

On Statistical Approximations of Geographical Maps

Jerry Platt

School of Business, University of Redlands
Redlands, CA 92373
Tel: +1 909-335-4041
Fax: +1 909-335-3400
Email: Jerry.Platt@redlands.edu

Abstract

Maps can be generated directly from coordinate systems, or indirectly from distance measures or other dissimilarity indices among objects. This paper reviews methods for generating two-dimensional maps, and for comparing maps drawn from alternative data sources. Particular attention is given to using the open-source R system for statistical computing to assess the statistical significance of lack-of-fit of any one map to another.

1. Introduction

There are many situations where one wants to generate a simple map to visually represent empirical evidence, and there are other situations where one wants to compare and contrast two maps of presumably related information. Here are some examples:

- An ancient map from some long-forgotten European cartographer is discovered, and the question is degree of accuracy relative to a modern rendition with carefully calibrated coordinates;
- The financial geography of investment houses shifts following September 11, 2001, and the question is whether the “location” of Wall Street has changed;
- Attributes of a face are compared against a data bank of known faces, and the question is whether a match can be found with a desired degree of confidence;
- Firms are mapped in space relative to their perceived strengths among consumers, and again relative to their self-perceived strengths, and the question is whether perceptions differ;
- Maps are drawn of things people fear, and the question is whether maps of male subjects differ from those of females, or whether Americans differ from Asians;
- Political candidates are mapped in political attribute space, and the question is whether candidate positioning has changed following a televised debate.

Note that geo-referenced space is not necessary for the consideration of map construction and comparison that is the subject of this analysis.

2. Map Constructions with Principal Components and Multidimensional Scaling

When there are measures on exactly two attributes for each of many objects, and when orthogonal projection onto a space makes sense, map construction is a simple task. However, when the number of attributes is greater than the number of dimensions to the desired map, or when data at the object level is not directly measurable, the task is less straight-forward.

2.1. From Data Observations to Maps

Most maps are generated from raw data. The simplest example is to draw a map from the coordinates of latitude and longitude for each of several locations. Given n locations, the data matrix is of dimension $n \times 2$. More generally, there may be p attributes associated with each of n locations, in which case the $n \times p$ matrix can serve as input to a principal component analysis, and the first two principal scores can serve as coordinates in drawing a two-dimensional map approximation of the original p -dimensional space.

2.2. From Dissimilarity Measures to Maps

Sometimes, however, data is not available at the specific location level. For example, attributes about individual locations may not be available, but distances between locations are observable. Multidimensional scaling allows one to reverse engineer the generating process for a map, and the first two dimensions of an MDS solution often do quite well at reproducing a two-dimension map.

3. Map Comparisons with Statistical Measures and Tests

3.1. Classical Tests v. Permutation Tests

A classical test of statistical significance relies upon parametric distributional assumptions to test a null hypothesis of independence between sampling mechanisms in two populations. Computer-intensive tests address the same issue, but substitute raw computing power for the more elegant mathematics but less applicable assumptions that underlie parametric distributions. Examples of computer-intensive methods include the jackknife, the bootstrap, and permutation tests, and Mantel developed a permutation test of the null hypothesis that two distance matrices have independent generating mechanisms.

3.2. Mantel Test

The null hypothesis is that distance matrix A is independent of distance matrix B . A measure of association between the matrices is constructed, such as the cophenetic correlation, which is a function of the sum of products of corresponding elements in the lower triangle of the respective matrices. To determine whether the observed correlation is significantly different than 0.0, as implied by independence, the rows and columns of matrix A are randomly permuted, while those of B remain fixed. The correlation is calculated between the randomly permuted matrix A and distance matrix B , and note is taken of whether this correlation is greater than the original cophenetic correlation. If,

after many repetitions of this randomization procedure, the incidence of observed correlations that exceed the original cophenetic correlation is below a set threshold for error tolerance, one concludes that the null hypothesis is not supported by the data.

