INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Exploration of spatial flow patterns using projection pursuit methods and dynamic visualization

Liu, Lin, Ph.D.
The Ohio State University, 1994
EXPLORATION OF SPATIAL FLOW PATTERNS
USING PROJECTION PURSUIT METHODS AND DYNAMIC VISUALIZATION

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By
Lin Liu, B.S., M.S.

The Ohio State University
1994

Dissertation Committee
Duane F. Marble
Morton E. O'Kelly
Randall Jackson

Approved by
Duane F. Marble
Department of Geography
DEDICATION

To My Family
ACKNOWLEDGEMENTS

I would like to express my sincere thanks to Professor Duane Marble for his guidance and insight throughout the dissertation research. His encouragement and support in the past four years have been the inspiration to my professional pursuit. Gratitude goes to Professors M. E. O'Kelly and R. Jackson for their suggestions and comments in this dissertation research. Thanks go to Professors E. Casetti, and L. A. Brown for their help. Thanks also go to Dave Wilcox for proof-reading the draft of this dissertation. I am especially indebted to my friend Dr. Zaiyong Gou, who gave me valuable suggestions and allowed me to use the data set he collected. Thanks also go to my other friends Drs. Yong Lao, Qin Tang and Demin Xiong for the wonderful time we had during the years we spent together at the Ohio State University. I am deeply grateful to my parents, who have been encouraging me to pursue higher-level excellence since I was young. To my wife, Rong, I thank you wholeheartedly for your unconditional support. Also, I like to thank my daughter, Yong Xu, for bringing happiness and joy into my life.
VITA

January 19, 1965 .............. Born - Xiangtan, Hunan Province, P.R. China

July, 1984 .................... B.S. in Geography at Peking University, Beijing, China

July, 1987 .................... M.S. in Remote Sensing and Cartography at Peking University

1987-1989 .................... Assistant Professor in the Institute of Remote Sensing
                        at Peking University, Beijing, China

1989-1990 .................... Computer Software Engineer in the Information Center
                        at the State Land Administration, Beijing, China

Summer, 1991 .................. Student Intern at Environmental Systems Research
                        Institute (ESRI), Redlands, California

1990-Present ................... Graduate Teaching Associate in the Department of
                        Geography at The Ohio State University, Columbus, OH

PUBLICATIONS

                        Proceedings of GIS/LIS '93.

Xu, Xiru, Jicheng Cheng, Zheng Wang, Kuengqin Xie, and Lin Liu, 1987. The Design of
                        GIS Software System and Its Application to Estimation of Winter Wheat Yield.
                        Proceedings of Beijing International GIS Meeting, May, 1987

                        Youth Geography, No.4, Vol.1, December, 1985

FIELDS OF STUDY

Major Field: Geography
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEDICATION</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>VITA</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>x</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xi</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>I INTRODUCTION</td>
<td></td>
</tr>
<tr>
<td>Complexity of Spatial Flow Data Sets</td>
<td>1</td>
</tr>
<tr>
<td>Flow Data</td>
<td>2</td>
</tr>
<tr>
<td>O-D Attribute Data</td>
<td>4</td>
</tr>
<tr>
<td>Spatial Data with Respect to Origins and Destinations</td>
<td>4</td>
</tr>
<tr>
<td>Current Status of Spatial Flow Analysis</td>
<td>7</td>
</tr>
<tr>
<td>Context of the Research</td>
<td>11</td>
</tr>
<tr>
<td>Methods for Spatial Flow Analysis</td>
<td>13</td>
</tr>
<tr>
<td>Current methods for Spatial Flow Data Analysis in Geography</td>
<td>13</td>
</tr>
<tr>
<td>SV, EDA, Dynamic Graphics and Projection Pursuit</td>
<td>14</td>
</tr>
<tr>
<td>A New Methodology for Spatial Flow Analysis</td>
<td>15</td>
</tr>
<tr>
<td>Implementation</td>
<td>17</td>
</tr>
<tr>
<td>Organization of the Dissertation</td>
<td>18</td>
</tr>
<tr>
<td>II LITERATURE REVIEW</td>
<td>19</td>
</tr>
<tr>
<td>Cartographic Representation of Spatial Flows</td>
<td>19</td>
</tr>
<tr>
<td>Computer Mapping of Spatial Flows</td>
<td>22</td>
</tr>
<tr>
<td>Graphic Representation of Multiple Flow Variables</td>
<td>26</td>
</tr>
<tr>
<td>Summary</td>
<td>29</td>
</tr>
<tr>
<td>Complexity Reduction Methods as Applied to Flow Studies</td>
<td>29</td>
</tr>
</tbody>
</table>
Flow Mapping .......................................................... 72
  Graphic Representation of a Single Flow Variable .... 73
  Graphic Representation of Multiple Flow Variables ... 73
Statistical Plot of Flow Variables ........................................... 75
  Histogram of a Flow Variable ........................................ 75
  Scatterplots of Flow Variables ....................................... 76
Statistical Plot of O-D Attribute Variables ......................... 77
Choropleth Mapping of O-D Attributes ................................. 77
Linking the Flow Map, Choropleth Map, Statistical Plot Flow
Variables, and Statistical Plot of O-D Attribute Variables
  for Detecting Spatial Flow Patterns ................................. 78
    Detection of Dominant Flows ...................................... 83
    Detection of Zero-flows ............................................ 83
    Detection of Similar Flows ....................................... 83
    Detection of Relationships between Flow Variables .... 83
    Detection of Relationships between Flow Variables
      and O-D Attribute Variables .................................... 84
    Detection of Relationships between Flow Variables
      and Distance ....................................................... 84
    Detection of Changing Flow Patterns in Time ................ 84
Summary ............................................................................. 85

IV IMPLEMENTATION OF THE PROTOTYPE SOFTWARE: PPFLOW ..... 87

  Structure and Organization of PPFLOW ............................. 87
  System Requirement ..................................................... 88
  Data Organization and File Structure ............................... 89
  Flow Chart and Modular Design of PPFLOW ..................... 90
  User Interface ........................................................... 93
  Hardware and Software Consideration ............................ 95
Modules and Functionality of PPFLOW ................................. 98
  Data Input Module ........................................................ 98
  Flow Mapping Module .................................................. 99
  Flow Representation Module ......................................... 100
  Flow Magnitude Adjusting Module ................................ 102
  O-D Distance Adjusting Module .................................... 103
  Choropleth Mapping Module ......................................... 104
  Flow Variable Plotting Module ...................................... 104
  O-D Attribute Variable Plotting Module ......................... 105
  Projection Pursuit Module ........................................... 106
  Dynamic Brushing Module ............................................. 105
Projection Pursuit Optimization ........................................ 108
  Calculation of the PP Index and its Partial Derivative .... 108
D Usage of PPFLOW ................................................................. 161
E The First Derivative of the Moment Index ............................. 177
BIBLIOGRAPHY ................................................................. 180
LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Comparison of factor analysis and projection pursuit methods. Four points in a three-dimensional space are projected onto one dimension using factor analysis and projection pursuit methods. On each of the two factors, four points are projected to two points, while there are four separate points on the PP variable.</td>
</tr>
<tr>
<td>2</td>
<td>PP_FLOW1 produced by PP from MIGRAT60, MIGRAT70 and MIGRAT80. Initial position is set at (1,2,3)</td>
</tr>
<tr>
<td>3</td>
<td>PP_ATT1 generated by PP from UNEMPLOY, CRIME and GNP/PC. Initial direction is (1,2,15)</td>
</tr>
<tr>
<td>4</td>
<td>PP_ATT2 generated by PP from UNEMPLOY, CRIME and GNP/PC. Initial direction is (3,2,3)</td>
</tr>
<tr>
<td>FIGURE</td>
<td>PAGE</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>1. A stack of O-D matrices. Each matrix represents a flow variable. The value of cell (i, j) represents the magnitude of a flow from origin i to destination j.</td>
<td>3</td>
</tr>
<tr>
<td>2. U.S. Foreign Trade in 1948. Because the contiguity of earth's surface is destroyed in the two-dimensional map, the flows across Indian Ocean can not be displayed. [Source: Ullman 1957]</td>
<td>6</td>
</tr>
<tr>
<td>3. Air traffic flows. Gradients of shading in corridors with constant width are used to portray flows. [Source: Wittick 1976]</td>
<td>7</td>
</tr>
<tr>
<td>4. Air traffic flows. Lines and numbers are used to portray flows. [Source: Wittick 1976]</td>
<td>8</td>
</tr>
<tr>
<td>5. The map of factor 1, which is most closely related to the major flows of transportation equipment between census regions. [Source: Black 1973]</td>
<td>9</td>
</tr>
<tr>
<td>6. Exploratory data analysis versus confirmatory data analysis in the studies of spatial flows.</td>
<td>12</td>
</tr>
<tr>
<td>7. A methodology for detecting spatial flow patterns. It integrates complexity reduction method and data visualization/exploration method.</td>
<td>16</td>
</tr>
<tr>
<td>8. Railway Traffic. Lines with varying width are used to represent railway traffic. [Source: Ullman 1957]</td>
<td>21</td>
</tr>
<tr>
<td>9. Barge and Raft Traffic. Lines with varying width are used to represent water routes and traffic volumes. Two-way traffic is shown on the map. [Source: Ullman 1957]</td>
<td>23</td>
</tr>
</tbody>
</table>

