Choropleth maps considered harmful: scale and spatial data analysis

Martin Charlton
Chris Brunsdon

National Centre for Geocomputation
Maynooth University
Maynooth
Co Kildare
IRELAND
Outline

• Class intervals: algorithm archaeology

• Scale effects on multivariate relationships

• WTF
CLASS INTERVAL CHOICE
Can anyone explain...

• ... how the Jenks *natural breaks* classification works?

• It's a question I ask the Maynooth MSc in GIS and Remote Sensing students weekly until I get an answer.

• One of my professorial colleagues told me it was used because it gave the *best* classification...
Origins

• Early packages produced outputs using overprinted symbols in line printers
  – SYMAP (Harvard ~1970)
    • Equal interval
    • Quantile
    • User supplied
  – GIMMS (Edinburgh ~1973)
    • Equal interval
    • User supplied
• Arcane commands on punched cards
A = AIRPORT
1980s

• In the mid-1980s ArcINFO became available at three UK Universities
• Still only equal-interval, quantile, and user-supplied classes
• By 1988 GIMMS offered 12 different methods...
  – Equal interval, equal integer interval, curvilinear, geometric arithmetic, reciprocal, trigonometric, percentiles, standard deviation, quantile, nested means... And user supplied.
  – Coup de grace was the CHEST deal
1990s

- ArcINFO 6.0 (ESRI ~1990)
  - Equal interval
  - Quantile
  - User supplied
- ArcView 2.0 (ESRI ~1994)
  - Quantile
  - Equal Interval
  - Unique value
- ArcView 3.0 (ESRI ~1997)
  - Equal interval
  - Quantile
  - **Natural Breaks**
  - Equal Area
  - Standard deviation

**Natural Breaks**
This is the default classification method in ArcView. This method identifies breakpoints between classes using a statistical formula (Jenk's optimization). This method is rather complex, but basically the Jenk's method minimizes the sum of the variance within each of the classes. Natural Breaks finds groupings and patterns inherent in your data.
George F Jenks (1916-1996)

- PhD (Syracuse) 1947
- Joined Geography faculty at University of Kansas in 1949
- Cartography research dominated his output
- Reading Jenks' papers it is clear he wanted map-makers to think very carefully about the nature of the data that was to be symbolised, and the manner of that symbolisation.
Important papers


• Jenks GF, **1967**, The data model concept in statistical mapping, *International Yearbook of Cartography*, 7, 186-190


• Jenks GF, **1977**, *Optimal data classification for choropleth maps*, Occasional paper No 2, Dept of Geography, University of Kansas
FAQ: What is the Jenks optimization method?

Answer

Some Esri products use the Jenks optimization method to classify features using natural breaks in data values. The Jenks optimization method is also known as the goodness of variance fit (GVF). It is used to minimize the squared deviations of the class means. Optimization is achieved when the quantity GVF is maximized:

1. Calculate the sum of squared deviations between classes (SDBC).

\[ \text{GVF} = \text{SDBC} \]

2. Calculate the sum of squared deviations from the array mean (SDAM).

3. Subtract the SDBC from the SDAM (SDAM - SDBC). This equals the sum of the squared deviations from the class means (SDCM).

The method first specifies an arbitrary grouping of the numeric data. SDAM is a constant and does not change unless the data changes. The mean of each class is computed and the SDCM is calculated. Observations are then moved from one class to another in an effort to reduce the sum of SDCM and therefore increase the GVF statistic. This process continues until the GVF value can no longer be increased.
Wikipedia on Jenks

Method [edit]

The method requires an iterative process. That is, calculations must be repeated using different breaks in the dataset to determine which set of breaks has the smallest in-class variance. The process is started by dividing the ordered data into groups. Initial group divisions can be arbitrary. There are four steps that must be repeated:

1. Calculate the sum of squared deviations between classes (SDDC).
2. Calculate the sum of squared deviations from the array mean (SDAM).
3. Subtract the SDBC from the SDAM (SDAM-SDBC). This equals the sum of the squared deviations from the class means (SDCM).
4. After inspecting each of the SDBC, a decision is made to move one unit from the class with the largest SDBC toward the class with the lowest SDBC.