3.3. Bidimensional Regression

Originally developed in the geography literature but recently rediscovered in psychology, bidimensional regression is a method for measuring the degree of fit of one map to another, after applying transformations, rotations and translations to maximize the degree of fit. The coordinate pairs for observations depicted in map A are regressed on the coordinate pairs for observations depicted in map B. The pair of intercepts estimate the optimal degree of horizontal and vertical translation. The pair of slopes can be transformed to measures of appropriate scale and angle transformation to maximize the extent to which the resulting variation of map A approximates map B.

3.4. Procrustean Analysis

According to Greek mythology, Procrustes was a robber on the outskirts of ancient Athens who lured unsuspecting travelers to his home, where he then tied them to his bed and provided a perfect fit by stretching or chopping the victims as needed. Similarly, Procrustean analysis is a means of applying legal and painless mathematical stretching and twisting to matrix A until it has minimum distortion relative to matrix B. Because maximization of ordinary least squares fit, as measured by R-square, must necessarily lead to the same result as minimization of sum of squared errors, bidimensional regression and Procrustean analysis produce equivalent results from differing directions.

4. Singular Value Decomposition as a Unifying Concept

Multivariate statistical techniques like principal components analysis, multidimensional regression, bidimensional regression, and Procrustean analysis all attempt to obtain reasonable degrees of fit of high-dimension problems in low-dimension space, thereby facilitating visual presentation and interpretation. It should not be a surprise, therefore, to discover that there is a unifying concept that underlies their constructions. That concept is the singular value decomposition, by which any matrix can be recast as the product of three fundamental component matrices. Table 1 shows the derivation of these multivariate methods from building blocks contained in a singular value decomposition.

For those unfamiliar with this remarkable decomposition and its properties, it might be instructive to briefly peek ahead to Figure 4, where the open-source R system for statistical computing is introduced. Note in the example there that matrix *a* is generated in the first line using commands with consequences you might surmise, then object *b* is created to house the three component matrices (*u*, *d*, *v*) from the singular value decomposition of matrix *a*, and finally matrix *c* completes the reconstruction of matrix *a* by following the multiplication process noted in Table 1, presumably confirming your expectation relative to the generation of matrix *a*.

Eigen-Analysis:

C = an NxN variance-covariance matrix

Find the N solutions to $C\alpha = \lambda\alpha$

λ = the N Eigenvalues, with $\lambda_1 \geq \lambda_2 \geq \dots$

α = the N associated Eigenvectors

$C = LDL'$, where

L = matrix of α s

D = diagonal matrix of λ s

Singular Value Decomposition:

Every NxP matrix A has a SVD

$$A = U D V'$$

Columns of U = Eigenvectors of AA'

Entries in Diagonal Matrix D = Singular Values
= SQRT of Eigenvalues of either AA' or $A'A$

Columns of V = Eigenvectors of $A'A$

Principal Component Analysis:

A is a column-centered data matrix

$$A = U D V'$$

V' = Row-wise Principal Components

D ~ Proportional to variance explained

UD = Principal Component Scores

DV' = Principle Axes

Ordinary Least Squares Regression:

$$Y = X\beta + \varepsilon$$

$$Y = Xb + e$$

$$X = UDV'$$

$$b = V_r D^{-1} U_r' Y, \text{ where } r = \text{first } r \text{ columns (} N > P \text{)}$$

$$b = (X'X)^{-1} X'Y$$

$$b = V_r V_r' \beta$$

$$\text{Estimated } Y \text{ values} = U_r U_r' Y$$

Multidimensional Scaling:

A is a column-centered dissimilarity matrix

$$B = -\frac{1}{2} \left[I - \frac{1}{N} ii' \right] A^2 \left[I - \frac{1}{N} ii' \right]$$

$$B = U D V'$$

$$B = XX', \text{ where } X = UD^{1/2}$$

Limit X to 2 Columns

→ Coordinates to 2d MDS

Procrustean Analysis:

