Demonstration of Ranking Issues for Students: A Case Study

I. Elaine Allen
Babson College

Norean Radke Sharpe
Babson College

Journal of Statistics Education Volume 13, Number 3 (2005),
ww2.amstat.org/publications/jse/v13n3/sharpe.html

Copyright © 2005 by I. Elaine Allen and Norean Radke Sharpe, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Data Analysis; Demographics; Graphics; Rank methods.

Abstract

This article uses a case study of 2001 town and city data that we analyzed for Boston Magazine. We use this case study to demonstrate the challenges of creating a valid ranking structure. The data consist of three composite indices for 147 individual townships in the Boston metropolitan area representing measures of public safety; the environment; and health. We report the data and the basic ranking procedure used in the magazine article, as well as a discussion of alternative ranking procedures. In particular, we demonstrate the impact of additional adjustment for the size of population, even when per capita data are used. This case study presents an opportunity for discussion of fundamental data analysis concepts in all levels of statistics courses.

1. Introduction

Rankings have been used to compare nearly all variables that can be quantified in the interest of demonstrating differences for the public. In the area of education, it has become popular to rank business programs (Business Week 2002); other graduate programs (U.S. News & World Report 2003); and even science programs on the PhD production rate of women (Sharpe and Fuller 1995). In sports, the issue of ranking is dominant, with widespread financial implications for many involved in the industry. As a result, the issue of ranking in both amateur and professional sports has received more attention than other areas in the statistical literature (see, for example, Harville and Smith 1994; Groeneveld 1990; and Naik and Khattree 1996).

Ranking has also been a popular tool to compare towns and cities on the basis of demographic
dimensions. Often these dimensions are indices created by the analysts in an attempt to quantify a demographic concept. While the creation of indices is useful in obtaining a ranking, it is important to examine the variables used to create the index, as well as how the individual indices relate to the overall ranking. Nissan (1994) developed a new composite index for educational attainment that could be used to rank metropolitan areas. The advantage of this index was that it created data on a continuous scale, thereby allowing more options for statistical analysis.

In addition to the creation of indices, rankings are dependent on the raw data used to create the indices – particularly if these are survey, or perceptual, data. A recent report compared the WHO rankings of national health-care systems for industrialized countries with the rankings of the perceptions of users of the same health-care systems (Blendon, Kim and Benson 2001). This article demonstrated that the rankings change dramatically, depending on whether the perceptions of the provider, or the consumers, are considered in the rankings. The point was clearly made that multiple methods should be used in important rankings that are going to be used to determine public policy and distribution of financial resources.

One of the most controversial reported rankings was probably the one reported in the Places Rated Almanac (Boyer and Savageau 1985). These rankings were immediately critiqued for their lack of consumer weights placed on the demographic dimensions used as indices to create the overall ranking (Pierce 1985). In fact, in 1986 the Section on Statistical Graphics of the American Statistical Association (ASA) invited its members to examine data from the Almanac to compare ranking methods and outcomes. The data for this project, which consisted of nine composite variables for 329 metropolitan areas of the United States, was analyzed for alternative ranking methods and graphical approaches for presentation. One group of researchers investigated the validity of the components of the indices; distributions of the indices; and bivariate and multivariate relationships among the indices (Becker, Denby, McGill and Wilks 1987).

Because the presence of linear relationships does not necessarily indicate how users of the rankings weight the relative worth of indices, these authors turned to the concept of dominance (traditionally used in the theory of decision analysis) for assistance. (Becker et al., 1987) defined a city as dominating another city if each of the standardized indices for one city was greater (i.e., better) than the other city. This concept of dominance provided an interesting graphical (and geographical) opportunity for identifying those cities that dominate other cities. However, there was no direct relationship between ranking and dominance, although the dominators tended to be ranked in the top half and the dominated tended to be ranked in the lower half of all geographical areas (Becker et al., 1987). Finally, Becker and his co-workers compared several standard ranking methods for the demographic data to investigate differences from those ranks published in the Places Rated Almanac. This article demonstrated, in particular, the importance of considering alternative ranking methods; the importance of considering the impact of population on rankings; and the general importance of investigating published rankings more deeply using proven statistical techniques.

