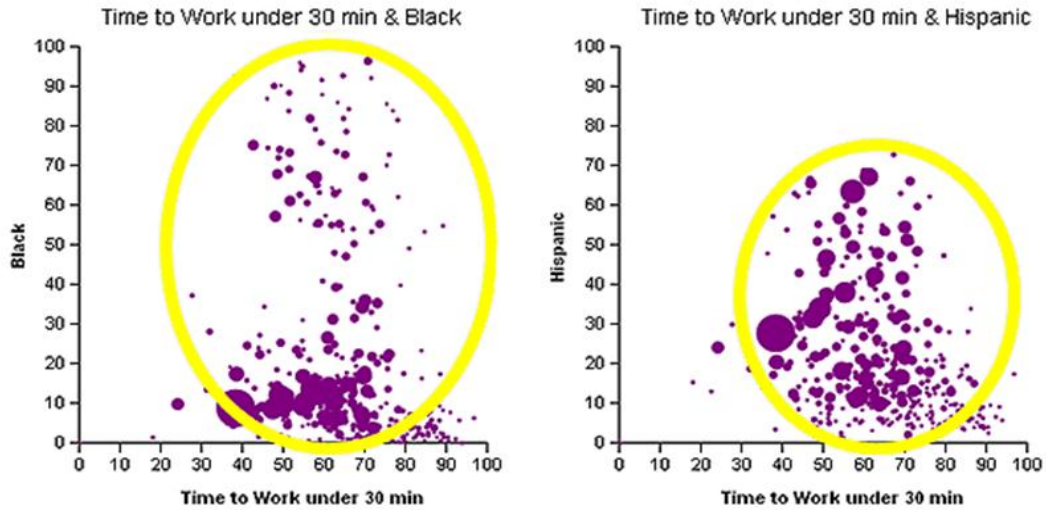


Figure 9. Compactness (wider range on left shows more variability than compact area on right.)

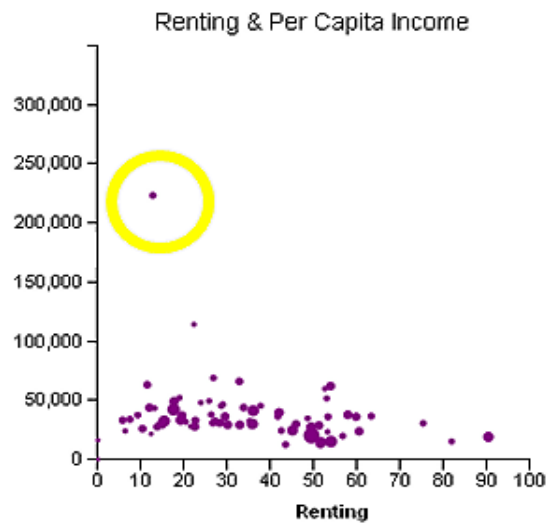


Figure 10. Scatterplot with an outlier

Bivariate legends that use scatterplots representing these statistical metrics can be generated with the D3 software. Users can access pre-existing default code for a simple scatterplot, and then apply graphic edits in proprietary software such as ArcMap. The code for the DIY Florida Census Data Scatterplot Maker is available for download and can be converted for use with any spatial data, not just data from the population census. The D3 scatterplots can be screen captured, edited, and saved as a raster image for further use. The downside of this approach is that it is labor-intensive, requiring editing, and only suitable for developing a

limited number of bivariate statistical legends for print or electronic maps. Figure 11 shows two scatterplots, one unedited and the other with enhancement into classes.