Two NxP data configurations, X and Y

$$X'Y = U D V'$$

$$H = UV$$

$$\text{OLS} \rightarrow \text{Min SSE} = \text{tr} \sum (XH - Y)'(XH - Y)$$

$$= \text{tr}(XX') + \text{tr}(YY') - 2\text{tr}(D)$$

$$= \text{tr}(XX') + \text{tr}(YY') - 2\text{tr}(VDV')$$

Bidimensional Regression:

(Y,X) = Coordinate pair in 2d Map 1

$$Y = \alpha_0 + \beta_0 X$$

(A,B) = Coordinate pair in 2d Map 2

$$\begin{bmatrix} E[A] \\ E[B] \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_1 \end{bmatrix} + \begin{bmatrix} \beta_1 & -\beta_2 \\ \beta_2 & \beta_1 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

α_1 = Horizontal Translation

α_2 = Vertical Translation

ϕ = Scale Transformation = $\text{SQRT}(\beta_1^2 + \beta_2^2)$

θ = Angle Transformation = $\text{TAN}^{-1}(\beta_2 / \beta_1) + 180^\circ$
if $\beta_1 < 0$

Table 1. Singular value decomposition as a unifying concept

5. A Small Worked Example in the R Statistical Computing Environment

5.1. A Sample Problem

Southern California (Figure 1) is comprised of seven counties that encompass approximately 350 zip code areas. In this simple example, we use the tools described previously to construct and compare maps of eight zip codes within the region.

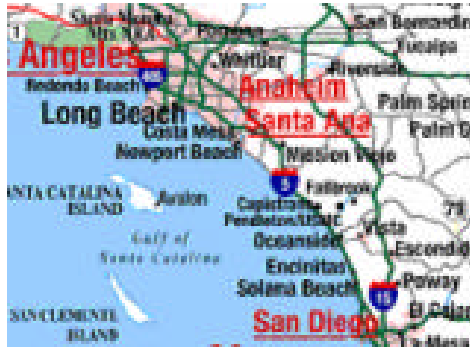


Figure 1. A map of Southern California

5.2. Data Sources

As stated, the seemingly straight-forward way to construct a geographic map is to plot geographic coordinates. Estimates of coordinates for many locations are available through multiple sources, on the internet and elsewhere. One example is shown in Figure 2. Of course, different sources provide different estimates, if only because in this case a zip code area is a polygon rather than a single point. As Figure 3 demonstrates, estimates of latitude and longitude for zip codes vary, as do estimates of distances between zip codes.

Application of the tests described in Section 3 indicates that these differences tend not to be statistically significant. Selecting a single source, geographic coordinate estimates for the eight locations are presented in Table 2. Other data sources are introduced later.

Results for ZIP Code 92373-	
Status Code	5
Latitude	34.040704
Longitude	-117.183107
Census Tract	
Census Block	
County Name	San Bernardino
County FIPS Code	06071

Figure 2. Coordinate data for one data point in a map construction

MELISSA DATA Products & Services Downloads Lookups Support Contact Site map

My Account Search Go

Distance between ZIP Codes

This program displays the distance between any two 5-digit ZIP Code in the United States. Enter the two ZIP Codes and click on **Calculate**.

Enter First ZIP Code: Enter Second ZIP Code:

REDLANDS, CA SAN DIEGO, CA

Distance from first to second ZIP Code is **87.5 miles**

Zip Code Database Download: Digital Zip Code Database & Maps Wednesday, May 18, 2005

ZIP-CODES.COM

Home About United States ZIP Codes FAQs The Two "Mr. ZIP Codes" MSA/CBSA Information Download our Digital Zip Code Maps Download our Zip Code Database Account Login

Free Search

Zip Code:
 Area Code:
 State:
 County:
 Town/City:

Only ONE field is required for your search.

Distance Calculator

Calculate the distance between two U.S. ZIP codes

Starting ZIP:
 Ending ZIP:

Zip Code Distance Calculator

Use this form to calculate the distance between two U.S. ZIP codes.