12 A perspective-view map of a transit service. Speed limits of bus lines are represented by the height of blocks. Peak hour bus load, average day bus load, and night bus load are displayed at each bus stop. [Source: Noguchi and Schneider 1977] .................................................................28

13 Comparison of the results from PP and PCA. P-index is the index for choosing projections. The larger the P-index, the better the projected result tends to be. (a) PCA result: horizontal axis shows the largest principal component P-index is 1.00x10^5. All the 975 data points are in the same cluster. (b) PP result: horizontal axis shows the projection with a P-index of 3.37x10^5. There are obvious clusters. For the purpose of clustering, (b) is apparently better than (a). [source: Friedman and Tukey 1974] ............................................................................................................35

14 Artist's three-dimensional rendition of diabetes data in PRIM-9 [Source: Reaven and Miller 1979] .......................................................................44

15 Brushing a scatterplot matrix. The brush is highlighting points with low values of hardness. The (3,2) panel (row 3, column 2) shows the dependence of abrasion loss on tensile strength for hardness held fixed to low values. [Source: Becker, et al 1987b, p. 365] ............................................46

16 Airline delays. This map shows hypothetical state to state airline delay data on a day where bad weather conditions are affecting most of the United States. There are so many line segments on the map that it is difficult to determine anything from the display. [source: Becker, et al 1990a] ......................... 51

17 The most serious airline delays (using dynamic link display technique). This map shows the state to state airline data as in Figure 16, but the slider on the lower scale has restricted the map to showing only the routes on which delays are greater than 100 minutes. This map makes it obvious that the primary delays on this day are into the Northeast, perhaps as a result of a snowstorm. [source: Becker, et al 1990a] ................................. 52

18 Node map (using dynamic node display technique). The rectangles encode the average air traffic delays at each location. The horizontal extent of the rectangle represents delays for inbound traffic, the vertical extent for outbound traffic. This map shows one particular point in time, hour 14:30. The dynamic node map tool displays maps like this in a rapidly varying time sequence, much like a movie, to allow the analyst to see time
are displayed. ................................................................. 118

32 Flow map using the line and number method. Two directional
flows between Ohio, California and Florida are displayed ........................ 124

33 Flow map using the proportional width method. Two-directional flows
between Ohio, Louisiana and Wyoming are displayed. .............................. 125

34 Flow map using the quantile width method. Two-directional flows
between Ohio, Louisiana and Wyoming are displayed. .............................. 126

35 Choropleth map of unemployment rate of 1980, overlaid by the
PP-synthesized migration flows between Ohio, California and Florida. ...... 127

36 A flow map containing all migration flows from or to Ohio in
the period from 1975 to 1980. ..................................................... 128

37 A flow map containing all migration flows from or to Ohio, and
that are within 502 miles of Ohio. .................................................. 129

38 A flow map containing all migration flows that are connected with Ohio,
California and Florida, and that are within 528 miles from each of
these three states. ........................................................................ 130

39 A flow map showing all migration flows that are within 425 miles
from each of the 48 states. .............................................................. 131

40 A flow map containing migration flows that are connected to Ohio
and are larger than 31377 persons. .................................................... 132

41 Dynamic brushing initiated from flow map/choropleth map.
The anchor states are California and Florida. The brushed state is Ohio.
The PP-synthesized migration flows between the anchor states and the
brushed state are displayed. .............................................................. 133

42 Dynamic brushing initiated from the histogram of the flow
variable PP FLOW1. Three bars of the histogram, and the flows within
the brushed bars are displayed on the flow map. ..................................... 134

43 Dynamic brushing initiated from scatterplot triangle of flow variables.
Two outliers in panel (4, 2) are brushed, and the flows represented
Dynamic brushing initiated from scatterplot triangle of O-D attribute variables. Five outliers in panel (2, 1) are brushed, and the flows between the states represented by the brushed outliers are displayed on the flow map. ............................... 136

Large migration flows during the period from 1955 to 1960. .................... 138

Large migration flows during the period from 1965 to 1970 ..................... 139

Large migration flows during the period from 1975 to 1980 .................... 141


A methodology for general geographic research .................................. 153

File selection menu ................................................................. 162

Flow map menu ................................................................. 163

Flow variable selection menu ................................................. 163

Color selection menu .......................................................... 163

Flow representation selection menu ........................................ 165

Choropleth map menu .......................................................... 166

Attribute variable selection menu .......................................... 166

Flow variable plot menu .......................................................... 167

Attribute variable plot menu ..................................................... 168

Projection pursuit menu .......................................................... 170

Flow PP save selection menu ..................................................... 170

Flow PP direction selection menu ................................................. 171
<table>
<thead>
<tr>
<th>Page</th>
<th>Section Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>62</td>
<td>Attribute PP save selection menu</td>
<td>172</td>
</tr>
<tr>
<td>63</td>
<td>Attribute PP direction selection menu</td>
<td>172</td>
</tr>
<tr>
<td>64</td>
<td>Brushing menu</td>
<td>173</td>
</tr>
<tr>
<td>65</td>
<td>Slide bars for flow magnitude and distance adjustment</td>
<td>175</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

Spatial pattern analysis has long been a major focus in geographic research. Geographers are always interested in identifying and explaining various kinds of spatial patterns across regions, both in physical geography and in human geography. This research will not deal with spatial patterns in general, but will focus specifically on interregional flow patterns.

It is important to distinguish network flows from interregional flows. Network flows are channelized through a specific network, e.g., a railroad network, while interregional flows do not follow any specific channel, and their paths are usually represented by straight lines connecting origins and destinations. Network flows are not dealt with in this research, and the term spatial flows refers specifically to interregional flows in this document.

Studying systems of complex spatial flows has always been challenging to geographers. Many researchers have attempted to develop new methods to tackle this problem. Some tried statistical methods, such as factor analysis and principal component analysis, to reduce the complexity of flow data. Others used cartographic methods to better represent spatial flows. But their contribution to the effective exploration of the spatial patterns hidden in complex flow data sets has not been substantial.
Fortunately, geographers are not alone in tackling the problem of spatial flows. Researchers in transportation and telecommunication have also been trying to develop new methods for studying transportation and telecommunication flows. New developments in scientific visualization (SV), exploratory data analysis (EDA) and dynamic graphics, if properly integrated, have the capability to provide new tools that can be effectively extended to the study of spatial flows. For example, Becker, et al. (1990a, 1990b, 1991a, 1991b) from AT&T Bell Laboratories applied dynamic graphics techniques to study airline traffic and the number of telephone calls in the United States. Although these new techniques facilitate the identification of flow patterns, the graphic representations of spatial flows are still weak.