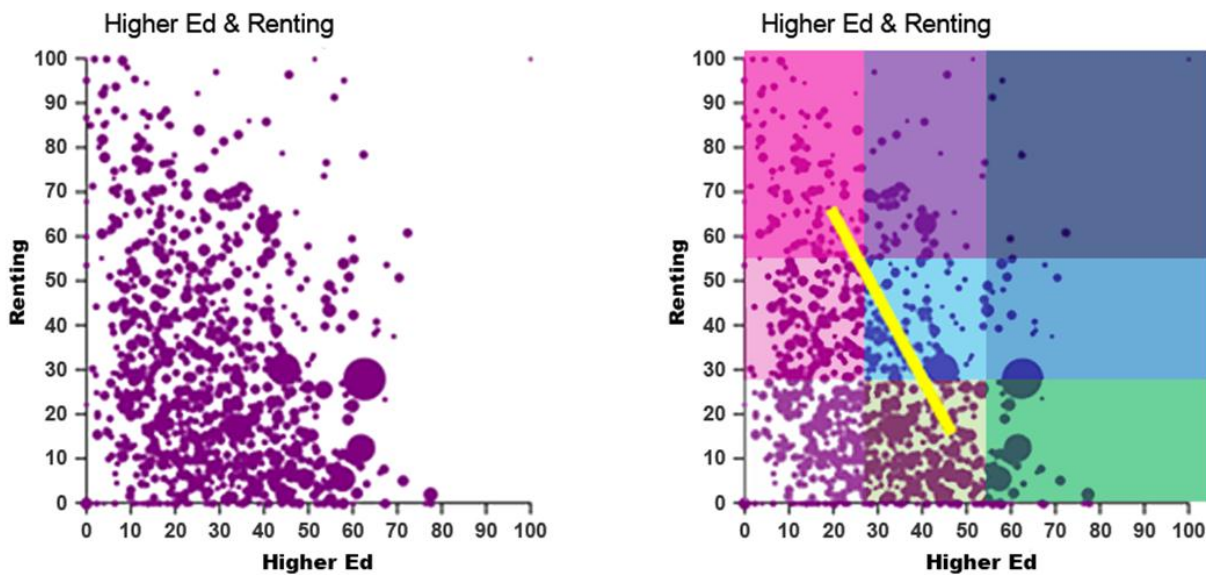


Figure 11. Statistical bivariate legends showing the results of graphic editing (left is an unedited result of D3, right has been edited to show classifications, colors, and trend line.)

Effects of Class Intervals

As in all cartographic representation choosing class intervals is one of the most important decisions in univariate choropleth mapping. Determining the class breaks can considerably alter the appearance of the map and in turn how variables are grouped categorically. The equal interval method divides the data independently of their statistical distribution, while quantiles attempt to place the same number of observations in each class. The natural breaks method uses data clustering to reduce variance to find a “natural” order in the data (Jenks, 1967). Compromises in all three are unavoidable given that it has now been established that typically five or six classes is the optimum number that most map readers can handle. Bivariate mapping is no exception to this rule, and in fact the use of two variables only compounds the problem. We explore equal interval, quantile and natural breaks in our next three figures (12, 13, and 14), where each visualizes the relationships of percentage Hispanic and percentage renting, both in Orlando, Florida. They illustrate the effect each has on the construction of bivariate maps, along how statistical legends provide the reader with a clearer picture of aspatial data distribution and categorical divisions. Figure 12 uses an equal interval classification method which sets breakpoints evenly within the data range irrespective of the shape of the data distribution, and as such it is best applied to data with a relatively even

distribution in all data bins, and not a good choice for data that is highly clustered or skewed. The scatterplot shows that most of the data values lie within the lowest x and lowest y classifications, and produces a map with block groups of the lowest bivariate class (in pink) and few in the highest (shown in green). The statistical legend communicates this unequal representation with a scatterplot that is heavily populated towards the bottom-left corner.

Figure 13 illustrates a quantile classification method that attempts to separate observations into roughly the same number of classes. This method is particularly suited for linearly distributed datasets, and not for identifying individual observations that deviate as outliers. The bivariate map contains an even distribution of classes because each contains an equal number of observations. This is reflected by the statistical legend which re-sizes automatically to adjust to the scatterplot where each class has roughly the same number of observations. Users can visually inspect the distribution and even determine if narrower ranges are necessary.