Starting Location Ending Location
 ZIP Code: ZIP Code:

REDLANDS, CA 92373 Longitude: -117.172 Latitude: 34.849 Database Lookup: ZIP Code 92373	SAN DIEGO, CA 92108 Longitude: -117.137 Latitude: 32.774 Database Lookup: ZIP Code 92108
The distance between REDLANDS and SAN DIEGO is 80.10 miles or 141.79 kilometers .	

Figure 3. Coordinates and distance estimates from different sources differ

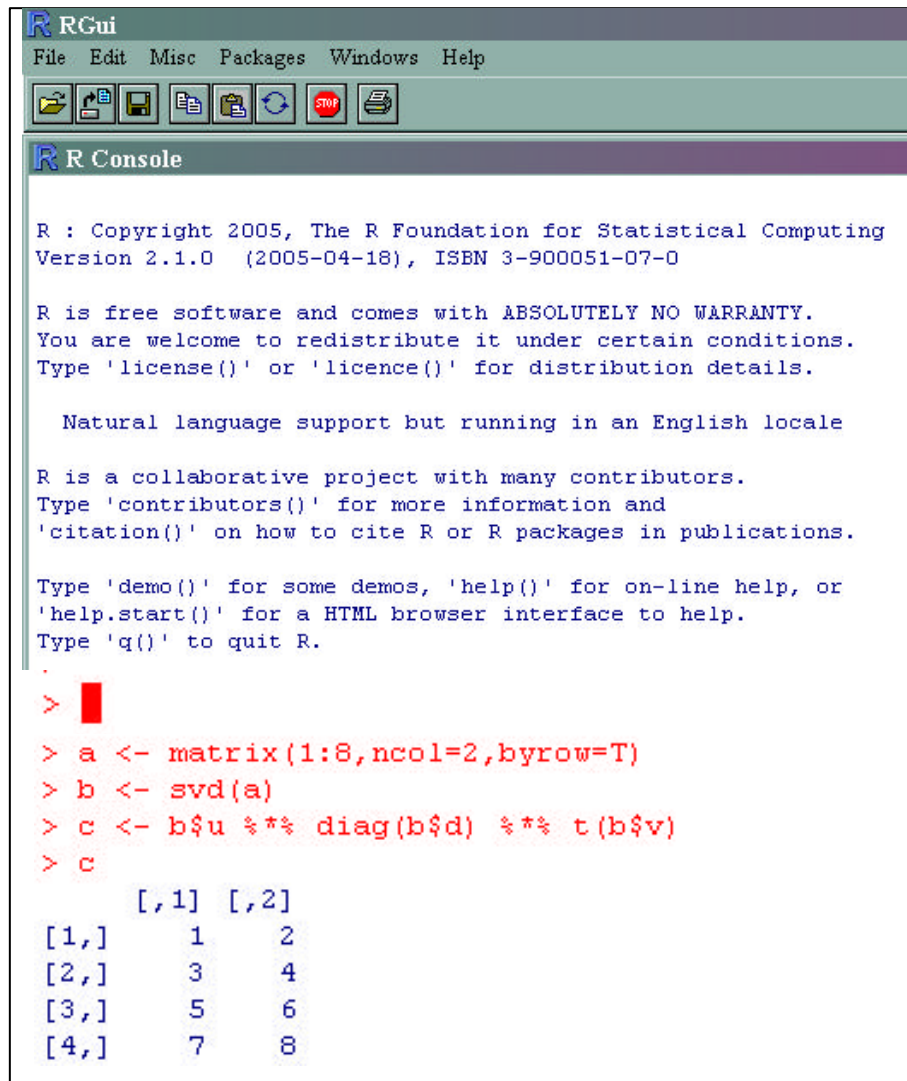
Name	State Code	ZIP Code	(North) Latitude	(West) Longitude
Redlands	CA	<i>RD</i> 92373	34.0057	117.1506
Burbank	CA	<i>BK</i> 91502	34.1771	118.3096
Torrance	CA	<i>TO</i> 90502	33.8359	118.2928
Cucamonga	CA	<i>RC</i> 91730	34.1001	117.5826
Santa Ana	CA	<i>SA</i> 92707	33.7155	117.8711
Riverside	CA	<i>RV</i> 92506	33.9325	117.3533
San Diego	CA	<i>SD</i> 92108	32.772	117.1465
Temecula	CA	<i>TA</i> 92590	33.4798	117.2258