New class deviations are then calculated, and the process is repeated until the sum of the within class deviations reaches a minimal value.[1][5]

Alternatively, all break combinations may be examined, SDCM calculated for each combination, and the combination with the lowest SDCM selected. Since all break combinations are examined, this guarantees that the one with the lowest SDCM is found.

Finally, the GVF statistic (goodness of variance fit) is calculated. GVF is defined as \( (SDAM - SDCM) / SDAM \). GVF ranges from 0 (worst fit) to 1 (perfect fit).

Identical language and abbreviations (SDBC, SDAM, SDCM in steps 1...4) to the ESRI FAQ
SDAM, SDBC, TSS, WSS, WTF...

- According to ESRI Jenks' approach minimises the SDCM.

\[
SDBC = \sum_{k=1}^{K} (\bar{x}_k - \bar{x})^2 = BSS
\]

\[
SDAM = \sum_{i=1}^{N} (x_i - \bar{x})^2 = TSS
\]

\[
SDCM = SDAM - SDBC = \sum_{k=1}^{K} \sum_{j \in k}^{n_k} (x_j - \bar{x}_k)^2 = TSS - BSS = WSS
\]

\[
GVF = \frac{SDAM - SDCM}{SDAM} = \frac{TSS - WSS}{TSS} = 1 - \frac{WSS}{TSS}
\]

- When the WSS is small, the GVF approaches 1. Minimising the WSS is the same as maximising the GVF.
Further reading

• The ESRI FAQ on Jenks optimization and the Wikipedia entry sends us to:

Further Reading

However

• Before you run out to get a copy...

• It's pretty rare...
What does Jenks 1967 actually say?

• I tracked a copy down to a secondhand bookseller in a suburb of Köln.

• My copy had been bought for the library of the Institute for Cartography and Topography at the University of Bonn.

• Its condition suggests that there haven't been many readers dipping into its pages in the last 48 years...
INTERNATIONALES JAHRBUCH FÜR KARTOGRAPHIE

Herausgegeben von
Professor Dr. Erwin Raisz, Frankfurt am Main
in Verbindung mit
Dr. F. Arndt, Wiesbaden – J.C. Beutler, Zürich
Dr. W. Bremer, Stuttgart – J. C. Bühler, Zürich
Prof. Dr. W. Frank, München – Prof. W. Rössler, München
Prof. Dr. P. H. Schaefer, Hannover – Dr. P. H. Schaefer, Hannover
Prof. Dr. H. T. Schmalz, München – Prof. Dr. H. T. Schmalz, München
Prof. Dr. F. J. Schouten, Utrecht – Prof. Dr. F. J. Schouten, Utrecht
Prof. Dr. A. H. Schurmann, Mainz ([U.S.] und Ben)

International Cartographic Association
Internationaler Kartographen-Verband
Interantional Cartographic Association
Interantionaler Kartographen-Verband

— VII 1967 —

G. BERTRAM-ESSEN AT THE VISUAL ARTS CENTER, WASHINGTON, D.C.
The general procedure which we have to follow in preparing choropleth maps is very simple. When we are to be guided, as we are, by the principles governing the use of statistical maps, the need for exact conformity to visual and the nature of the data, we are compelled by the need to match characteristics of the choropleth maps with both the statistical behavior and the visual presentation of the data. Therefore, we have to present the data in such a way that visual techniques work best to get the maximum amount of information from the map. This is usually done by using qualitative colors to represent the different levels of the data, thereby allowing the reader to quickly identify the areas with higher values.

In this case, we are dealing with a choropleth map representing the distribution of population density. The map is divided into five classes, and each class is represented by a different color. The largest class, representing the highest population density, is shown in a dark color, while the smallest class, representing the lowest population density, is shown in a light color.

To understand how the data is represented, we can look at the legend provided on the map. The legend shows the range of values for each class, along with the corresponding color. By comparing the color of a particular region to the legend, we can estimate the population density of that region.

In conclusion, choropleth maps are a useful tool for visualizing the distribution of data across different areas. They allow us to quickly identify areas with higher values and understand the overall pattern of the data.