Finally, Figure 14 uses the Jenks natural breaks method to produce class intervals that represent obvious clusters in distribution of observations. As method seeks to identify breaks in the data to use as natural classification breakpoints it relies on variance between classes, and does not perform well for data with low variance. It also does not perform well for data with a single large cluster. The bivariate map in Figure 13 looks similar to the quantile map in Figure 12 because of data distribution clusters in multiple areas that happen to align with a quantile classification. A data set with a stronger correlation would produce more uneven representation of the data across various classes. The accompanying statistical legend visually conveys the data distribution and the classification breakpoints.

Class interval selection is an important part of mapmaking. Traditional bivariate legends often do not provide the reader with adequate information. Many have equal sized classes that imply an equal interval classification when this may not be the case. Statistical legends can effectively communicate both data distribution and classification choice to the reader.

Examining Hispanic and Rental Rates

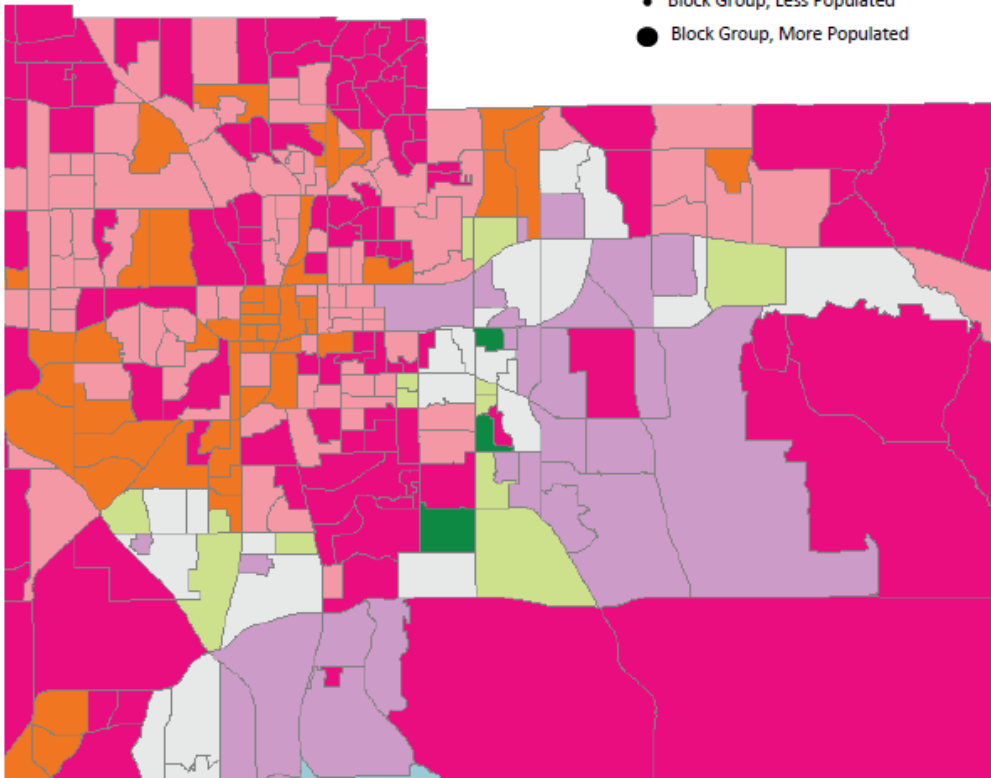
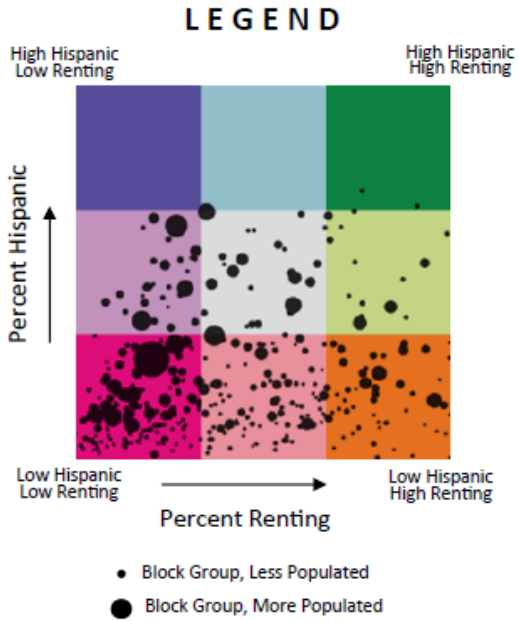
Orlando, Florida