2. Case Study

2.1 Motivation and database for the case study

In February 2003 the authors were contacted by the Editor of Boston Magazine regarding appropriate methodology for ranking towns and cities surrounding Boston for a story on ‘healthiest towns.’ Data on towns and cities were obtained from public databases. The database sources include: U.S. Census, Executive Office of Environmental Affairs, Massachusetts Department of Revenue, Massachusetts Department of Education, FBI, U.S. Environmental Protection Agency, Massachusetts Municipal
Association, Massachusetts Beverage Business, Massachusetts Department of Public Health, Registry of Motor Vehicles and the Banker & Tradesman. The database for the Boston Magazine study was compiled by the magazine’s researchers and examined and error-checked by the authors, with a final validity check and audit by the magazine prior to analysis.

The overall database included demographic variables: town size in square miles, population density, school standardized test scores, tax rate, educational cost per pupil, median home price in 2002 and percent change in home price since 2001. Variables included in the ranking of ‘healthiest town’ were: violent crimes, public safety spending per capita, motor vehicle deaths and structure fires per capita, air pollution sources, number of contaminated sites, radon potential, percent of open space, different cancer rates, HIV/AIDS per 100,000 people, and sexually transmitted disease rate. All cancer rates were reported as standard incident ratios (SIR). The Standard Incidence Ratio (SIR) is calculated as the observed number of deaths for a particular cancer divided by the age, race, and gender adjusted death rate for the state of Massachusetts (the standard population) times 100. See www.mass.gov/dph/bhsre/mcr/00/supplement/citytown1996_2000.pdf for complete information. A value of 100 would indicate that a town's cancer rate was identical to the state rate. While there were more variables provided than those listed here, not all variables were included in the construction of the ranks, primarily due to missing values, inequity in reporting, or inappropriate application.

2.2 Construction of the ranks

In order to form rankings, the variables were divided into three categories, or indices: public safety, health, and environment. Some variables provided to us were not included because their relationship to the town’s occupants could not be clearly established, such as persons per doctor, fast food restaurants, liquor licenses, and number of health clubs/gyms. In addition, a few variables were excluded because of an overwhelming amount of missing data for towns included in the analysis. Missing data occurred for some of the smaller towns, because of no regulatory requirement for reporting (e.g., suicide rates) or very small numbers that are truncated to a categorical level (e.g., “< 5”). In one instance there was a small amount of missing data for a variable so the missing data for a town was replaced with the mean for that variable. Data were standardized by a town’s population size when not reported per capita or for a town’s size (for pollution measures) to get variables standardized per square mile. In addition, the standardized cancer ratios were combined into one overall SIR by adding them together to obtain an overall cancer incidence rate, rather than using individual cancer rates. Table 1 displays the variables included in the rankings.

Table 1. Variables included in Town rankings for the three categories.

**Public Safety:**
Violent crimes per capita in 2001
Structure fires per capita in 2001
Motor vehicle deaths per capita in 2001
Public spending per capita in 2001

**Health:**
Incidence per 100,000 of HIV in 2001
Incidence per 100,000 of Sexually Transmitted Diseases in 2001
SIR of Cancer (Leukemia, Breast, Lung, Bronchus, Prostate, Colon, and Rectal) in 2001
After variables were selected within categories, all variables were standardized to create variables with the same metric, and the resulting standardized variables were averaged to create a mean standardized score per category (public safety, health, and environment). Radon level, a scalar, was recoded as -1 (low), 0 (medium) and 1 (high) prior to averaging. The towns were ranked within the categories of public safety, health, and environment. An overall rank was calculated by averaging the mean standardized scores for the three categories. This method was used as it preserves the magnitude of the differences between the variables in the individual and overall indices. No weighting was applied to any of the categories or variables within the categories, nor were the towns weighted by size. For standardized values representing ‘unsafe’ or ‘unhealthy’ conditions, the greater the z-score, the less healthy; for ‘safe’ or ‘healthy’ conditions, the z-scores were inverted (the more negative the z-score, the healthier the town). Thus, the town ranked first (the best) in each category had the most negative z-score. The robustness of using this method is discussed in the next section.