[Figure 1: Choropleth map of population density]

[Figure 2: Legend for the choropleth map]

[Figure 3: Close-up view of the map]

[Figure 4: Alternative representation of the data]

[Figure 5: Comparison of different map styles]
Example data

- Jenks provides some example data for the 48 contiguous states of the USA – except he omits Rhode Island from the listing!
- But not the worked examples
Jenks criterion

And here's what he proposes:

The classing procedure is started by dividing the range of the data by the number of classes to be used, and then calculating the means and the average deviations for each class. After inspecting the average deviation values for all classes a decision is made to move one areal unit from the class with the largest average deviation toward the class with the lowest average deviation. New class means and average deviations are then calculated and the process is reiterated until the average deviations for all classes are equal or nearly equal. At this point the sum of the within class deviations for the series should approach, but not necessarily reach a minimal value.

So he's optimising on the variance of the average deviation over the groups
# Worked examples

## Equal Average Deviation Worktable

<table>
<thead>
<tr>
<th>Class</th>
<th>Class limits</th>
<th>N</th>
<th>( \bar{x} )</th>
<th>( \bar{x} )</th>
<th>( \bar{d} )</th>
<th>a.d.</th>
<th>r.a.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3-12</td>
<td>18</td>
<td>117</td>
<td>6.5</td>
<td>40.0</td>
<td>2.222</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>13-22</td>
<td>17</td>
<td>272</td>
<td>16.5</td>
<td>40.0</td>
<td>2.353</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>23-33</td>
<td>7</td>
<td>202</td>
<td>28.9</td>
<td>19.3</td>
<td>2.757</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>34-42</td>
<td>3</td>
<td>119</td>
<td>39.7</td>
<td>5.3</td>
<td>1.767</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>43-50</td>
<td>3</td>
<td>143</td>
<td>47.7</td>
<td>7.3</td>
<td>2.433</td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>48</td>
<td>853</td>
<td>111.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Equal Relative Average Deviation Worktable

<table>
<thead>
<tr>
<th>Class</th>
<th>Class limits</th>
<th>N</th>
<th>( \bar{x} )</th>
<th>( \bar{x} )</th>
<th>( \bar{d} )</th>
<th>a.d.</th>
<th>r.a.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3-5</td>
<td>8</td>
<td>33</td>
<td>4.1</td>
<td>3.4</td>
<td>.425</td>
<td>.104</td>
</tr>
<tr>
<td>2</td>
<td>6-11</td>
<td>8</td>
<td>60</td>
<td>7.5</td>
<td>9.0</td>
<td>1.125</td>
<td>.150</td>
</tr>
<tr>
<td>3</td>
<td>12-19</td>
<td>17</td>
<td>252</td>
<td>14.8</td>
<td>30.8</td>
<td>1.812</td>
<td>.122</td>
</tr>
<tr>
<td>4</td>
<td>20-30</td>
<td>7</td>
<td>181</td>
<td>26.9</td>
<td>23.9</td>
<td>3.414</td>
<td>.127</td>
</tr>
<tr>
<td>5</td>
<td>31-50</td>
<td>8</td>
<td>327</td>
<td>40.9</td>
<td>43.0</td>
<td>5.375</td>
<td>.131</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>48</td>
<td>853</td>
<td>110.1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• Nothing about SDAM, SDBC, GVF &c &c
Another rarity:

http://www.macwright.org/2013/02/18/literate-jenks.html [illustration on left]

Jenks, GF, 1977, **Optimal data classification for choropleth maps**, Occasional Paper No. 2, Lawrence, Kansas: University of Kansas, Department of Geography
Jenks, 1977

- It contains code...
- ... which few people would appear to have seen

The "jenks" style has been ported from Jenks’ Basic code, and has been checked for consistency with ArcView, ArcGIS, and MapInfo (with some remaining differences); added here thanks to Hisaji Ono.

- ... writes the creator of the R \textit{classInt} package ...
- ... with that magisterial authority of people who are wrong ...
BUT...

• There's a problem...

• It's not BASIC...
AND...

- Jenks himself says:

It is the author's view that the Fisher algorithm is the best external grouping technique. The algorithm was first proposed by W.D. Fisher in a paper entitled, "On Grouping for Maximum Homogeneity," and later discussed by Hartigan in his book, CLUSTERING ALGORITHMS (see bibliography).