Map Purpose

To highlight areas where Hispanic population and rental rates do not align. This “corner” color palette emphasizes the interaction of high and low areas.

Equal Interval Classification

Data are divided into classes of equal size regardless of data distribution.



Data Source: Census Block Groups 2010

Figure 12. Equal interval classification

Examining Hispanic and Rental Rates

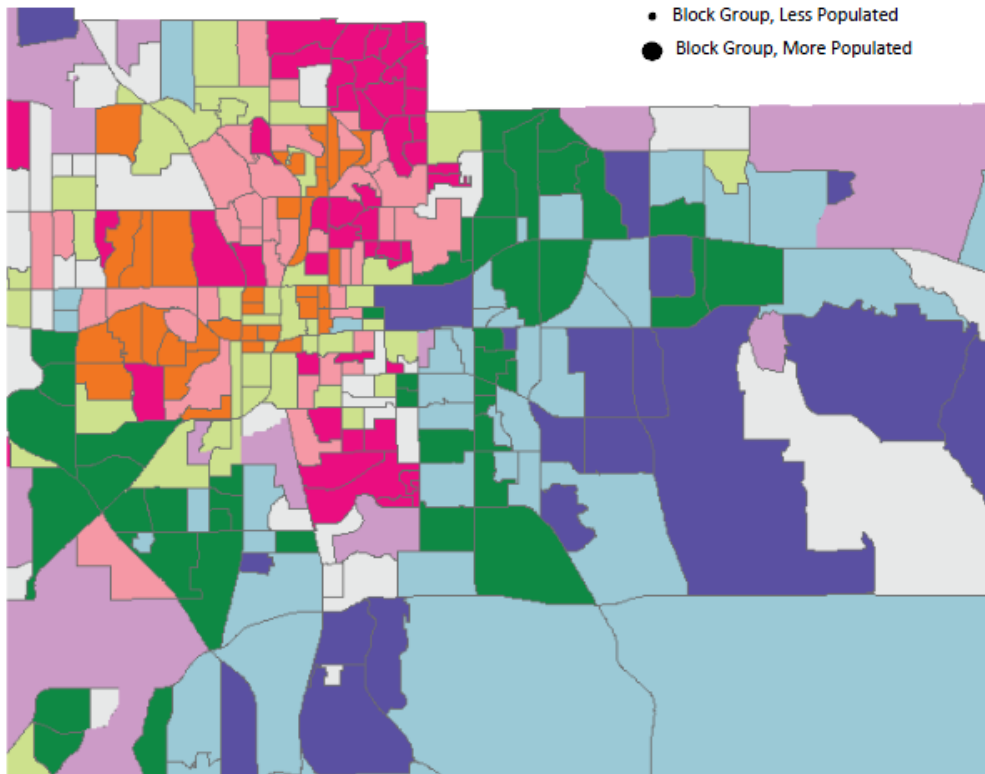
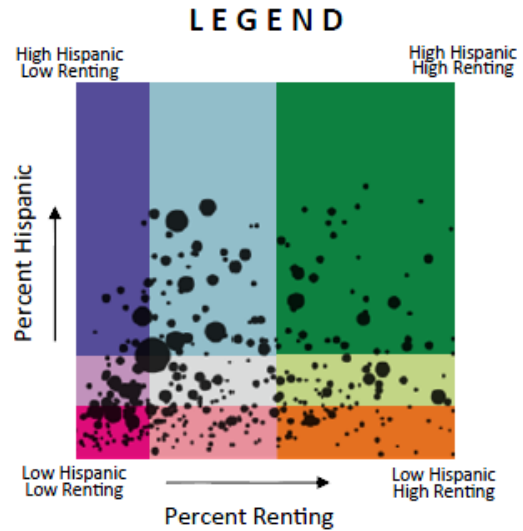
Orlando, Florida