3. Ranking Methods

3.1 Mean-Standardized Rank Method

The first method we used was the traditional mean-standardized rank method, where we standardized each of the indices (mean of zero, standard deviation of one), so that the differences between each of the indices are re-scaled to be consistent across each index. We then averaged the standardized indices and ranked the cities according to the mean (Becker et al., 1987 referred to this approach as the rank-scaled method). For example, to obtain the ranking of the health index, we first obtained the z-scores for each of the health variables (incidence of HIV, STD, and Cancer) for each town, then averaged these z-scores and sorted the averages giving the most negative z-score a rank of 1. (The towns with the highest incidence of each of these diseases would end up with the most positive z-scores and be ranked near the bottom.) This standardized method is preferred over the more accessible rank-mean approach (which ranks each of the variables without standardizing, then averages the ranks to obtain an overall rank for the index), because the rank-mean method does not maintain the magnitude of the differences between the original measurements for the variables.

Using this method for the Boston Magazine data set, all three indices were weighted equally. See www.bostonmagazine.com/documents/publications/april03_gate2.pdf for the rankings. Although a prior study has revealed differences in the importance of each demographic dimension by survey respondents in a stratified random sample (Pierce, 1985), transferring the weights to our data set provided additional decisions. First, the emphasis of our data set and ultimate ranking was on factors that affected a person’s well-being (measures of safety, environmental factors, and the state of health of the town’s population). The Pierce study included the additional dimensions of the state of the economy, climate, housing, education, recreation, transportation, and the arts in the survey. Second, the dimension of ‘health’ was defined differently from ours; Pierce used a measure of health care provided in the metropolitan area, as opposed to incidence of disease. Finally, the dimension of health and environment were combined in the Pierce study. Thus without additional guidance from an updated survey on public perception and

Environment:
Percent open space in 2001
Air pollution sources per square mile in 2001
Number of contaminated sites per square mile in 2001
Presence of Radon (low, medium, high)
importance of these demographic dimensions, we decided to equally weight our three indices.

Although we had used per capita data for most of the variable components of the indices, we noticed that a relationship seemed to exist between the population of the city or town, and the placement in the overall ranking: the larger cities seemed to be at the bottom of the ranking, while the less populated towns seemed to be at the top of the ranking. We suspected that, although our variables adjusted for population (e.g., number of cancer cases per 1000 residents), a “penalty” may be paid by larger towns for their greater density of population. For example, a greater density of residents suggests side effects of over-crowding and financial constraints that cannot be accounted for simply by examining incidence, as opposed to prevalence.

To investigate this suspected relationship between population and the ranking, we graphed the ranking versus the log of the population of each town (see Figure 1) and computed the correlation. Although, Figure 1 does not seem to indicate a particularly strong relationship between the log of the population and rank (with the possible exception of Worcester and Boston, the two largest cities in the data set) the correlation between these two measures is 0.49 (p < 0.01). (Omitting Worcester and Boston only drops the correlation to 0.46, p < 0.01).

We then investigated the correlation between the log of the population and each of the individual indices, since differences in the relationships are likely to exist across each index. The graphs of the standardized scores for each index versus the log of the population of each town clearly show that the strongest relationship exists between the health measure and the number of residents in the town (see Figure 2, Figure 3, and Figure 4). Although the largest city in the data set (Boston) seems to be an unusual observation in Figure 2 and Figure 3, it seems to support this relationship between the composite measure for health and population – again, although incidence rates were used in the computation of this index. The correlations between the log of the population and the standardized scores for safety, environment, and health were -0.04, 0.15, and 0.75 (p<.01), respectively. (None of the indices were significantly correlated with each other.)
Figure 2. Standardized scores for the safety index versus log of population.