- http://lib.stat.cmu.edu/cmlib/src/cluster/fish.f

ON GROUPING FOR MAXIMUM HOMOGENEITY*

WALTER D. FISHER
Kansas State College

Given a set of arbitrary numbers, what is a practical procedure for grouping them so that the variance within groups is minimized? An answer to this question, including a description of an automatic computer program, is given for problems up to the size where 200 numbers are to be placed in 10 groups. Two basic types of problem are discussed and illustrated.

Hartigan's code dates from the early 1970s...
• And here's John Hartigan's coding of Fisher's algorithm:

```fortran
C... COMPUTE SG,MG ITERATIVELY
    DO 30 I=2,N
    SS=0.
    S=0.
    DO 31 II=1,I
        III=I-[I+1]
        SS=SS+X(III)**2
        S=S+X(III)
        SN=II
    VAR=SS-S**2/SN
    IK=III-1
    IF (IK.EQ.0) GO TO 31
    DO 32 J=1,K
        IF (J.EQ.1) GO TO 32
        IF (SG(I,J).LT.VAR+SG(IK,J-1)) GO TO 32
        MG(I,J)=III
        SG(I,J)=VAR+SG(IK,J-1)
    32 CONTINUE
    31 CONTINUE
    SG(I,1)=VAR
    30 MG(I,1)=1
```
• And if you look at

• You'll discover Jenks' adaptation of Hartigan's coding of Fisher's algorithm
Jenks, 1977

• Jenks' FORTRAN code has two subroutines of interest:
  – FAVAR: "This subroutine creates optimal data classes as measured by variance"
    • The sum of squared deviations about class means
  – FAABS: "This subroutine creates optimal data classes as measured by absolute deviations"
    • The sum of absolute deviations about class medians

• Cindy Brewer (Penn State) and Linda Pickles (National Cancer Institute) note in a paper in the AAAG (2002):

  ESRI's documentation did not explain the specifics of their algorithm, but the ArcView natural-breaks method produced the same class breaks as did the Jenks algorithm that minimizes the sum of absolute deviations from class means (Terry Slocum, personal communication, e-mail, May 2000).
Criterion

- At each step the algorithm computes the cluster errors: the sum of the cluster diameters over the clusters
- That with the smallest error is selected
- The cluster diameter can be:
  - Sum of the within class squared deviations from the mean
  - Sum of the within class absolute deviations from the median [easier for hand calculation, p132]
- Thus the cluster errors can be plotted from a range of value of K to suggest the most appropriate
6.1 INTRODUCTION

A partition of a set of objects is a family of subsets such that each object lies in exactly one member of the family. If \( R(M, K) \) is the number of partitions of \( M \) objects into \( K \) sets, it can be seen by induction on \( M \) that

\[
R(M, K) = KR(M - 1, K) + R(M - 1, K - 1).
\]

Also \( R(M, 1) = R(M, M) = 1 \). As a result, the number of partitions increases approximately as \( K^M \); for example, \( R(10, 3) = 9330 \).

NOTE. The partition is guaranteed optimal, but it is not necessarily unique.

Hartigan, 1973, p130
Scree plots

- Hartigan uses scree plots but he also suggests a stopping rule based on the ratio of the mean square error for $K$ and $K+1$ clusters.
What else does Jenks 1977 recommend?

1. Use standardised data – by which he means counts standardised by some suitable denominator (population/area, farm yield/area, area income/population)
2. Subdivide the data into a limited number of classes
3. Use colours or patterns that are visually related
4. Strive for simplicity
5. Select the method of data classification with care.
But where are the Natural Breaks?

(Fig. 5), or they may follow the “natural break” theory set forth by Alexander and Zahorchak.\textsuperscript{13} Natural break classes are obtained from histogram plots of the data, and the class limits are taken at low spots on the histogram (Fig. 6). The two histograms in Figure

- Jenks GF and Caspall FC, \textbf{1971}, "Error on choroplethic maps: definition, measurement, reduction, AAAG, 61(2), 217-244

- Alexander and Zahorchak were reporting their work to create maps of population density in the USA in 1943

- Jenks and Caspall use measures based on the within class sum of deviations from the mean to evaluate a range of classing methods
Jenks and Caspall, 1971

- Their natural breaks are based on the first four local low points in the histogram.
- Although the NB has the lowest tabular error.
- ... there are 1481 solutions with a lower tabular error.