It is the intention of this research to draw upon existing work on spatial flow analysis in geography and to integrate it with new techniques developed in EDA, SV and dynamic graphics. From this basis, new methods will be developed to detect spatial flow patterns in complex flow data sets. With the assistance of the methods developed here, a user should be able to identify major flow patterns such as clusters showing the similarity of flows, outliers representing extremely large flows, and relationships between flows and regional social-economic characteristics in a complex flow data set.

Complexity of Spatial Flow Data Sets

The difficulties in spatial flow analysis arise mainly from the complexity of spatial flow data sets. A spatial flow data set may contain not only the flow data itself, but also a wide range of social, economic, and physical data associated with the origins and destinations. The following sections describe three types of data: flow data, origin-destination (O-D) attribute data, and spatial data pertaining to origin and destination regions.
Flow Data

Conventionally, interregional flow data are stored as O-D matrices. Figure 1 shows a stack of such O-D matrices. Each O-D matrix has one row for each flow origin and one column for each flow destination. This research deals only with data sets in which the origins and destinations are the same. For example, interregional flow data within the fifty states of United States generates a 50 by 50 O-D matrix. The value of cell \((i, j)\) represents the magnitude of the flow from origin \(i\) to destination \(j\). As each O-D matrix represents a single flow variable, representation of multiple flow variables requires multiple O-D matrices. For example, three 50 by 50 matrices are needed to represent the flow of rice, wheat and corn between fifty states. Annual interstate migration from 1951 to 1990 would require forty 50 by 50 matrices.

![Figure 1. A stack of O-D matrices. Each matrix represents a single flow variable. The value of cell \((i, j)\) represents the magnitude of the flow from origin \(i\) to destination \(j\).](image-url)
The data volume of an O-D matrix is very large. A 50 by 50 matrix, which describes the state-level, interregional flow data in the United States, has 2,500 entries. If the flow data were disaggregated to the county level, the O-D matrix would have more than 10 million entries because there are more than 3,000 counties in the United States. Recall that each O-D matrix represents only one flow variable; to represent N different flow variables, N such matrices are required. Thus, representing N interregional flows across U.S. counties would require more than N*10 million entries.

**O-D Attribute Data**

In addition to O-D matrices, which represent flow data, a spatial flow data set may also contain a wide range of social, economic, and physical data associated with the individual origins and destinations. These additional data can play an important role in the study of spatial flows. For example, in gravity models the flow between a given O-D pair is a function of the distance between the O-D pair and the attractiveness of the origin and destination. The attractiveness of a place is often related to its characteristics, which reflect local social, political, economic, and physical conditions. Studying relationships between spatial flows and O-D attributes can clarify spatial flow patterns.

**Spatial Data with Respect to Origins and Destinations**

In the study of spatial flow patterns, O-D matrices and O-D attributes alone are not sufficient because they do not provide the spatial data, such as the distance between O-D pairs, required for spatial analysis. Spatial data can be provided in different forms. The simplest form is an O-D distance matrix, in which the value of cell (i, j) is the distance between origin i and destination j. When gravity models are applied to the study of flow
data, such distance data can be used to examine the regularity of "distance decay", which suggests that the flow magnitude drops as the distance between origins and destinations increases. The disadvantage of using an O-D distance matrix in spatial flow analysis is that flow patterns can not be shown in a two-dimensional space because no locational information is provided.

Spatial data pertaining to spatial flow analysis can also be provided as locational coordinates of individual origins and destinations. The coordinates can be given by an (x, y) pair in a cartographic coordinate system such as UTM or the State Plane Coordinate System, or as spherical coordinates such as latitude and longitude. Both can be used to calculate O-D distances. The difference between the two is that the former provides two-dimensional position data while the latter has global position data. In this research, cartographic coordinates are used, and regional flows are portrayed using two-dimensional graphics.

Spherical coordinates are required for global flow studies, and global flow patterns must be represented with three-dimensional graphics. It is not difficult to understand that a two-dimensional map can not represent a three-dimensional globe without destroying its contiguity; as a result, continuous global flows cannot be portrayed on a two-dimensional map. Figure 2 shows a map of U.S. foreign trade in 1948. Since the United States is located at the center of the map, displaying the flows from the U.S. to Asia and Africa does not pose a problem. A problem occurs when one attempts to add the flows between Asia and Africa across the Indian Ocean. This is impossible because the Indian Ocean is divided by the edge of the map.
Figure 2. U.S. Foreign Trade in 1948. Because the contiguity of earth's surface is destroyed in the two-dimensional map, the flows across Indian Ocean can not be displayed. [Source: Ullman 1957]
Current Status of Spatial Flow Analysis

Geographers have long applied gravity models to study of flow data. In a gravity model, the magnitudes of the flows are expressed as a function of the distances between O-D pairs and social-economic variables that are associated with individual origins and destinations. The work along this line is not "true" spatial flow analysis because only distance is considered which alone can not fully define the spatial flow relations.

A number of researchers (Kern and Rushton, 1969; Wittick, 1976; Tufte, 1983; Tobler, 1987) have studied the graphic representation of spatial flows. Three methods for flow representation have been developed. The first connects O-D pairs by using bands with width proportional to flow magnitude (Figure 2). The second uses constant width corridors

Figure 3. Air traffic flows. Gradients of shading in corridors with constant width are used to portray flows. [Source: Wittick 1976]
connecting O-D pairs, and then fills these corridors with shading gradients to represent flow magnitude (Figure 3). The third method for quantitative portrayal of spatial flows uses lines to connect O-D pairs and displays the exact flow volumes beside the lines (Figure 4). Although this method provides accurate flow information, it is not as visually effective as the first two methods.

Figure 4. Air traffic flows. Lines and numbers are used to portray flows. [Source: Wittick 1976]

Gou (1993) combined scientific visualization, exploratory data analysis and dynamic graphics to explore spatial flow patterns. He implemented a two-way brushing technique to link the map representation of spatial flows to a statistical plot of O-D attribute data, hoping to reveal spatial flow patterns based on relationships between flows and O-D attributes. His work is indeed a step forward compared to previous studies. But, his approach focused only
on analysis of a single flow variable, and is not effective when multiple flow variables are involved.

To deal with a flow data set containing multiple flow variables, Black (1973) used dyadic factor analysis to derive major factors from multiple flow variables, and then drew the factors onto separate two-dimensional maps using lines and arrows to connect origins and destinations. Figure 5 displays his factor 1, which is primarily related to the major flows of transportation equipment. It is clear that the flows are mainly from Middle Atlantic (New York is in this region) to other regions. However, the magnitude of flows is not represented

Figure 5. The map of factor 1, which is most closely related to the major flows of transportation equipment between census regions. [Source: Black 1973]
on the map. The presence of state boundaries confuses the match between centroids and boundaries of the nine census regions.

The literature on global flow pattern exploration has yet to appear, perhaps because of the difficulty in representing flows on three-dimensional graphics. Tobler (1987) addressed the difficulty of studying international movement and stated the following:

"..., but truly international movement is different because the spherical nature of the earth has not been taken into account. Thus one can imagine drawing the flow arrows along great circles on an oblique orthographic view of a hemisphere; the map projection problem becomes even more difficult when a movement pattern over the entire world must be shown. (pp. 162-163)

The problem is not the cartographic projection because no projection can show a movement over the entire world. The solution to this problem will depend upon a three-dimensional portrayal of the earth, which can be implemented by using the high speed, three-dimensional computer graphics. Dynamic rotation of a three-dimensional earth was implemented by Sandhu (1990) and Wentz (1990), and was used to display temporal earthquake data. This technique has not been used to study global flows, perhaps because of difficulties in the graphic representation of global flows.