Map Purpose

To highlight areas where Hispanic population and rental rates do not align. This “corner” color palette emphasizes the interaction of high and low areas.

Quantile Classification

Data are divided so that each class holds the same number of occurrences.



Data Source: Census Block Groups 2010

Figure 13. Quantile classification

Examining Hispanic and Rental Rates

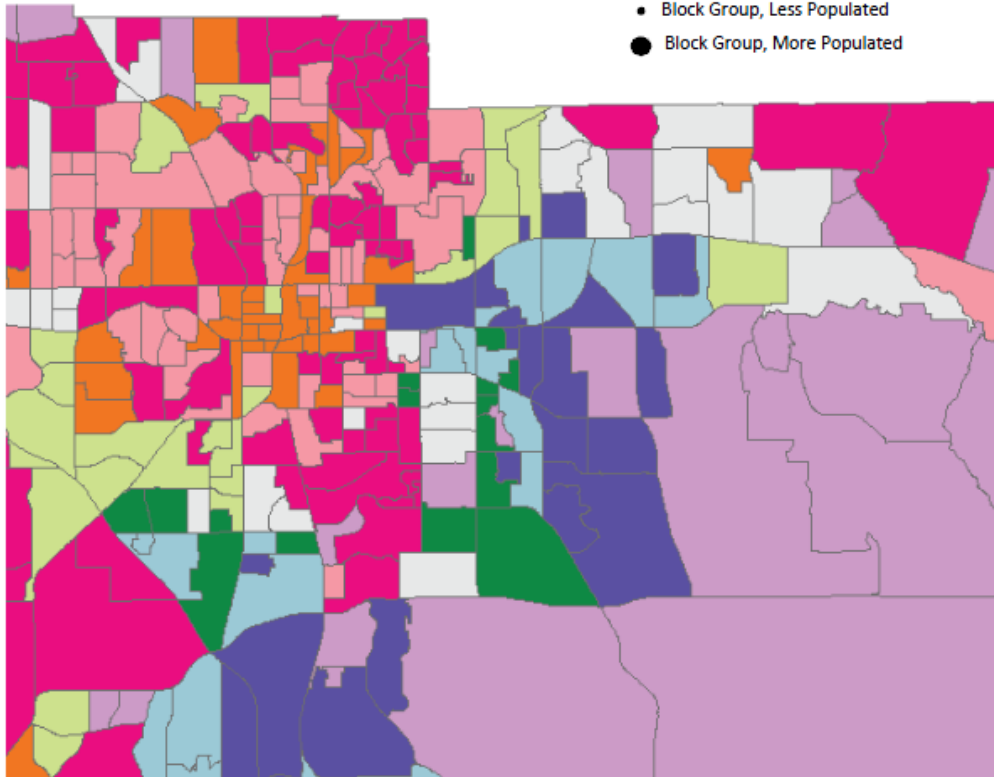
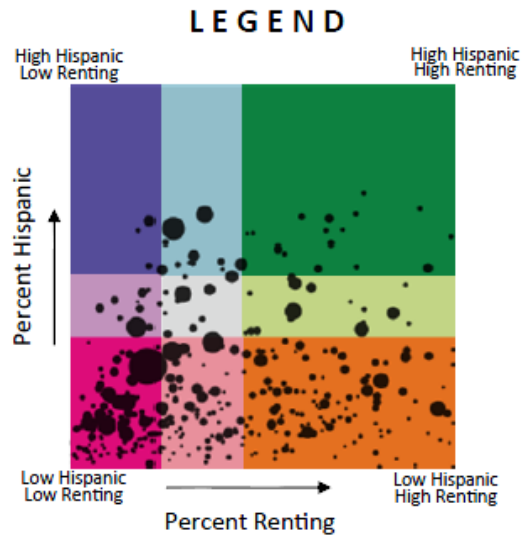
Orlando, Florida