Figure 3. Standardized scores for the environmental index versus log of population.
The question then arises, how should this relationship between population and the resulting rank-order of each of the indices be handled? One approach that is fairly common is to rank each city or town within a set of towns of similar size based on census definitions and villages, towns, and cities. (This is analogous to the approach taken by *U.S. News & World Report* when they rank institutions in each size and selectivity category.) The advantage to this approach is that it is easily understood by readers of the popular press. The disadvantage of this approach is that it does not produce a definitive ranking – nor does it contribute to the understanding of how to adjust for the size of the population, even after per capita data have been used.

### 3.2 Population-Adjusted Rank Method

One approach to adjusting for population is to run a regression between the log of the population and each of the indices and use the standardized residuals of each regression as the standardized measures for each index: safety, environment, and health (see Becker et al., 1987). These new standardized scores are then averaged to obtain an overall ranking. This method essentially uses the difference between the suspected (predicted) relationship between population and the index as a measure of how far below (or above) the town is measuring up on this demographic dimension from what is expected, given the town’s population. The result is that each town is provided the opportunity to rise (or fall) in the rankings if they exceed (or fall below) expectations for the size of the town.

*Figure 5* shows the relationship between the mean-standardized rank method and the population-adjusted method. While the relationship is positive and significant ($r = 0.83$, $p < 0.01$), it is clear that for many towns, the difference between the population-adjusted rank and the mean-standardized rank can be dramatically different. *Figure 6* shows the relationship between the difference between the two ranks (population-adjusted rank minus mean-adjusted rank) and the average of the two ranks. A positive difference on the Y axis indicates that the town “dropped” in rank, while a negative difference indicates that the town “rose” in rank. It is clear that those towns in the “middle of the pack” are affected the most. *Table 2* shows the top fifteen towns for the two different ranking methods and *Table 3* shows the bottom fifteen towns for the two different methods.
Figure 5. Relationship between the mean of standardized ranks and population-adjusted ranks.

Figure 6. Difference between population-adjusted ranks and mean-standardized ranks versus the average of the two ranks.

Table 2. Top fifteen towns using different ranking methods

| Rank | Mean-Standardized Rank Town | Population-Adjusted Rank Town |
|------|-----------------------------|------------------------------|
|      |                             |                              |
Table 3. Bottom fifteen towns using different ranking methods

| Rank | Mean-Standardized Rank Town | Population-Adjusted Rank Town |
|------|-------------------------------|-------------------------------|
| 147  | Chelsea                       | Chelsea                       |
| 146  | Lawrence                      | W. Bridgewater                |
| 145  | Worcester                     | Avon                          |
| 144  | Boston                        | Lawrence                      |
| 143  | Lowell                        | W. Newbury                    |
| 142  | Brockton                      | Topsfield                     |
| 141  | Lynn                          | Worcester                     |
| 140  | New Bedford                   | Wilmington                    |
| 139  | W. Bridgewater                | Brockton                      |
Note, that there is a fair amount of agreement between the two ranking methods in the top and bottom towns. A total of eight towns place in the top fifteen for both ranking methods, whereas a total of nine towns place in the bottom fifteen for both ranking methods. Is there a relationship between the difference in the rankings and the population of the town? Figure 7 shows a graph of the difference in the rankings versus the log of the population. Note that a positive difference implies that the town dropped in the population-adjusted ranking and a negative difference implies that the town rose in the population-adjusted ranking. The outlying towns represent towns that did not rise (or fall) as much as expected. For example, the towns of Boston and Worcester did not rise in the population-adjusted ranking as much as was predicted by the relationship. The towns of W. Newbury and Avon did not drop in the population-adjusted ranking as much as predicted. This was most likely a function of the constraint on how far they could drop in the rank, because both of these towns ended up in the bottom fifteen according to population-adjusted rank.

|   | Cambridge | Lynn |
|---|-----------|------|
| 138 | Avon      | Rowley |
| 137 | Everett   | Lowell |
| 136 | Somerville | Rockland |
| 135 | Wilmington | Everett |
| 134 | Haverhill | Littleton |

Figure 7. Relationship between the difference between ranks and log of population.