Lacking a good spatio-temporal data structure, geographers and GIS specialists have not been able to effectively bring the temporal dimension into geographic analysis. Even though some recent progress has been made on theoretic spatio-temporal data models recently (Langran, 1989, 1991; Peuquet, 1993), a practical spatio-temporal GIS is still not available. Like other geographic analysis, studying the temporal aspects of spatial flows remains challenging.
Context of the Research

It is clear that there are many unsolved problems in the study of spatial flows. Because of time limitations, it is unrealistic to attempt to solve all the problems in this dissertation. Instead, the present research focuses on the detection of spatial flow patterns in three selected areas:

1) improving the methods that are currently used in the study of interregional flow data sets containing a single flow variable;

2) developing new methods to facilitate the study of interregional flow data sets containing multiple flow variables, e.g., the study of overall flow patterns of rice, wheat, and corn between fifty states of United States;

3) applying the methods developed in 2) to regional flow data sets containing flow data incorporating time series, e.g., the study of changing migration patterns between the fifty states of the United States from 1950 to 1990.

A major contribution of this research to the study of spatial flow phenomena is the development of new methods to deal with multiple flow variables. Previous research on flow exploration and visualization has focused on the application of dynamic graphics and the graphic representations of a single flow variable. These techniques are not sufficient when the flow data set contains multiple flow variables.

A complete study of spatial flow patterns requires a recursive flow pattern detection and examination process (Figure 6). This involves two different kinds of data analyses:
Figure 6. Exploratory data analysis versus confirmatory data analysis in the study of spatial flows.

Exploratory data analysis (EDA) and confirmatory data analysis (CDA). The former utilizes visual tools and numeric tools to explore flow patterns such as clusters, outliers and relationships, leading towards hypothesis generation. The latter uses statistical tools to test the significance of newly generated hypotheses. When a hypothesis is tested to be insignificant and is rejected, one may need to go back to the visual and numeric tools to generate new hypotheses. This process is repeated until hypotheses are accepted.
This research focuses on the detection of spatial flow patterns, which can lead to the generation of hypotheses. The testing of hypotheses will not be addressed here.

Methods for Spatial Flow Analysis

Current Methods for Spatial Flow Data Analysis in Geography

Currently, two different methods are used for the study of geographic flows. In statistics, factor analysis and principal component analysis are used to reduce the number of flow variables, regression methods are used to build relationships between flow variables. One important characteristic of statistics is that it rests upon a series of assumptions such as random sampling and tightly specified distributions (usually normal distributions). Since not all geographic flow data sets have these statistical properties, the application of statistical methods often does not yield promising results. Cartographic methods, on the other hand, are concerned with the graphic representation of spatial flows. Most of the present representations are static, and are not effective when flow data volume is large or for multiple flow variables.

Current methods are not sufficient for exploring spatial flow patterns, and more effective methods need to be developed. Along with the rapid development of computer science, some tools such as scientific visualization (SV) and dynamic graphics are have been developed for scientific research. In statistics, exploratory data analysis (EDA) is a new method that complements traditional confirmatory data analysis. The next section introduces these new tools and describes how they have been adapted to the study of spatial flows.
**SV, EDA, Dynamic Graphics and Projection Pursuit**

The rapid development of computer graphics has created a new approach, scientific visualization (SV), for scientific research. SV consists of a set of tools that permits researchers to explore complex data sets interactively within environments that enhance and enrich the researchers' own intuition. Some scientists regard SV as "an act of cognition" (MacEachren, et al. 1992). In the past four years, SV tools have been applied to study spatial flows (Becker, et al. 1990a, 1990b, 1991a, 1991b; Gou, 1993). But the previous work was mostly confined to simple data sets with a single flow variable. This research will extend the scientific visualization to the study of spatial flow patterns in a complex data set with multiple flow variables.

Exploratory Data Analysis (EDA), first introduced by Tukey (1977a), provides comprehensible pictures of a data set so that one can objectively analyze data without any pre-assumptions. The exploratory nature of EDA fits well into the study of spatial flow patterns. On the basis of EDA, Becker, et al. (1987a, 1987b) developed a dynamic brushing technique to link scatterplots of multiple variables. Recently, a number of geographers (Haslett, et al. 1991; Tang 1993; Gou 1993) have extended the dynamic brushing technique to build linkages between the map display and scatterplots of multiple geographic variables. For example, Gou (1993) used a brushing technique to link the map of a single flow variable and the scatterplots of multiple O-D attribute variables. This research will show how EDA and dynamic graphics can be extended to the study of spatial flow patterns in a data set with multiple flow variables.
The term "projection pursuit" was first introduced by Friedman and Tukey (1974). Unlike SV, which is a graphic method, projection pursuit (PP) is based on numeric analysis. The projection pursuit method aims to uncover interesting structures hidden in a multi-dimensional\(^1\) data set by projecting it onto a lower dimension for visual inspection (Jones and Sibson, 1987). The selected projections are similar to the principal components in principal component analysis (PCA) in that both are linear combinations of the original variables, but PP is more effective than PCA at revealing structures such as clusters and outliers (Friedman and Tukey 1974; Jones and Sibson 1987). The capability of PP methods to uncover structures is attractive in the study of multiple flow variables, especially when the number of flow variables is very large. Although it appears that PP methods have not been used in geography, there is no reason to ignore this powerful structure revealing tool in spatial flow analysis. This research applies PP methods to reduce the complexity of flow data sets.

*A New Methodology for Spatial Flow Analysis*

Based on existing work on spatial flow analysis in geography, we know that conventional statistical methods and existing cartographic methods alone are not adequate for spatial flow analysis. This research suggests a new methodology for detecting spatial flow patterns. As displayed in Figure 7, the methodology consists of two stages. The first stage is complexity reduction, which can be skipped if the flow data set contains only a small number of flow variables and O-D attribute variables. In this stage, projection pursuit

---

\(^1\) The term dimension here refers to the number of variables in a data set, and is not related to the dimensions of geographic space, which can be one, two, or three-dimensional.
methods are adapted to reduce the number of flow variables and O-D attribute variables in a complex flow data set. The second stage is flow visualization and exploration. Spatial flows are explored in an interactive environment that integrates SV, EDA and dynamic graphics. Such an environment can enrich a researcher's intuition in identifying spatial flow patterns, which can lead to new hypotheses.
Implementation

The proposed new methodologies are implemented as an X-Window based prototype system PPFLOW (Projection Pursuit FLOW Data Analysis System) on a Silicon Graphics workstation. The system is written in C, with the Xlib and Motif libraries and has an interface with the look-and-feel of Motif.

PPFLOW consists of ten modules. The Data Input Module loads flow data sets into the system. The Flow Mapping Module maps a selected flow variable onto a flow map with proper spatial flow representation. The Flow Representation Module allows the user to select a different graphic representation for spatial flows at any stage in PPFLOW. The Query by Flow Magnitude Module allows the user to explore spatial flow distribution based on the magnitude of flows. The Query by O-D Distance Module gives the user the ability to examine the spatial organization of flows based on O-D distance. The Choropleth Mapping Module shows the distribution of selected social-economic characteristics of the origin destination regions. The Flow Variable Plotting Module draws scatterplots of selected flow variables in a scatterplot triangle. The O-D Attribute Variable Plotting Module draws scatterplots of selected social-economic variables that are associated with individual origins and destinations in a scatterplot triangle. The Projection Pursuit Module is designed to reduce the complexity of spatial flow data by producing PP-synthesized variables as substitutes for the original ones. The Dynamic Brushing Module links the flow map, choropleth map, flow variable plot, and O-D attribute variable plot in an animated fashion through dynamic brushing. The integration of the ten modules makes PPFLOW a powerful tool for exploratory analysis of spatial flows.
Organization of the Dissertation