[complete enumeration of all 3921225 break combinations]

**Table 3.—Tabular Error Indices**

<table>
<thead>
<tr>
<th>Generalizing Technique</th>
<th>Total Tabular Error</th>
<th>Mean or Average Tabular Error</th>
<th>Tabular Error Index (TEI)</th>
<th>Tabular Accuracy Index (TAI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting Number</td>
<td>$602.72</td>
<td>5.91</td>
<td>.31199</td>
<td>.68801</td>
</tr>
<tr>
<td>Quintile</td>
<td>$645.25</td>
<td>6.33</td>
<td>.33400</td>
<td>.66600</td>
</tr>
<tr>
<td>Natural Breaks</td>
<td>$569.42</td>
<td>5.58</td>
<td>.29475</td>
<td>.70525</td>
</tr>
<tr>
<td>Clinographic</td>
<td>$773.18</td>
<td>7.63</td>
<td>.40022</td>
<td>.59978</td>
</tr>
<tr>
<td>Standardized</td>
<td>$595.52</td>
<td>5.84</td>
<td>.30826</td>
<td>.69174</td>
</tr>
</tbody>
</table>

Source: calculated by authors.
Classing

- Choosing the class intervals is a data transformation:
  - Ratio variable $\rightarrow$ ordinal variable

- Consequences:
  - Reduce the range of the data: 1...5 is common
  - Change the data type: ratio $\rightarrow$ ordinal
  - Lose information: only a small number of classes
Jenks 1977

- Recall that he asserts that the 'best' algorithm is Fisher's Exact Optimisation algorithm
- He suggests two objective functions:
  - Sum of absolute within-class deviations from median
  - Sum of squared within-class deviations from the mean
Objective functions

• Fisher's algorithm finds the partition which minimises the within-class sum of squares (WSS).
  – The WSS is known as the *objective function*.

• The problem then is:
  – Which allocation of objects to classes yields the smallest value of the objective function?
  – We call the search for the smallest (or largest) *optimisation*. 
Optimisation

- So, if the combination of an algorithm and an objective function gets you a minimum/maximum, what objective functions are there besides WSS/GVF for choosing class intervals?

- On other words, let's turn this on its head!
### Objective functions

<table>
<thead>
<tr>
<th>Objective function</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance of class ranges</td>
<td>Equal interval</td>
</tr>
<tr>
<td>Variance of class sizes</td>
<td>Quantile</td>
</tr>
<tr>
<td>Variance of class ranges of z-scores</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>Variance of difference between class ranges</td>
<td>Arithmetic</td>
</tr>
<tr>
<td>Variance of class ranges of logged values</td>
<td>Geometric</td>
</tr>
<tr>
<td>Variance of upper:lower class limit ratios</td>
<td>Geometric</td>
</tr>
<tr>
<td>Variance of class ranges of reciprocal values</td>
<td>Reciprocal</td>
</tr>
<tr>
<td>Sum of within-class squared deviations from the mean</td>
<td>Fisher-Jenks</td>
</tr>
<tr>
<td>Sum of within-class absolute deviations from the median</td>
<td>Jenks -Caspall</td>
</tr>
<tr>
<td>Sum of within-class absolute deviations from the mean</td>
<td></td>
</tr>
</tbody>
</table>

We tend to think in terms of the **method** of constructing the intervals. Thinking of them as objective functions in an optimisation problem makes them arise on a *consistent* basis. Nested means don't have an obvious functional form...
Optimal classifications

- So... many of the commonly encountered methods are optimal

- It just depends on what you want to optimise...
So, whose software does what?

- **QGIS**
  - Port of Hartigan's code – **minimise WSS**
  - Sampled if N > 1000

- **GeoDa**
  - Lines 1185:1237 and 93:118 of https://github.com/GeoDaCenter/geoda/blob/master/Explore/CatClassification.cpp
  - **Maximise GVF** (which is **minimising WSS**)

- **R**: Package `classInt`
  - **Minimise WSS** using Fisher's algorithm: Fortran and R port
Well, there's Jenks and there's Jenks...