Is there a method to report which towns are over performing, and which towns are under performing
(w.r.t. safety, environment, and health), based on the size of their population? If we run a regression for the two variables in Figure 7, we obtain the predicted amount of change in rank for each town, given its population ($R^2 = 67\%$). Note, that those towns above the regression line are under-performers, and those towns below the regression line are over-performers, given the size of their population. More specifically, in Figure 8 all towns above the regression line and the horizontal line $Y = 0$, have dropped in rank more than expected after adjusting for the size of population; these towns are under-performers. The towns above the line $Y = 0$, but below the regression line are over-performers, because they did not drop in rank as much as expected. All towns with a negative residual (below the regression line) are over-performers.

![Figure 8. Display of Under Performing and Over Performing Towns.](image)

Table 4 shows the fifteen towns with the most negative residuals, $Y_i - \hat{Y}_i$, or the greatest over-performers for their size and the most positive residuals (greatest under-performers for their size). The actual change in rank from the traditional mean-standardized approach and the population-adjusted method are shown in Table 4; a positive change means that the town actually moved closer to the bottom of the ranking (i.e., closer to the largest rank of 147) and a negative change means that the town moved closer to the top of the rank (i.e., closer to the top rank of 1). Note, that there are both under-performing towns that are large, as well as small. Clearly, the actual change in rank is negative for the larger towns – they did not rise as much as expected, and the actual change in rank is positive for the smaller towns – they fell in rank more than expected. (The actual changes in rank do not appear in descending or ascending order, because these towns were chosen based on the value of their residual in the regression in Figure 8, and are in descending order by residual.)
Table 4. The fifteen best over-performing towns and worst under-performing towns.

| Actual Change in Rank (positive means they moved closer to the bottom of the rankings) | Town/City | Actual Change in Rank (negative means they moved closer to the top of the rankings) | Town/City |
|---|---|---|---|
| -27 | Boston | -44 | Billerica |
| -4 | Worcester | 8 | Avon |
| -3 | Brockton | -60 | Newton |
| -3 | Lynn | -51 | Waltham |
| -7 | Lowell | -40 | Beverly |
| -8 | New Bedford | 13 | W. Newbury |
| -2 | Lawrence | -33 | Lexington |
| 54 | Stow | 7 | Dover |
| 62 | Nahant | -36 | Marlborough |
| 46 | Newbury | -40 | Methuen |
| 41 | Hamilton | -31 | Andover |
| 39 | Southborough | -38 | Arlington |
| 56 | Sherborn | -47 | Framingham |
| 49 | Manchester-by-the-Sea | 7 | W. Bridgewater |
| 59 | Essex | -25 | N. Andover |

4. Teaching Implications

Rankings are a common tool used in the comparison of towns, institutions, teams, individuals, countries – and almost every entity that can be quantified in some way. As a result, they are frequently used to make financial, policy, and political decisions. It is well known that most rankings can be improved – either by investigating the ranking method, the composite indices, or the raw data used to create the
indices. Thus it is important that our students learn about ranking issues in their statistics courses.

Since our students are future consumers, and perhaps creators, of surveys and rankings, it is essential that students are educated in the dangers of rankings. It is important that our students understand that there exist multiple ranking methods – each of which will yield different (although perhaps similar) results. An advantage to using both size and per capita measures in an analysis is that the results indicate, not only the advantage of living in a specific town, but also whether you are ranked as an over performer or under performer. Perhaps by including a discussion of ranking methods in our statistics courses, we can enhance the ethical creation and consumption of rankings, in general.