This first chapter has been an introduction to the dissertation research. In the next chapter, the relevant literature is reviewed, and the current status of flow pattern studies in geography, projection pursuit methods, exploratory data analysis, and scientific visualization are reviewed in detail. Chapter III proposes a new methodology for the study of spatial flow patterns. The new methodology is based on the integration of numeric methods such as the projection pursuit method and graphic methods such as scientific visualization, exploratory data analysis and dynamic brushing. The implementation of the new methodology is addressed in Chapter IV, in which system design, user interface, hardware and software environment, data organization, and other implementation issues are discussed. Chapter V demonstrates some practical results achieved by applying the prototype system PPFLOW to a few geographic problems. This research has led to several conclusions which are addressed in the last chapter. Appendices A, B and C describe the format of the files used by PPFLOW. Appendix D describes the usage of the PPFLOW. The details on how to derive the first partial derivative of the moment-based PP index are addressed in Appendix E.
CHAPTER II
LITERATURE REVIEW

As indicated in the introduction, the objective of this research is to reveal spatial flow patterns, which may lead to new hypotheses. To achieve this goal, this research proposes a new methodology, which is based on a number of areas: traditional cartographic methods and complexity reduction methods as applied to spatial flow analysis, projection pursuit methods, exploratory data analysis, scientific visualization, and dynamic graphics. This chapter provides a selective review of each of these areas, and draws them together within the perspective of spatial flow analysis.

Cartographic Representation of Spatial Flows

To associate spatial flows with geographic space, geographers have devised various ways of representing spatial flows graphically. Traditionally, the mapping of spatial flow data was done manually, and the variety of flow representations was limited. Early flow maps were often simplistic and the data volumes involved were relatively small. In the recent decades, computer-based flow mapping has become available and some new graphic techniques for representing spatial flows graphically have been invented.

Graphic representation of flows can be traced back for more than a century. For example, a famous flow map (Tufte 1983), drawn in 1861, shows Napoleon's invasion to and
withdrawal from Russia in an effective fashion. Lines are used to represent the paths over which Napoleon's army passed, and the width of the lines shows the number of surviving soldiers. The two flows (to and from Moscow) are distinguished by different shading. No explicit direction of flows is given, but one can easily recognize that the flow with a sharply decreasing number of live soldiers shows the withdrawal from Moscow. However, only a single origin-destination (O-D) pair is involved in this nicely designed map, and its data volume is not at all comparable to the complex flow data sets that are the subject of this dissertation research. Another early example is the "Currents of Migration" map, drawn by Ravenstein in 1885 (Tobler 1987, 1994). This map shows that migration occurred mainly between neighboring areas in the United Kingdom. The direction of migration flow is clearly shown on the map by using lines with an arrow head at the end; yet, no flow magnitude data is given.

As time moved on, geographers began to create more complicated flow maps. One of these maps is the Railway Traffic Map (Figure 8), originally included in Ullman's American Commodity Flow (1957). This map shows all railway traffic flows larger than 1,000,000 net tons per mile per year in the forty-eight contiguous states of the United States and the southern portion of Canada. Lines with varying widths are used to represent railway routes and railway traffic. The map clearly shows the high concentration of railway traffic in the Middle Atlantic and Midwest, but does not include any directional traffic information. The representation of directional flows, which is a problem to be addressed in this research, is considerably more difficult.
Figure 8. Railway Traffic. Lines with varying width are used to represent railway traffic. [Source: Ullman 1957]
Another map in Ullman's book shows barge and raft traffic in 1949 along major the rivers in the United States (Figure 9). In this case, directional aspects of the flows are represented by using two parallel bands with each band showing traffic in one direction. The direction of the flow is shown by a separate, arrow-headed, straight line. The map clearly shows that north-bound traffic is much larger than south-bound traffic. Since there are only a few major rivers drawn on the map, the map is easily interpreted. However, this method of showing direction would not be appropriate for a data set with a much larger number of rivers.

*Computer Mapping of Spatial Flows*

Along with the development of computer technology, computer flow mapping began to appear starting in the late 1960's. Among others, Kern and Rushton (1969) developed a computer program called MAPIT to produce flow maps, dot maps and graduated symbol maps. MAPIT, written in Fortran 63 for the CDC 3600 computer, did not display output on a computer screen, but drawn on a Calcomp plotter (Model 563). Of interest to this research is a map that shows the beauty-care center patronage in Christchurch, New Zealand. Lines are used to represented spatial interaction by connecting the locations of customers to the locations of beauty-care centers. Because of its appearance, this kind of map is often referred to as a "spider map." It clearly shows that most customers went to the largest center downtown. No other flow information except connections is displayed on the map.

Wittick (1976) developed a computer system called FLOW for mapping and analyzing transportation networks. FLOW, written in Fortran, was able to draw flow maps on either a digital plotter or a graphics terminal, which allowed the researcher to examine
Figure 9. Barge and Raft Traffic. Lines with varying width are used to represent water routes and traffic volumes. Two-way traffic is shown on the map. [Source: Ullman 1957]
the map before its final output to a digital plotter. He implemented two approaches to the quantitative portrayal of spatial flows. One draws constant width corridors connecting O-D pairs, and then fills the corridors with shading gradients to represent flow magnitude. In Figure 3, which shows air traffic flows, the flow magnitude is classified into four intervals, each of which is associated with a distinct shading density. Although the requirement of grouping flows to several intervals is considered a drawback compared to the proportional line width method, Wittick argues that the corridor shading method has more flexibility in displaying all flows and their values. His second approach to quantitative portrayal of spatial flows uses lines to connect O-D pairs, and displays exact flow volumes along the lines (Figure 4). Although this method provides accurate flow information, its visual effect is not as good as his first method. In addition, both methods fail to depict the directional component of the flows.

Tobler (1987) posed the question "should the width of the flow band be proportional to the magnitude of the movement?" (p. 158), and listed the following alternative flow representations:

1) make all flow bands the same width, and use a variable intensity shading to represent the magnitude;
2) make the shaded area proportional to the magnitude;
3) make the product of the shading density and the shaded area proportional to the magnitude.
By comparing these three alternatives with the original proportional width method, Tobler reported his impression that "widths proportional to flow magnitudes are interpreted more correctly" (p. 158).

In the same study, Tobler also proposed three methods (Figure 10) for the representation of bi-directional flows, which was not addressed in the foregoing literature. He used:

1) two arrows, with the smaller one on the top of the larger one, with the width of each arrow proportional to the respective movement magnitude;

2) half-barbed, separated arrows, with the width of each arrow proportional to the respective movement magnitude;

3) half-barbed, abutting arrows, with the width of each proportional to the respective movement magnitude.

But, Tobler felt that none of the three methods worked very well (p. 158). Hence, the representation of simultaneous bi-directional flows along a single path remains an unsolved problem.

Figure 10. Three representations of bi-directional flows. [Source: Tobler 1987]
Graphic Representation of Multiple Flow Variables

Graphic representation of aspatial multi-variate data is not new. Weathervane symbols have long been used to represent wind speed, wind direction, etc. in climatology. Chernoff (1973) used face features, such as length of nose, curvature of mouth, and curvature of eyes, to represent the magnitude of multiple variables. In addition, Anderson's glyphs and Kleiner-Hartigan trees can also be used to graphically represent multiple variables (Tukey 1981, 1982a). Figure 11 shows a collection of these symbols.

Figure 11. Graphic representations of multi-variate data. [Source: Tukey, in Cleveland 1988, p. 264]
The symbols described above are well-suited for representing multi-variate data that is associated with point or area entities. It is fairly easy to put one such symbol at each point or at the center of an area. The focus of this research, however, is spatial flows, which are linear entities. Can these symbols be used to represent multiple flow variables associated with a link connecting an O-D pair? It is possible to put one such symbol at the midpoint of a link. This works well when there are only a few well-separated links on a map, but when the number of links becomes large or the links are cluttered, a symbol may be crossed by many links, making it difficult to identify which link a symbol is attached to. The combination of point symbols and linear entities will not work for the complex flow data set used in this research.

