2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery

A Method for Implementing a Statistically Significant Number of Data Classes in the Jenks Algorithm

Dr. Matthew A. North
Washington & Jefferson College
mnorth@washjeff.edu

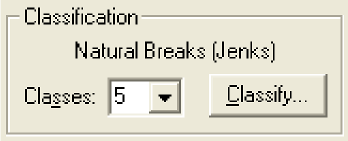
Abstract

The Jenks Natural Breaks algorithm is a standard method for dividing a dataset into a certain number of homogenous classes. The algorithm is commonly used in Geographic Information Systems (GIS) applications. One major drawback to the use of Jenks in this context is that the number of desired classes must be indicated before the algorithm is applied to the dataset. Without a mechanism for determining the appropriate number of classes for a given dataset, the results of Jenks classification may be inaccurate, or worse, arbitrary. This paper proposes a method for determining, through iterative tests of statistical significance, the appropriate number of classes for a data set of any given number of observations. Pseudo-code for the method is provided.

1. Introduction

The Jenks Natural Breaks algorithm (hereafter, "Jenks") was introduced in 1977 as a method for "optimal data classification" [2]. The design of the algorithm is based primarily upon Fischer's "Exact Optimization" method, developed in 1958 [7]. Jenks was specifically developed for use in analysis of geographic data, and has emerged as a standard geographic classification algorithm, as evidenced by its selection as the default classification method in the industry-leading software package ArcGIS from Environmental Research Systems Institute (ESRI). Numerous tests for validity over time have ensured the reliability of the algorithm, resulting is its use in numerous published works, primarily in the field of GIS [1, 3].

One drawback to the use of Jenks in modern GIS systems is that the number of classes desired in a result set must be provided before the algorithm is applied to the dataset [3, 6]. Figure 1 illustrates this requirement.



Natural Breaks (Jenks)

Classes: 5

Figure 1. Jenks number of classes indicator in ESRI's ArcGIS ArcMap software.

This predefinition of the desired number of classes prior to Jenks analysis creates a bit of a paradox [8]. The objective of Jenks is to create homogeneous classification groups, but the number of groups resulting from Jenks analysis must be predefined. Users who are familiar with common benchmarks, divisions, or strata in their data may be able to provide a reasonable number of categories for Jenks analysis, but what if the user is unfamiliar with the nature of existing divisions within their dataset [5]? Further, what if the assumptions made by an individual familiar with a dataset are insufficient to select a truly meaningful number of classes before applying the Jenks algorithm [4]?

2. Statement of the Problem

Because of the need to indicate a target number of classes prior to the Jenks algorithm being applied to a dataset, a method for reliably

978-0-7695-3735-1/09 $25.00 © 2009 IEEE
DOI 10.1109/FSKD.2009.319

determining a statistically accurate and appropriate number of classes is needed.

3. Proposed Solution

When attempting to determine whether or not two groups of numbers are statistically different from one another, a simple *t*-test has long been established as a statistically valid and reliable way of determining whether or not homogeneity exists [6]. The method proposed here for identifying a statistically significant number of classes for use with the Jenks algorithm relies upon the basic *t*-test, and the use of number arrays which are dynamically constructed, compared to one another, and then re-constructed.

The first step in this method is to analyze the variable in the data set to be used for categorization. The variable will be reduced to distinct values and rank ordered from highest to lowest. The purpose of using distinct values is to avoid skewing of the *t*-test if multiple instances of the same value occur in the dataset, while other values may occur only once. With the variable values reduced and sorted, the first value is selected and compared to the second value via *t*-test, and a p value is generated. Depending on the desired alpha level for the *t*-test, that value is either considered homogeneous, and thus part of the first category, or considered heterogeneous and thus part of a new category. Once a value from the list is found to be statistically different, that value is used as the comparison value for the next category, and the subsequent values are compared until significant difference is found again. A counter variable is set to start at zero, and increment by one for each new category which is identified through the cyclical process until all values in the selected variable have been evaluated. The value from the counter is returned to the user as the statistically significant number of classes to be used for Jenks classification in their data set.