<table>
<thead>
<tr>
<th>JCFig7</th>
<th>JCFig15</th>
<th>JCFig20</th>
<th>Jenks</th>
<th>Fisher</th>
<th>GeoDa</th>
<th>QGIS</th>
<th>ArcMap104</th>
<th>ArcView32</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>4</td>
<td>35</td>
<td>38</td>
<td>35</td>
<td>35</td>
<td>34</td>
<td>35</td>
<td>10</td>
</tr>
<tr>
<td>b</td>
<td>31</td>
<td>29</td>
<td>40</td>
<td>33</td>
<td>33</td>
<td>37</td>
<td>33</td>
<td>27</td>
</tr>
<tr>
<td>c</td>
<td>43</td>
<td>20</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>15</td>
<td>18</td>
<td>41</td>
</tr>
<tr>
<td>d</td>
<td>15</td>
<td>13</td>
<td>4</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>e</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

- Using the Illinois farm output data in Jenks and Caspall (1971)
  - Jenks-Caspall Fig 7 - uses an arbitrary selection of the first four natural breaks in the data histogram to give 5 classes
  - Fig15 - optimises on 'tabular accuracy index' - (sum of absolute deviations from mean)
  - Fig 20 - optimises on the 'boundary accuracy index'
  - Jenks - Hisaji Ono's R port of Jenk's "probably Basic" code (R `classInt` library)
  - Fisher - Hisaji Ono's compiled version of Hartigan's FORTRAN code
  - GeoDa - revision 1.8.10.14 (July 2016) of Luc Anselin's software
  - QGIS - 'Essen' - revision 2.14.0
  - ArcMap104 - ArcMap for Desktop Advanced 10.4.1
  - ArcView 32 - ArcView GIS 3.2 (1999)
Fig. 15. One of the best, if not the best map of the Illinois data for the tabular map-user. The TAI for this map is .73455, which is nearly three percentage points higher than any of the “traditional” maps (Table 5).

- "One of the best, if not the best..."
- But there are 1222 solutions with a better TAI...
• It would seem that as an optimisation algorithm, the Jenks- Caspall one might not be very good.

• Their objective function is the sum of the absolute deviations from the mean.

• Natural breaks are obtained from a histogram.
Illinois Farm data – ArcGIS plots

(a) (b)
(c) (d)

(a) Fig 7
(b) Fig 15
(c) Fig 20
(d) ArcGIS

Manual classifications for (a) ... (c) taken from Jenks GF and Caspall FC, 1971, Error on choropleth maps: definition, measurement, reduction, *Annals of the Association of American Geographers*, 61(2), 217-244.
Beware...

- One of the most dangerous buttons on any GUI is **Classify**, particularly when it's allied to **Natural Breaks (Jenks)** as an option.

- For a start, there's no such thing as Jenks' Natural Breaks classification.
~1998...

- ArcView 3.x
• The ESRI implementation has remained remarkably stable over nearly 20 years.
What happens in practice

- We often skip the analysis of the data and proceed straight to a choropleth map.

- And if Jenks’ Natural Breaks isn’t the default classification, we often press this button with the vague notion that it’s the best one...
Good practice...

• Whatever you do don't ... accept your software defaults
• Long before you map anything, you should understand the characteristics of the data distribution
• If it's done automatically, it's probably wrong
SCALE and AGGREGATION
It had been known since 1931 that larger areas would yield larger correlation coefficients than smaller ones.

Experiments on the relationship between juvenile delinquency rates and median income in the 252 census tracts of Cleveland OH.
Results: scale

- What they found was with fewer, larger, spatial units, the more negative is the relationship between juvenile delinquency rates and median income.
Variations in the size of the correlation coefficient seem conditioned upon changes in the size of the unit used, with a smaller value of $r$ associated with the smallest unit rather than with the largest.

The ultimate question is whether a geographical area is an entity possessing traits, or merely one characteristic of a trait itself.
MAUP

• This phenomenon was rereported in the 1950 by the sociologist Robinson

• Yule and Kendall (1950) remarked that spatial units could be regarded as 'modifiable' and noted:
  – we seem able to produce any value of the correlation from 0 to 1 merely by choosing an appropriate size of the unit of area for which we measure the yields. Is there any "real" correlation between wheat and potato yields or are our results illusory?