Graphic representation of multiple spatial flow variables is seldom addressed in the body of literature. The only reference that has been located is an article by Noguchi and Schneider (1977). This work presents an interesting transit service map (Figure 12) that is produced by a computer program called CENVUE(S). The map shows a perspective view of three bus lines with the speed limit shown on each road segment along the three bus lines, and the three bus load variables given at each bus stop. The bus lines are represented by a number of connected blocks, each representing a road segment with the height of the block drawn proportional to the speed limit. Three bus load variables, including peak hour bus load, average bus load and night bus load, are displayed at each bus stop by using three blocks, the height of which is proportional to the number of passengers.
Figure 12. A perspective-view map of a transit service. Speed limits of bus lines are represented by the height of blocks. Peak hour bus load, average day bus load, and night bus load are displayed at each bus stop. [Source: Noguchi and Schneider 1977]

Although this method of representing multi-variate flow data produces satisfactory results in this specific case, it creates a cluttered, incomprehensible map for a large transit service network which may include hundreds of bus lines. Also, only the multiple flow variables that are associated with nodes (bus stops) are displayed on the map. The focus of the present research is the representation of multiple flow variables associated with links.
routes), which is a much more difficult problem. In fact, the transit service map displays only one route-associated variable, the speed limit.

**Summary**

The foregoing discussion has focused on the graphic representation of spatial flows. There are three widely used methods, i.e., bands with width proportional to flow magnitude, constant width corridors with shading-gradient representing flow magnitude, and lines with numbers, for representing a single flow variable. The first two methods produce a better visual effect than the third, which is intended for detailed quantitative analysis. Based on the impression of Tobler (1987), the first method produces maps that are interpreted more correctly.

Not all the problems in the graphic representation of spatial flows have been solved. For example, the graphic representation of bi-directional flows needs to be improved, and an effective representation of multiple flow variables has not yet been discovered.

More fundamentally, the objective of this research is to detect spatial flow patterns and to generate new hypotheses. Graphic representations alone can not reach this objective. It is easy to see that when the volume of flow data displayed on a flow map is large, the map will be cluttered no matter what representation method is used. This leads to the issue of reducing the complexity of a flow data set, which is the topic of the next section.

**Complexity Reduction Methods as Applied to Flow Studies**

In addition to the development of effective graphic displays of spatial flows, Tobler (1987) also conducted some experiments directed toward reducing the complexity of the O-D matrix. In his migration example, he suggested a number of ways of reducing the number
of migration flows which need to be shown. First, instead of showing all the migration flows in an O-D matrix, only those flows from or to a single place are shown. Secondly, migration flows that are under a specific threshold can be deleted. The problem is determining the size of the threshold. Since the entries of an O-D matrix generally follow a Pareto distribution (i.e., there are many small ones, but only a relatively few large ones), a good deletion strategy is to remove migration flows that are below the average flow magnitude in the O-D matrix (Tobler 1987, p. 160). A third strategy he suggested is to the collapse an O-D matrix by combining geographically adjacent places; thus reducing the number of entries in the O-D matrix. Of course, data aggregation is always accompanied by a loss of resolution.

These methods are effective for a single O-D matrix, but are unable to handle the multiple O-D matrices used to represent multiple flow variables. It is the intention of the present research to develop better methods for reducing the complexity of multiple O-D matrices.

To reduce the complexity of a multi-variate data set, geographers usually use principal component analysis (PCA) and factor analysis (FA) to reduce the number of variables. These two methods have been used to study spatial flows. For example, Marble and Garrison (Marble, 1993) were among the first researchers to apply FA and PCA to flow data analysis. In the 1960s, they studied airline traffic data to reveal the structure of airline networks and the relationship between this spatial structure and national economic development. Two types of matrices were generated in their study. In the first type, the cells in the matrix are filled with 0/1 entries, where 0 represents non-connectivity and 1 represents connectivity between an O-D pair. The R and Q mode principal components of the matrix
were examined as possible indices of the structure of the national airline network. They applied these techniques to the airline networks of several countries, generating a set of principal components for each country. Then they used linear regression to examine the relationship between economic development and these components, and found significant correlations. In the second type of matrix, the cell values represented flows in terms of flights/week. Marble and Garrison repeated the original process with this matrix, and found similar results. They concluded that the structured complexity of each country's airline network was coarsely related to its level of economic development.

A slightly different method that has been used to reduce the complexity is dyadic factor analysis (Berry 1966, 1968, 1971; King 1969; Black 1973). In a dyadic flow matrix, rows are O-D pairs and the columns represent different flows. Multiple flow variables can be included in a single dyadic flow matrix. Factor analysis can be used to extract major flow factors from the multiple flow variables in the matrix. Black (1973) applied this methodology to a data set which contained 24 interregional commodity flows between 9 aggregated census regions for the United States. He extracted five factors, each representing a combination of original commodity flows. For example, factor I was mostly related to the major flows of transportation equipment. These factors are then mapped into a geographic space. Figure 5 shows that the flows of transportation equipment are mainly from the Middle Atlantic (New York is in this region) to other regions. Black identified only the directions of certain groups of commodity flows among the nine census regions, but the magnitude of these flows was still unknown.
Is PCA/FA the best complexity reduction method for the purpose of the present research? PCA transforms originally correlated variables to new components, which are orthogonal to each other. The components with large variance can be used in future analysis such as cluster analysis. In this research, what is desired is the informative projections (a projection is a direction onto which multi-dimensional data are mapped) in the dimensions of multiple flow variables. Are the principal components the most informative projections? This question will be addressed later in this chapter and also in the next chapter. This research focuses on exploring spatial flow patterns through visualizing informative projections. Any informative projections are desired and they do not have to be orthogonal. It can be argued that projections produced without the orthogonal constraint may be more informative. Hence, PCA/FA may not be the best complexity reduction method for the particular purpose of this research. A search was made for better alternatives.

The Projection Pursuit Approach

*What is Projection Pursuit (PP)?*

The term "projection pursuit" was first introduced by Friedman and Tukey (1974). The first projection pursuit (PP) method was implemented to reveal clusters in a multi-dimensional data set. But in general, PP is used to "name a technique for the exploratory analysis of reasonably large and reasonably multivariate data sets" (Jones and Sibson 1987), and it "reveals structure in the original data by offering selected low-dimensional orthogonal projections of it for inspection" (Jones and Sibson 1987). Eslava-Gomez (1989) had slightly different description of projection pursuit:
Projection pursuit is a technique specifically directed at finding non-linear structure through linear projections from a p-dimensional to q-dimensional space (p > q and in practice q=1,2). (p.26)

**PP versus PCA**

A brief review of principal component analysis will help us to better understand PP. PCA involves an orthogonal transformation of the original variables (some of which are presumably correlated) into a set of new uncorrelated components. The components with large variance (a measure of the dispersion of data values around the mean) are normally selected for follow-up analyses such as cluster analysis and regression.

From the above description, it can be noted that PP is similar to principal component analysis in that both of them are used to reduce the dimensionality of data sets. In PP, as in PCA, each of the selected projections can be treated as a new variable, and can be expressed as a linear combination of the original variables.

Yet, PP and PCA are significantly different. PP produces projections from multi-dimensional data for the purpose of visual analysis in an exploratory data analysis environment. Visual inspection of the projection reveals informative structures such as clusters, outliers. PCA derives components for follow-up statistical analysis such as clustering and regression. Cluster analysis of PCA components may yield clusters, but not other structures such as outliers and relationships. This research focuses upon exploration of spatial flows, aiming to uncover possible flow structures, including clusters, outliers and relationships by visualization. Hence, what we need are the projections (or components) that reveal most structures. Now, the question is which dimension reduction method is capable of revealing more structures for visual analysis: PP or PCA? The difference between PP and
PCA in detecting structures such as clusters and outliers has been studied by many scientists, e.g., Friedman and Tukey (1974), Jones and Sibson (1987), and Eslava-Gomez (1989), just to name a few. The ineffectiveness of PCA for this task is clearly pointed out by Eslava-Gomez (1989), who states:

Principal components have been used to detect grouped data. However they are not designed to do so, and are rather ineffective at detecting non-linear structure such as clusters. (p. 12)

Friedman and Tukey (1974) used both a PP method and PCA to study a 14-dimensional simplex data set, containing 975 data points, which occur in a number of clusters. The PP method clearly shows two clusters (Figure 13 b), while no clustering is shown in the largest principal component (Figure 13 a). The difference is also reflected in the value of the PP index (P-index in Figure 13). The PP index in the direction of the selected projection (Figure 13 b) is $3.37 \times 10^5$, which is much larger than the $1.00 \times 10^5$ in the direction of the largest principal component. Their results suggest that PP is a better method for visual analysis which is the focus of this research.