3.1 Pseudo-code

Consider a data set containing variable *pop*, which is the population of each city in the data set. The data set contains n observations, or in other words, n number of cities and their respective populations.

```
1  Let array $y$ contain the values of variable
   pop.
2  Select distinct values from $y$ into array $z$
3  Sort $z$ in numeric order.
4  Create variable $c$ with data type integer
5  Set $c = 0$
6  Create variable $m$ with data type double
7  Set $m = .05$  //this is the alpha level for the
                        t-test
8  Create $j$ as an array
9  Create $k$ as an array
10 Let $j$ contain the first observation in $z$
11 Let $k$ contain the second observation in $z$
12 Compare $j$ to $k$ using t-test
13 Create variable $p$ with data type double
14 Let $p$ equal the $P$ statistic from the t-test in
   line 12
15 If $p > m$ then let $j$ contain two instances of
   the first value in $z$, and let $k$ contain the
   second and third values in $z$
16 Loop: Repeat lines 12 through 15, adding
   the next ordinal value in $z$ to $k$
17 When $p < m$, then $j$ and $k$ are significantly
   different from one another; Let $c = c+1$
18 Let $j$ contain the value of the final
   observation in $j$
19 Let $k$ contain min($z$) where min($z$) > $j$
20 Loop: Repeat lines 12 through 15, adding
   the next ordinal value in $z$ to $k$
21 When $p < m$, then $j$ and $k$ are significantly
   different from one another; Let $c = c+1$
22 When all observations of pop in $z$ are
   evaluated, exit loop
23 Print "There are " + $c$ + "statistically
   significant categories in this data set."
```

4. Discussion

This method for implementing a statistically significant number of classes in the Jenks algorithm is a simple but effective preprocessing step which can be taken to ensure that the resultant number of classes created during Jenks classification is meaningful. It does not negate the need for the Jenks algorithm, since the method does not suggest *where* the divisions between classes should occur, only *how many* there should be. Thus, it works in tandem with the Jenks algorithm by informing the user what value should be provided before the natural breaks are created.

This proposed method is both a work in progress and a theoretical solution to the problem posed at the beginning of this paper, so the logical next step is to build a system within which to test the method. Since many data sets in geographic information systems nowadays are stored in geodatabases, one potential option for implementing and testing the proposed method would be to combine structured query language queries with a scripting or procedural database language to create, populate and evaluate the necessary variable as outlined in the pseudo code. Although some scripting and programming can be conducted within the ArcGIS software, the method would be most easily tested external to the map software environment.

Assuming the method is proven to work effectively and results in a reliable method for identifying the significant number of classes in any given data set, it would be logical to then implement it into systems which use Jenks Natural Breaks, such as ArcGIS. This preprocessing mechanism could be fairly easily integrated into software packages which use Jenks, given that it need not interact with any other part of the software. It would simply need to access the file containing the data set, evaluate the variable to be categorized using this proposed method, and return the significant number of classes to the user.

5. Conclusions

This paper presents a suggested method for determining the statistically significant number of categories in any given data set. The method is useful given the need for the Jenks Natural Breaks algorithm to have a target number of classes provided before the algorithm is applied to the data set. While number of classes may be set manually based upon knowledge of the user related to the data set, this method provides a more structured approach and solution to the question of how many categories are appropriate when categorizing data. This method will be particularly useful to individuals who work in GIS environments, where the Jenks algorithm is widely used.

6. References

[1] C.A. Brewer and L. Pickle (2002). Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series. *Annals of the Association of American Geographers, 92*(4), pp. 662-681.

[2] G.F. Jenks (1977). Optimal data classification for
choropleth maps. Occasional paper No. 2. Lawrence, Kansas: University of Kansas, Department of Geography.

[3] M. Fariweather and K.G. Fairweather (1984). Choropleth Mapping: The Problems of Classification and Data Presentation. Resources in Education, 1984, pp. 13-27.

[4] P. Hurvitz (2004). Classifying Map Displays. *Spatial Technology, GIS, and Remote Sensing*. Seattle, Washington: University of Washington, College of Natural Resources.

[5] R.G. Mcgarvey, E.A. Lehtihet, E.D. Castillo and T. M. Cavalier (2001). On the Frequency and Location of Set Point Adjustments in Sequential Tolerance Control. *International Journal of Production Research, 39*(12), pp. 2659-2674.

[6] R.M. Smith (1986). Comparing Traditional Methods for Selecting Class Intervals on

Chloropleth Maps. Professional Geographer, 38(1), pp. 62-67.

[7] W.D. Fisher (1958) On Grouping for Maximum Homogeneity. *Journal of the American Statistical Association*, 53, pp. 789-798.

[8] Y. Martin and M. Church (2004). Numerical Modeling of Landscape Evolution: Geomorphological perspectives. *Progress in Physical Geography, 28*(3), pp. 317-339.