Table 2. Coordinates for 8 zip codes in Southern California

5.3. The R Statistical Computing Environment

Analysis can proceed using many spreadsheet or statistical programs. However, the preference is to use a single system, and to select an adaptable and open-source environment, so that procedures can be modified as needed and shared among interested users. The R statistical computing environment provides one such system. It has several additional advantages, including a wide array of available packages that perform functions ranging from interfaces (to Oracle databases, GRASS geographic information systems, and ESRI shapefiles), to map objects and other visualization tools, to computational procedures (including Mantel testing, Procrustean analysis, etc.).

R is a descendent of the S language developed at Bell Labs (R = S-1 = “Almost S”), as is its commercial counterpart, S-Plus. Figure 4 shows the graphic user interface to R, and briefly demonstrates the ease of interactive statistical computing. The base system and supporting packages are available for download at <http://www.r-project.org>.



```
R RGui
File Edit Misc Packages Windows Help

R Console

R : Copyright 2005, The R Foundation for Statistical Computing
Version 2.1.0 (2005-04-18), ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.
.
> 
> a <- matrix(1:8,ncol=2,byrow=T)
> b <- svd(a)
> c <- b$u %*% diag(b$d) %*% t(b$v)
> c
      [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
[4,]    7    8
```

Figure 4. A singular value decomposition and matrix reconstruction in R

5.4. Map Constructions and Comparisons

Given the coordinates in Table 2, the plot function in R faithfully produces in Figure 5 a crude but accurate map of the eight Southern California locations.

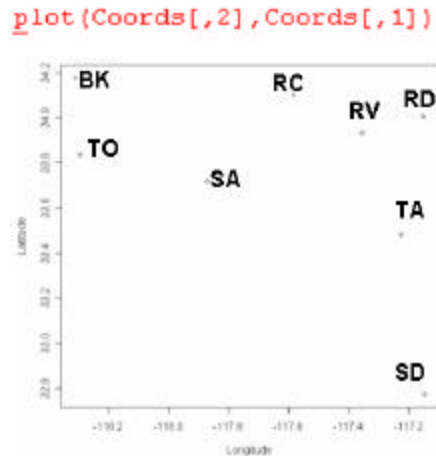


Figure 5. From coordinates to a map

Virtually the same display can be generated by plotting and rotating the first two dimensions of the multidimensional scaling solution to the matrix of estimated distances between locations. A gap analysis of the difference between the two displays is provided in Figure 6. The virtual reproduction of one map by the other is confirmed by the Mantel test statistic of 0.96, and by the correlation estimate of 0.995 following Procrustean analysis. One thousand permutations confirm the significance of these results at < 0.001 .

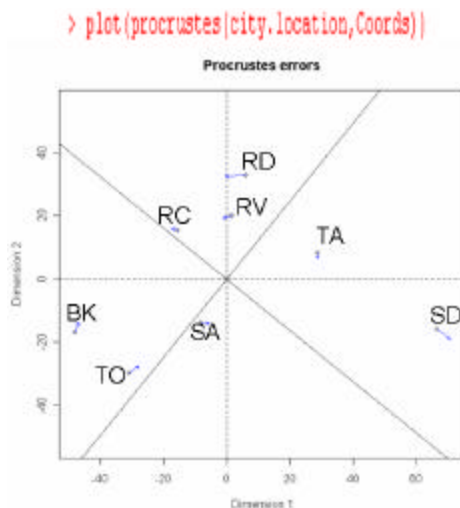


Figure 6. Gap display between coordinate map and MDS reconstruction based upon straight-line distance estimates

Driving distances between locations also are readily available, and are likely to differ from straight-line distances to the extent roads are influenced by terrain. The gap display in Figure 7 indicates slightly less degree of fit between maps than in Figure 6. Indeed, the Mantel test statistic drops to 0.94, and the bidimensional regression R-square is 0.98.