**Development History of PP**

Aiming to solve the problem of mapping multivariate data onto lower-dimensional manifolds, Kruskal (1969, 1972) first developed the initial idea of projection pursuit (PP), although he did not use the term "projection pursuit." Kruskal defined an "index of condensation" to guide the linear transformation of a multidimensional data set onto lower dimensions, hoping to reveal clustering. The term "projection pursuit" was first introduced by Friedman and Tukey (1974), who, along with other people, implemented PP in a
computer system called PRIM-9 (Fisher, Friedman and Tukey 1974), which was capable of reducing up to nine dimensions to a one or two-dimensional subspace.

Later, important contributions to PP were made by a number of statisticians. Friedman and Stuetzle (1981) introduced projection pursuit regression, a general nonparametric multiple regression method, in which the regression surface is "a sum of general smooth functions of linear combinations of the predictor variables in an iterative manner" (Friedman and Stuetzle 1981). Huber (1985) formalized the PP concept, which then
could be proved theoretically. In Huber's view, the interesting structures of a multidimensional data set are presented by the projections in which the distribution of the projected data represents a significant departure from the normal distribution. Jones and Sibson (1987) introduced two projection pursuit indices. One is defined as Shannon entropy, which requires heavy computation. The other is a simplified version of the first, based on the third and fourth order moments of the projected data, and is very efficient in terms of computation.

Some of the major studies involving PP are reviewed in more detail below.

**Kruskal's Work**

The goal of Kruskal's work (1969, 1972) was to find a linear transformation that mapped a multidimensional space onto a lower dimensional manifold with the hope of uncovering hidden structures. Based on the coefficient of variation of the inter-point distance, Kruskal (1969) first defined the index of condensation as:

$$ C = \frac{s}{m} $$

(1)

where $s$ and $m$ are the standard deviation and the mean of the inter-point distances. A large $C$ value tends to indicate condensation; this is clear in Kruskal's original explanation:

To see why this makes some kind of sense, imagine several rather loosely grouped clusters, and then consider what happens as every point in each cluster moves toward the center of its own cluster. The within-cluster distances all get smaller. Some between-cluster distances get smaller and some larger, but they seem to share no systematic tendency. Hence $s$ gets larger and $m$ gets smaller, both of which contribute to $C$ getting larger. This leads to the notion that a large value of $C$ tends to indicate condensation. (Kruskal 1969)
**Friedman and Tukey's Work**

Following Kruskal (1969, 1972), Friedman and Tukey (1974) introduced the term projection pursuit (PP) to name a dimensionality reduction method that is attained through linear transformation. The PP index developed by Friedman and Tukey is a combination of local density and global spread of the projected data:

$$I(a) = s(a) \cdot d(a)$$

(2)

where $a$ represents a projection, $s(a)$ measures the spread of the data, and $d(a)$ describes the "local density" of the points after projection onto $a$. Friedman and Tukey defined $s(a)$ as the trimmed standard deviation of the data from the mean as projected onto $a$. Because the objective is to reveal clusters, outliers are removed in the calculation of $s(a)$. Thus, $s(a)$ is robust against extreme outliers. The local density $d(a)$ is expressed by using an average nearness function. The locality is defined by a user-specified cutoff radius. Only the points within the range enclosed by the cutoff radius are used to measure local density. They compared PP and PCA, and demonstrated that PP is more effective than PCA in detecting clusters. Although Friedman and Tukey's PP method is very effective in uncovering clustering structures, it fails to detect outliers, which are also of interest in studying spatial flows.

**Huber's Work**

Huber's work marks a departure from Friedman and Tukey's in the definition of interesting structures hidden in multi-dimensional data. Friedman and Tukey (1974) defined an interesting structure as one in which projected data have a large local density and a wide spread (refer to the definitions following Equation (2) ) along a projection, thus revealing...
point clusters. Huber views interesting projections as those that have a large departure from
the normal distribution. By maximizing a PP index of this type, any non-normal structures
will be uncovered. The structures can be point clouds, concentrations along lines or curves,
or other irregular structures. Based on such a philosophy, Huber introduced various PP
indices, which include standardized absolute cumulants, standardized Fisher information,
and standardized negative Shannon entropy.

Jones and Sibson's Work

Their PP indices are measured by negative Shannon entropy, as suggested by Huber
(1985). The major contribution of Jones and Sibson (1987) lies in the techniques they
developed for a computationally efficient implementation. The heavy computation
requirement of the Shannon entropy index is significantly reduced by deriving a heuristic,
alternative version of the index. The new index uses the third and fourth order moments of
the projected data, and is defined as

\[(k_3^2 + k_4^2)/12\]  (3)

where \(k_3\) is the third cumulant, and \(k_4\) is the fourth order cumulant. \(k_3 = \mu_3, k_4 = \mu_4 - 3\). \(\mu_3\) and
\(\mu_4\) are the third and fourth moments. This moment index has a value of zero for the normal
distribution (Jones and Sibson 1987). Interesting projections are those for which the PP
index is maximized.

Evaluation of PP Indices

A projection pursuit (PP) index can be evaluated in two aspects: its effectiveness in
locating informative projections and its computational efficiency. Friedman and Tukey's PP
index (Friedman and Tukey 1974) is good for detecting point clusters, while the indices
suggested by Huber (1985), and Jones and Sibson (1987) can reveal any structures that are different from the normal distribution. These structures include point clusters, outliers, concentrations along a line, a curve, a circle, or any geometric shape. Apparently, the latter is better for spatial flow data analysis because we are interested in not only point clusters, but also in any possible non-normal structures. For example, outliers in flow data usually represent large flows, and clusters group flows that are similar.

In addition to choosing a proper PP index, optimization is another difficult problem. The optimization of a PP index is very complicated, and requires heavy computation. For convenience of notation, N is used to denote the number of observations, and K to denote the number of dimensions (or variables, parameters). The problem of PP is to maximize a PP index by choosing a K-vector $\alpha$ with $|\alpha|=1$ (a non-linear constraint). The computational complexity of optimizing Jones and Sibson's moment index is proportional to N (Jones and Sibson 1987), while that of Friedman and Tukey's PP index is proportional to $N\log(N)$ (Friedman and Tukey 1974). The former is more efficient.

Because PP methods are better than PCA in uncovering non-linear structures such as clusters and outliers (Friedman and Tukey 1974; Jones and Sibson 1987; Eslava-Gomez 1989), they are expected to be very effective in reducing the complexity of a spatial flow data set containing multiple flow variables. Based upon the computational efficiency and the capability of revealing non-linear structures such as clusters and outliers, Jones and Sibson's moment PP index appears to be a better candidate for spatial flow data analysis.
Applications of PP

PP methods have been applied to data sets in statistics, chemistry, physics, etc. Many PP application examples can be found in the papers written by Friedman and Tukey (1974), Friedman, et al. (1980), Friedman and Stuetzle (1981), Diaconis (1984), Li (1985), Jee (1985), Jones and Sibson (1987), Eslava-Gomez (1987) and Chen (1989), but most of the examples are designed to test the PP methods developed in the papers.