• The term Modifiable Areal Unit Problem (MAUP) was coined by Openshaw
**Scale and aggregation**

- Not only does variation in size appear to affect the value of the correlation coefficient.
- There are many different ways of aggregating 252 zones into 10.
  - Not all sensible; it has to be admitted (R Flowerdew).
- This has an effect too...
Openshaw and Taylor (1979)

• Openshaw and Taylor experimented with data for districts in Iowa
  – Proportion voting Republican
  – Proportion aged over 60

• In an exhaustive set of experiments, Openshaw and Taylor obtained correlations between -0.99 and 0.99 from different levels of aggregation
Scale effect


Aggregation effect
Analysis

• Gehlke and Biehl's original work was framed in terms of correlation
  – Much subsequent work has also focussed on correlation, notably Openshaw and Taylor

• We know what happens when we have two variables, but what happens when we have forty?

• G+B, O+T used artificial aggregations of units
• What happens with real ones?
Experiments with Irish Census data
Enabling environment

- Count data from the 2011 Irish Census of Population is freely available from the Central Statistics Office for 17 levels of geography!

<table>
<thead>
<tr>
<th>Geography</th>
<th>Availability Download Tables</th>
<th>Availability SAPMAP WFR Mapping Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counties (34)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Electoral Divisions (3,409)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Small Areas (18,488)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Dáil Constituencies 2007 (43)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Dáil Constituencies 2013 (10)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Gaeltacht Areas (7)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Legal Towns and Cities (85)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Towns/Cities (Settlements) (824)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Local Electoral Areas (171)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Regional Authority Areas (8)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Provinces (4)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Dioceses (24)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Dublin Parishes (197)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Garda Regions (5)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Garda Divisions (28)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Garda Districts (96)</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Garda Sub Divisions (563)</td>
<td>All</td>
<td>All</td>
</tr>
</tbody>
</table>
Spatial hierarchies

• Some of these data form a spatial hierarchy, others represent misaligned geographies

• Provinces (4)
  – Regional Authority Areas (8)
    • Counties (34)
      – Electoral Divisions (3409)
        » Small Areas (18488)

• Others do not
  – Dáil Constituencies straddle county boundaries
  – Higher level An Garda Síochána areas straddle some county boundaries but to a lesser degree than Dáil Constituencies
## Spatial units

<table>
<thead>
<tr>
<th>Unit</th>
<th>N</th>
<th>Hierarchy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Area</td>
<td>18488</td>
<td>✓</td>
</tr>
<tr>
<td>Electoral Division</td>
<td>3409</td>
<td>✓</td>
</tr>
<tr>
<td>Garda Sub District</td>
<td>563</td>
<td>Mostly ✓</td>
</tr>
<tr>
<td>Local Electoral Area</td>
<td>171</td>
<td>✓</td>
</tr>
<tr>
<td>Garda District</td>
<td>96</td>
<td>Mostly ✓</td>
</tr>
<tr>
<td>Dáil Constituency 2007</td>
<td>43</td>
<td>Overlaps</td>
</tr>
<tr>
<td>Dáil Constituency 2013</td>
<td>40</td>
<td>Overlaps</td>
</tr>
<tr>
<td>County</td>
<td>34</td>
<td>✓</td>
</tr>
<tr>
<td>Regional Authority Area</td>
<td>8</td>
<td>✓</td>
</tr>
</tbody>
</table>
Variables

- We have identified forty variables based on the set of general socio-economic indicators which form the basis of the UK's Output Area Classification
- Represent a broad set of domains
  - Age structure
  - Nationality
  - Family status
  - Housing
  - Employment/economy
  - Social class
- [https://rpubs.com/chrisbrunsdon/14998](https://rpubs.com/chrisbrunsdon/14998)
UNIVARIATE EXPLORATION
Univariate behaviour

• With 40 variables measured at 9 spatial aggregations for the same study area we can examine
  – Variations in the mean
  – Variations in the standard deviation
  – Variations in the coefficient of variation

• We are at the stage of exploration... What happens?
Visualisation

- Boxplots show the variation in the mean for each of the 49 variables with a separate boxplot for each set of spatial units.