Many types of analyses can be done with this dataset of health-related variables in an introductory statistics course, including descriptive statistics, variable creation for size of town and correlation, and simple or multivariate regression examining the relationship between housing, health, safety and environmental variables. However, it is in more advanced Applied Multivariate Statistics classes that the data can be fully examined. In this class the students have the opportunity to create factors from the individual variables, examine clusters of towns by different health factors, and use these results to predict and rank the towns. Using the radon variable, and creating categorical variables based on size, population, and density, students can use ANOVA to examine significant differences between groups. A typical exercise for students would be the creation of a model predicting housing prices to find the towns where buyers are getting the most value (in terms of health, safety, and environmental factors) and towns that are overpriced.

5. Conclusion

Additional research is needed in this often-neglected area to enhance existing tools and develop new methods. It is clear that we need to continue to 1) examine the composition of indices to assess validity; 2) compare multiple methods of rankings to assess reliability; and 3) adjust the rankings by population (if appropriate) to improve their usability and applicability.

In this article, a case study of a recent data set collected by Boston Magazine was used to demonstrate the impact of an additional population adjustment on the rankings. Although variables were used in the composition of the indices using per capita data and standardized prior to ranking, other effects of an increased population exist, which cannot be quantified. It is not surprising that the size of the population seemed to have greatest impact on the health index, even though incidence (not prevalence) of disease was used. In addition, since each index was weighted equally (without updated consumer perception results), it is important to investigate the impact of each index, since each consumer, may in fact, weight each index differently.

As with other statistical and graphical techniques, the outcome and presentation of the ranking analysis is highly dependent on the quality of the method used to obtain the rankings. Hopefully, with the use of case studies, such as the one discussed in this article, we can improve the general comprehension of rankings and motivate additional research in this area.

----------------------------------------

Acknowledgments

We thank the referees for their careful and thoughtful reading of our manuscript.
References

Becker, R.A., Denby, L., McGill, R., and Wilks, A.R. (1987), “Analysis of data from the Places Rated Almanac,” The American Statistician, 41(3), 169-186.

Blendon, R.J., Kim, M., and Benson, J.M. (2001), “The Public versus the world health organization on health system performance,” Health Affairs, 20(3), 10-20.

Boyer, R., and Savageau, D. (1985), Places Rated Almanac (rev. ed.), Chicago: Rand McNally.

Business Week (Oct. 21, 2002), “The Best B-Schools,” 84-110.

Groeneveld, R. A. (1990), “Ranking teams in a league with two divisions of t teams,” The American Statistician, 44(4), 277-281.

Harville, D. A., and Smith, M.H. (1994), “The home-court advantage: how large is it, and does it vary from team to team?” The American Statistician, 48(1), 22-28.

Naik, D. N., and Khattree, R. (1996), “Revisiting Olympic track records: Some practical considerations in the principal component analysis,” The American Statistician, 50(2), 140-144.

Nissan, E. (1994), “A composite index for statistical inference for ranking metropolitan areas,” Growth and Change, 25, 411-426.

Pierce, R. M. (1985), “Rating America’s metropolitan areas,” American Demographics, 7, 20-25.

Sharpe, N. R., and Fuller, C.H. (1995), “Baccalaureate origins of women physical science doctorates: relationship to institutional gender and science discipline,” Journal of Women and Minorities in Science and Engineering, 2(1 & 2), 1-15.

U.S. News & World Report (April 14, 2003), “America’s Best Graduate Schools,” 52-76.

I. Elaine Allen
Division of Mathematics & Statistics
Babson College
Babson Park, MA
U.S.A.
Allenie@babson.edu

Norean Radke Sharpe
Division of Mathematics & Statistics
Babson College
Babson Park, MA
U.S.A.
Sharpen@babson.edu