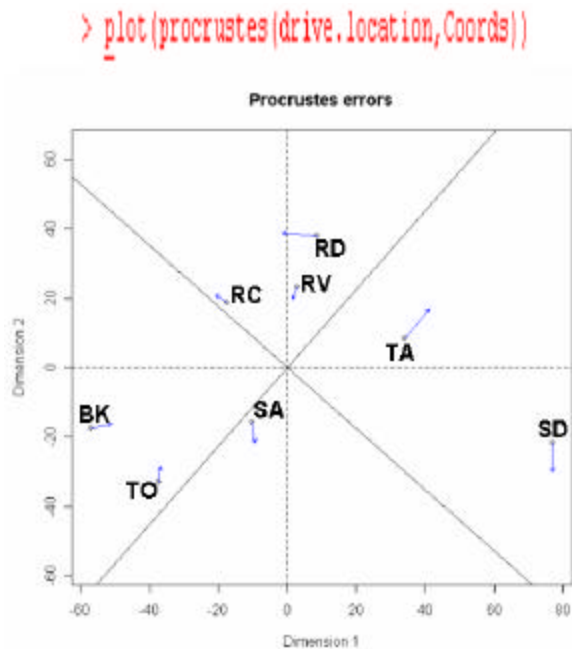


Figure 7. Gap display between coordinate map and MDS reconstruction based upon driving distance estimates

So far the maps have rather faithfully reproduced one another, whether generated from data at the object level or from dissimilarity measures among objects. The degree of fit is likely to drop substantially as non-geographic attributes of location are introduced to the mapping process. To demonstrate this, Table 3 introduces housing attributes for the eight locations, based upon recent information published on the Dataquest web site.

	SFRVol	SFRMed\$	SFR%Ch	CondoVol	CondoM\$	Condo%Ch	Pr/SQF
RD	38	\$450	32.4%	8	\$265	47.2%	\$245.38
BK	16	\$598	-8.1%	9	\$419	16.1%	\$423.73
TO	7	\$477	31.0%	16	\$311	25.9%	\$432.01
RC	75	\$396	20.7%	36	\$305	29.5%	\$253.59
SA	44	\$500	29.9%	28	\$301	18.7%	\$419.75
RV	69	\$384	23.9%	3	\$210	27.3%	\$239.77
SD	21	\$840	9.1%	25	\$477	15.7%	\$579.84
TA	68	\$403	13.4%	1	\$279	-15.5%	\$239.62

Table 3. Housing attributes for 8 zip codes in Southern California

Table 4 introduces information from the ESRI web site that portends to capture the community tapestry for each of the eight locations.

	Population	Household	White	Hispanic	Female	MedHHInc	HHIt50	Hlgt100
RD	35,011	2.36	79.3%	15.5%	53.00%	\$55,235	45.0%	25.6%
BK	11,723	2.43	56.5%	44.5%	51.70%	\$33,598	67.5%	9.4%
TO	17,314	3.08	38.3%	37.1%	51.70%	\$56,110	42.9%	21.8%
RC	57,069	3.11	55.4%	39.9%	49.80%	\$55,514	42.8%	17.6%
SA	63,568	4.76	44.0%	81.3%	48.40%	\$53,549	45.4%	16.2%
RV	45,600	2.85	72.7%	22.7%	51.70%	\$70,245	35.2%	32.6%
SD	13,704	1.71	73.0%	16.5%	51.20%	\$47,518	52.8%	13.0%
TA	3,532	3.06	66.6%	45.6%	48.90%	\$35,791	59.8%	25.2%