- The polylines in the parallel coordinate plots connect variables between each set of spatial units – they're coloured by the mean value for the Small Areas.
Means

Number of spatial units

Means

<-- Smaller ---- Area of Spatial Units ---- Larger -->
Standard deviations
Coefficient of variation

Coefficients of Variation

Coefficients of Variation

Number of spatial units

<-- Smaller ---- Area of Spatial Units ---- Larger -->
Univariate

- There's little to chose between the distributions of the means
  - Perhaps the pattern of outliers is more consistent for larger spatial units
- The standard deviations get smaller as the areas of the spatial units get larger
- The within-area variance increases as the units get larger
- This pattern is reflected, hardly surprisingly, in the values for the coefficients of variation
BIVARIATE EXPLORATION
Bivariate

- Gehlke and Biehl used two variables, as did Openshaw and Taylor.
- With 40 variables we can compute all possible pairwise correlations.
- As the correlation matrix is symmetrical, and the correlations of the variables with themselves are 1, this gives us 780 separate correlations for each set of spatial units.
- ... 7020 in all... somewhat fewer than Openshaw and Taylor.
All pairwise correlations (i)

- Each dot on the plot represents the correlation between two variables.
- The 780 correlations are larger, in general, as the spatial units get larger compared with the Small Areas.
All pairwise correlations (ii)

- The process continues as the number of spatial units decreases relative to the number of Small Areas.

- If only Charles and Katherine had had access to a computer...
• The correlations do increase in size, and there are fewer outliers, with an increase in the size of spatial unit.

• We've observed that the variance decreases with increase in spatial unit size, so the covariance must, in general, increase.

• It would be useful to be able to look at the variables as a set rather than as 780 separate pairs...
MULTIVARIATE EXPLORATION
Multivariate

• The forty variables can be thought of as the dimensions of the dataset
• However, many are highly correlated:
  – at Small Area level the correlation between the DINK and Age25_44 variables is 0.794
  – DINK and PrivateRenters: 0.777
  – DINK and Born_outside_Ireland: 0.665
  – So are these really 4 separate dimensions?
• The question arises as to the real dimensionality of the data
• A useful technique here is Principal Components Analysis
PCA

- Principal Components represent a transformation of an M variable dataset, where the M variables are correlated, to an M component dataset, where the components have a correlation of zero.
- All the variance in the M variables is retained in the M components.
- Each component is calculated as a linear combination of the original M variables.
- Components are mutually orthogonal – the angle between any pair of components is 90°.
1st and 2nd component eigenvalues

Eigenvalues of 1st Component

Eigenvalues of 2nd Component

Key
- Hierarchy
- Overlaps
- Partial Hierarchy

log(Number of Areas)
Visualising the changing structures

• A heatmap is a convenient method of visualising the patterns of loading on the eigenvectors as the spatial scale changes.

• As the areas get larger the eigenvectors account for more variance as the dimensionality reduces.

• Contrasts within each eigenvector appear to be dependent on scale.
1st Eigenvector

- The loadings as a heatmap.
- Tan represents positive loadings and green negative loadings.
- This component contrasts rural family areas with the younger urban renter areas.
- Whilst reasonably consistent across the spatial scales, the polarisation is less marked with larger areas.
The second component contrasts urban struggling families with the affluent suburban elite.

At the larger spatial scales the emphasis shifts to include older pensioner areas among the less affluent groups and away from a qualified workforce and larger houses.

Implications for geodemographics!
What have we found?

• This isn't the last word on the MAUP

• With larger areas we lose between-area variance
  – The total variance doesn't decrease so we get larger correlations

• With larger areas we lose dimensionality
  – Multivariate structures are also scale dependent, but not always in a predictable fashion
Implications

• Geodemographic classifications might not be as objective as we'd like them to be
  – Data relationships are scale dependent

• Choropleth maps
  – We regionalise smaller spatial units to get larger ones
    • Openshaw's AZP
    • Openshaw/ Coombes ERA/ Functional Regions
    • Masser- Brown Intramax algorithm
    • Community detection in mobile phone call interactions
      – using Thiessen polygons around masts as building blocks
Choropleth maps

- We have no metric to tell us what is 'good', even if such a concept exists
  - What is optimal depends on the objective function

- Aggregating spatial units loses information

- Data classing loses information

- We probably need to find an alternative to the classified choropleth map
Tobler suggests...

Choropleth Maps Without Class Intervals?

W. R. Tobler

It is now technologically feasible to produce virtually continuous shades of grey by using automatic map drawing equipment. It is therefore no longer necessary for the cartographer to “quantize” data by combining values into class intervals.