Map Purpose

To highlight areas where Hispanic population and rental rates do not align. This “corner” color palette emphasizes the interaction of high and low areas.

Jenks Natural Breaks Classification

The categories are determined by data clustering to reduce variance within each class.



Data Source: Census Block Groups 2010

Figure 14. Jenks Natural Breaks classification

Conclusion

Bivariate mapping facilitates immediate comparison of two variables. Traditional bivariate legends represent an aggregation of the bivariate distribution into categorical classes. By utilizing the simple scatterplot, bivariate legends can display the point clouds of the entire bivariate distribution allowing users to visualize covariance, trending and correlation. However, few mapping software packages allow such modifications to the bivariate legend. In response we have identified web-based open source software, D3 to allow users to develop their own legends, and have demonstrated examples from the DIY Florida Census Data Scatterplot Maker for creating scatterplots of variables of population, housing, and transportation for Florida at the census block group level.

Acknowledgements

The authors would like to thank Thomas Tricarico for his D3 code optimization and Brittney Gress for her graphic enhancements.

References

- Bartholomeus C.M. (Benno), Haarman, R. F., Riemersma-Van der Lek R.F., Willem A. Nolen a, Mendes, R., Drexhage, H.A., Burger, H. 2015. "Feature-expression heat maps – A new visual method to explore complex associations between two variable sets." *Journal of Biomedical Informatics* 53:156–161.
- Bostock, M. "D3.js - Data-Driven Documents." *D3.js - Data-Driven Documents*. N.p., n.d. Web. 03 Nov. 2015.
- Carstensen, L. W. 1984. "Perceptions of variable similarity on bivariate choroplethic maps." *The Cartographic Journal* 21.1: 23-29.
- Carstensen, L. W. 1986. "Bivariate choropleth Mapping: The Effects of Axis Scaling." *The American Cartographer* 13(1): 27-42.
- Cromley, R. G., Yanlin Y. 2006. "Ogive-based legends for choropleth mapping." *Cartography and Geographic Information Science* 33.4 (2006): 257+.
- Cleveland, W. S., McGill, R. 1984. "The Many Faces of a Scatterplot." *Journal of the American Statistical Association* 79: 807-822.
- Dunn, R. 1989. "A Dynamic Approach to Two-Variable Color Mapping." *The American Statistician* 43 (4):245-252.
- Dykes, J., Wood, J. Slingsby, A. 2010. "Rethinking map legends." *IEEE Transactions on Visualization and Computer Graphics* 16(6):890-899.
- Eyton, R. 1984. "Complementary-Color, Two-variable Maps." *Annals of the Association of American Geographers* 74 (3): 477-490.
- Few, S. 2013. "Introducing Bandlines: Sparklines Enriched with Information about Magnitude and Distribution." *Perceptual Edge, Visual Business Intelligence Newsletter*. January/February/March 2013.
- Haarman, C.M., Riemersma-Van der Lek, R.F., Nolen, W.A., Mendes, R., Hemmo A. Drexhage, Huibert Burger. 2015. "Feature-expression heat maps – A new visual method to explore complex associations between two variable sets." *Journal of Biomedical Informatics*, 53:156-161.
- Halliday, S. 1987. "Two-variable choropleth maps: an investigation of four alternate designs." Thesis.

Harrower, M. 2003. "Tips for designing effective animated maps." *Cartographic Perspectives*, 44:63–65.