Table 4. Community tapestry for 8 zip codes in Southern California

Maps are generated by plotting the first two dimensions of the multidimensional scaling solution to the 7-dimension housing dataset and the 8-dimension community tapestry dataset, respectively. As anticipated, correspondence between the two maps is relatively weak, as shown in Figure 8. Both the bidimensional regression correlation and the Procrustean analysis correlation are estimated at 0.55, producing an R-square estimate of 0.30 that is not significantly different from the null hypothesis of independence (based upon 1,000 permutations). As noted in the display, however, the lack of fit may be explainable by a single influential observation, so that substituting another fit criterion could improve the evidence regarding association.

```
> plot(procrustes(h.location,t.location))
```

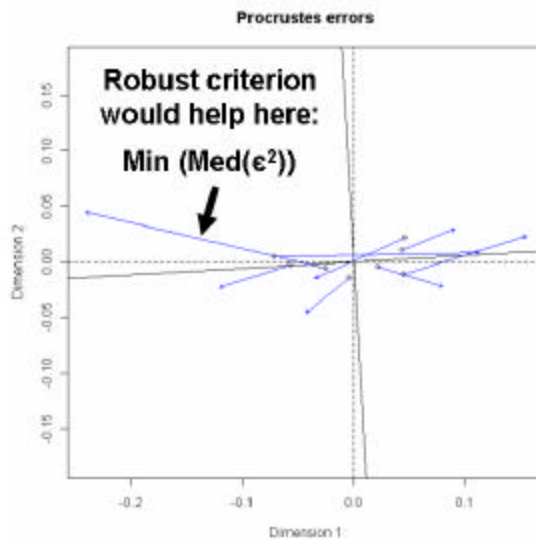


Figure 8. Gap display between maps created from housing attributes and from community tapestry attributes

5.5. Further Considerations

The analysis to this point has used some relatively less known tools and techniques to conclude that (1) maps drawn from actual geographic coordinates are not significantly different from those drawn based upon distance matrices, and (2) housing and community tapestry attributes do not map in a way that suggests association between the groups.

Traditional analysis reinforces and extends these conclusions. A linear regression of the housing price per square foot attribute on latitude, longitude, and their interaction produced an R-square of 0.73, with a p-value of 0.043. Statistical significance is not often supported for three predictors and eight observations! Adding the population density attribute and the percent Caucasian attribute to this model did not significantly improve the fit. Interestingly, substituting multidimensional scaling dimensions for the geo-reference predictors actually improves the fit, to an R-square of 0.83, with a p-value of 0.017. This is because the rotation of axes in multidimensional scaling reduces the magnitude of the error for one unusual observation (San Diego, from -107 to -36).

Finally, it should be noted that two-dimension maps often sacrifice information and accuracy to maintain simplicity. For example, the first two principle components capture 71.0% of variation in the straight-line distances and 70.3% of variation in the driving distances. The corresponding results for the first three principal components are 85.3% and 84.0%, respectively. Clearly, information is lost in dropping from three to two dimensions.

6. Conclusions and Extensions

Principal component analysis is useful in mapping nxp data matrices in two dimensions. Multidimensional scaling is useful in mapping pxp dissimilarity matrices in two dimensions. Procrustean analysis is useful in comparing maps, either in aggregate using permutation tests, or by examining errors among corresponding observations in the two maps. Robust procedures show promise for improving these strategies, and are easily implemented in a rich open-source statistical computing environment like R.

7. References

- Carlosona, A., Andrade, J.M., Kubista, M., and Prada, D., 1995. Procrustes Rotation as a way to compare different sampling seasons in soils. *Analytical Chemistry*, 67(14)2373-8.
- Davison, A.C. and Hinkley, D.V., 1997. *Bootstrap Methods and Their Application*.
- Dray, S., Chessel, D. and J. Thioulouse, 2003. Procrustean co-inertia analysis for the linking of multivariate datasets. *Ecoscience*, 10:110-119.
- Edwards, Jonathan and Oman, Paul, 2003. Dimension Reduction for Data Mapping, *R News*.
- Everitt, B.S. and Rabe-Hesketh, S., 1997. *The Analysis of Proximity Data*. Arnold Press.
- Fabrikant, S.I., 2001. Visualizing regions and scale in information spaces. *Proceedings, International Cartographic Conference*. 2522-9.

- Francel, K.E. and G.D. Schnell, 2002. Relationships among human disturbance, bird communities, and plant communities along the land-water interface of a large reservoir. *Environmental Monitoring and Assessment*, 73:67-93.
- Friedman, A. and Kohler, B., 2003. Bidimensional Regression: assessing the configural similarity and accuracy of cognitive maps and other two-dimensional data sets. *Psychological Methods*, 4:468-491.
- Gower, J.C., 1971. Statistical methods of comparing different multivariate analyses of the same data. Pages 138-149 in F.R. Hodson, D.G. Kendall & P. Tautu (eds) *Mathematics in the Archaeological and Historical Sciences*. Edinburgh University Press, Edinburgh.
- Gower, J.C., 1975. Generalized procrustes analysis. *Psychometrika*, 40:33-51.
- Jackson, D.A., 1993. Multivariate analysis of benthic invertebrate communities: the implication of choosing particular data standardizations, measures of association, and ordination methods. *Hydrobiologia*, 268:9-26.
- Jackson, D.A., 1995. PROTEST: A PROcrustean randomization TEST of community environment concordance. *Ecoscience*, 2:297-303.
- Jackson, D.A. and H.H. Harvey, 1993. Fish and benthic invertebrates: community concordance and community-environment relationships. *Canadian Journal of Fisheries and Aquatic Sciences*, 50:2641-2651.
- Jackson, D.A. and K.M. Somers, 1991. Putting things in order: The ups and downs of detrended correspondence analysis. *American Naturalist*, 137:704-712.
- King J.R., and D.A. Jackson, 1999. Variable selection in large environmental data sets using principal components analysis. *Environmetrics*, 10:67-77.
- Klingenberg C.P, and G.S. McIntyre, 1998. Geometric morphometrics of developmental instability: Analyzing patterns of fluctuating asymmetry with procrustes. *Evolution*, 52:1363-1375.
- Munoz, J. Felicisimo, AM, Cabezas, F. Burgaz, AR. and I. Martinez, 2004. Wind as a long-distance dispersal vehicle in the Southern Hemisphere. *Science* 304: 1144-1147.
- Oksanen, Jari, 2005. Multivariate Analysis of Ecological Communities in R. *Working Paper*.
- Olden, J.D., Jackson, D.A. and P. R. Peres-Neto, 2001. Spatial isolation and fish communities in drainage lakes. *Oecologia*, 127:572-585.
- Paavola, R. Muotka, T. Virtanen, R. Heino, J. and P. Kreivi, 2003. Are biological classifications of headwater streams concordant across multiple taxonomic groups? *Freshwater Biology*, 48: 1912-1923.

Paszkowski C.A., and W.M. Tonn, 2000. Community concordance between the fish and aquatic birds of lakes in northern Alberta, Canada: the relative importance of environmental and biotic factors. *Freshwater Biology*, 43: 421-437.

Peres-Neto, P.R. and D.A. Jackson, 2000. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia*, 129:169-178.

Soininen, J. Paavola, R., and T. Muotka, 2004. Benthic diatom communities in boreal streams: community structure in relation to environmental and spatial gradients. *Ecography*, 27:330-342.

Tobler, Waldo, 1994. Bidimensional Regression. *Geographical Analysis*, 26:3.

Waterman, S. and Gordon, D., 1984. A quantitative-comparative approach to analysis of distortion in mental maps. *Professional Geographer*, 36(3):326-337.

Wemelsfelder, F., Hunter, E.A., Mendl, M.T. and A.B. Lawrence, 2001. Assessing the 'whole animal': a Free-Choice-Profiling approach. *Animal Behaviour*, 62: 209-220.