Herrmann, D., Pickle, L. W. 1996. "A Cognitive Subtask Model of Statistical Map Reading." *Visual Cognition*, 3(2), 165-190. doi:10.1080/135062896395715

Jacoby, W. G. 1998. "Multiple-code plotting symbols in scatterplots." *Statistical Graphics for Visualizing Multivariate Data*. (pp. 11-18). Thousand Oaks, CA: SAGE Publications, Inc. doi: <http://dx.doi.org/10.4135/9781412985970.n2>

Jenks, G. F. 1967. The data model concept in statistical mapping. *International yearbook of cartography*, 7(1), 186-190.

Kumar, N. 2004. *Frequency Histogram legend in the Choropleth Map: A Substitute to Traditional Legends.* Cartography and Geographic Information Science 31:4, 217-236, DOI: 10.1559/1523040042742411

Leonowicz, A. 2006. "Two-variable choropleth maps as a useful tool for visualization of geographical relationship." *Geografija* 42(1): 33-37.

Marmolejo-Ramos, F. 2014. "Current topics in statistical graphics." *Revista Colombiana De Estadística*, 37(2): 1-4.

Meehan, G. B. 1984. An Alternative Legend for the Choropleth Map: The Box Plot Array (Statistical Graphics, Map Design, Cartography, Thematic) (Order No. 8508606). Available from ProQuest Dissertations & Theses Global. (303306698). Retrieved from <http://search.proquest.com/docview/303306698?accountid=4840>

Olson, J.M. 1975. "The organization of color on two-variable maps." *Proceedings AUTOCARTO II*. Falls Church, Va.: American Congress on Surveying and Mapping and the U.S. Bureau of the Census.

Olson, J.M. 1981. "Spectrally Encoded Two-variable Maps." *Annals of the Association of American Geographers*, Vol. 71, No. 2 (Jun., 1981), pp. 259- 276.

Perdue, N. 2013. "The Vertical Space Problem: Rethinking Population Visualizations in Contemporary Cities." *Cartographic Perspectives* 74: 9-27.

Pickle, L., Mungiole, M., Jones, G.K., White, A.A. 1996. Atlas of United States mortality. DHHS Publication No. (PHS) 97-1015. National Center for Health Statistics, Hyattsville, Maryland.

Reveiu, A., & Dârdala, M. 2011. "Techniques for statistical data visualization in GIS." *Informatica Economica*, 15(3), 72-79. Retrieved from <http://search.proquest.com/docview/912868818?accountid=4840>

Robertson, P.K.; O'Callaghan, J.F. 1986. "The Generation of Color Sequences for Univariate and Bivariate Mapping," in *Computer Graphics and Applications, IEEE*, vol.6, no.2, pp.24-32, Feb. 1986

doi: 10.1109/MCG.1986.276688

URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4056802&isnumber=4056791>

Robinson, A.H., Morrison, J. L., Muehrcke. P. C., Kimerling, A.J., Guptill, S. C. 1995. *Elements of Cartography*. Wiley, New York, NY.

Tobler, W. 2004. "Thirty five years of computer cartograms." *Annals of the Association of American Geographers*, 94:58–73.

Viau, Christophe (June 26, 2012), "[What's behind our Business Infographics Designer? D3.js of course](#)", *Datameer's blog*, retrieved October 19, 2015
<http://www.datameer.com/blog/author/cviau>

Wood, J., Dykes, J. 2008. "Spatially ordered treemaps." *IEEE Trans. Vis.Comp. Graphics*, 14(6):1348–1355.

Wood, D., Fels, J. 2008. "The natures of maps: Cartographic constructions of the natural world." *Cartographica*, 43(3):189–202.

Yu-Hsuan Chan, Correa, C.D., Kwan-Liu M. 2013. "The Generalized Sensitivity Scatterplot." *Visualization and Computer Graphics, IEEE Transactions* 19 (10):1768-1781.
doi: 10.1109/TVCG.2013.20