In addition, PP methods have been applied to the display of multi-dimensional medical image data, such as those produced by MRI or multimodality image fusion (Harikumar and Bresler 1992; Boesel and Bresler 1992). Instead of directly mapping multiple variables, they first projected the data onto a single dimension (which represents a new variable) and then mapped this new variable. They concluded that PP is effective in revealing non-linear clusters in multi-parameter medical data.

PP methods had not been used in geography until 1994 when Liu (1994) presented a paper on using PP methods to detect spatial flow patterns in a multi-dimensional spatial flow data set. The capability of PP in uncovering non-linear structures such as clusters and outliers in a flow data set was demonstrated in the presentation.

Exploratory Data Analysis, Dynamic Graphics and Scientific Visualization

As pointed out in Chapter I, exploratory data analysis, dynamic graphics and scientific visualization can be used to improve the study of spatial flows. The following sections provide a brief review of the body of literature in these two areas and their relation to geographic studies.
**Exploratory Data Analysis**

Exploratory data analysis (EDA) was first introduced by Tukey in his book *Exploratory Data Analysis* (1977a). In this book, Tukey described EDA as a "detective work," and he used the following example to explain the nature of EDA:

A detective investigating a crime needs both tools and understanding. If he has no fingerprint powder, he will fail to find fingerprints on most surfaces. If he does not understand where the criminal is likely to have put his fingers, he will not look in the right places. Equally, the analyst of data needs both tools and understanding. (p. 1)

EDA provides a kit of tools to investigate a data set and try to understand what it means. The tools described in Tukey's original book concentrate on "simple arithmetic and easy-to-draw pictures" (p. v).

EDA is completely different from confirmatory data analysis (CDA). Unlike CDA, which rests upon a series of strict assumptions such as random sampling and tightly specified distributions (usually normal distributions), EDA develops a comprehensible graphic picture of the data so that one can objectively analyze data without any assumptions. EDA emphasizes exploration and hypothesis generation, while CDA focuses on hypothesis testing and confirmation, such as model fitting and significance testing. To emphasize the importance of EDA, Tukey argued that "unless exploratory data analysis uncovers indications, usually quantitative ones, there is likely to be nothing for confirmatory data analysis to consider" (p. 3).

Today, the techniques of EDA have been extended far beyond the initial "simple arithmetic and easy-to-draw pictures." Following Tukey, Wang (1978), Barnett (1981), Becker and Chambers (1984), and Cleveland and McGill (1988) contributed to the
development of EDA techniques. Comprehensibility of graphic data representation and intuitive graphic analysis form the basis of EDA, as Barnett (1981) noted that David Andrews (1978) remarked:

EDA is the manipulation, summarization, and display of data to make them more comprehensible to human minds, thus uncovering underlying structure in the data and detecting important departures from that structure.

EDA techniques have been implemented in a number of computer systems, such as PRIM-9 (Fisherkeller, Friedman and Tukey 1974), ORION I (McDonald 1983) and MACSPIN (Donoho, et al. 1988). PRIM-9, implemented on an IBM 360/91 (a large mainframe computer), is an interactive data display and analysis system for multidimensional (up to nine dimensions) data through a combination of projection and rotation. It renders a three-dimensional picture based on the use of any three selected variables to define the subspace. ORION I and MACSPIN are the descendants of PRIM-9. ORION I was built on a graphics workstation with an integrated high resolution display, while MACSPIN was based on the Macintosh. In addition to the difference in platforms, both systems have more functions than PRIM-9. For instance, MACSPIN provides an animation function, which permits the user to study a fourth variable, e.g. time, on a three-dimensional display. These systems have been reviewed in great detail by Sandhu (1990), Tang (1993), and Gou (1993).

EDA is capable of playing an important role in the study of spatial flows. Without the exploration of flow patterns, which then leads to new hypotheses about spatial flows, there will be only existing theories to confirm.
Dynamic Graphics

An influential technique that has evolved from EDA is dynamic graphics (Becker, et al. 1987a, 1987b). Dynamic graphics has two important characteristics: direct manipulation and instantaneous change (Becker, et al. 1987b, p. 355). Direct manipulation means that the user can interactively manipulate data that are graphically displayed on a computer graphics display. Graphic elements of the data can be directly manipulated by pointing devices such as a mouse or track ball. Instantaneous change means that the graphic elements change instantaneously once they are manipulated. In other words, dynamic graphics refers to an interactive, real-time computer graphics environment within which the researcher can easily explore the properties of a data set.

In their dynamic graphics system, Becker, et al. (1978a, 1978b) implemented a set of techniques, including identification, deletion, scaling, rotation, dynamic parameter control and brushing. Each of these techniques is discussed briefly below.

Identification can be done in two different ways. One way is to find out the label associated with a graphic element by selecting that element on the computer screen. The other way is to find out which graphic element is associated with a given label.

One simple use of deletion is to delete an outlier in a scatterplot so that the researcher can better examine the relationship between the remaining points. Deletion of outliers should be done with caution since outliers may carry important information pertaining to a data set.

Scaling is necessary in many statistical graphics. For example, data points in a scatterplot are usually scaled so that they can fit into the plot. Also, scaling can be used to
control the ratio of the length of the Y axis over length of the X axis in a scatterplot. Changing the ratio can alter the slope in the plot; therefore affecting our ability to comprehend the information in the plot. According to Cleveland and McGill (1987), the accuracy of judging slope decreases when slope departs away from \(-45^\circ\) or \(+45^\circ\).

The technique of rotation is usually used to rotate a three-dimensional graph so that the research can examine structures in the three variables. Reaven and Miller (1979) used this technique to study chemical diabetes in PRIM-9. The data arising out of the observations on 145 adults including both diabetic patients and normal people, were displayed using PRIM-9 (Fisher, Friedman and Tukey 1974). The three-dimensional

![Three-dimensional diagram of diabetes data in PRIM-9](source: Reaven and Miller 1979)
image has a boomerang-like structure with two wings and a fat middle (Figure 14). The plump in the middle roughly corresponds to normal people, the left wing to overt diabetics, and the right wing to chemical diabetic patients. Without rotation, this structure cannot be easily identified.

Dynamic parameter control can be used to control the content of a graph. For example, the content of a population density map can be dynamically adjusted by controlling the low and high limits of population density.

Brushing was developed by Becker, et al. (1987a, 1987b) to link multiple views in a scatterplot matrix, which contains multiple scatterplots arranged in a matrix form. A brush is actually a selection box, which can be a square, a rectangle, a circle, or any other shape. When a brush covers a subset of points in a scatterplot, the corresponding points will be highlighted in other scatterplots. This is a very effective tool for revealing relationships between multiple variables. Figure 15 shows a scatterplot matrix that consists of six scatterplots of a data set containing three variables: tensile strength, hardness and abrasion loss (the amount of rubber rubbed off each specimen by an abrasive material) of 30 rubber specimens (Becker, et al. 1987b). When a rectangular brush covers the points with low hardness values in panel (3,1), the corresponding points are highlighted in all other five scatterplots. The highlighted points in panels (3,2) and (2,3) show a linear relationship between tensile strength and abrasion loss.

Monmonier (1989a) first introduced EDA concepts into geography. He discussed the technique of "geographic brushing," which can be used to enhance exploratory analysis of the commonly used scatterplot in geography. Buja, et al. (1991) used focusing and linking,
Figure 15. Brushing a scatterplot matrix. The brush in the (3, 1) panel (row 3, column 1) is highlighting points with low values of hardness. The (3, 2) panel (row 3, column 2) shows the dependence of abrasion loss on tensile strength for hardness held fixed to low values. [Source: Becker, et al. 1987b, p. 365]

which is similar to brushing, to link a scatterplot to a geographic map. For example, by linking a scatterplot of housing value and climate to a spatial distribution map of metropolitan areas in the United States, they easily found that metropolitan areas with mild climate and expensive housing are concentrated along the west coast.
The above techniques are very useful EDA tools for exploring relationships and structures in a multi-dimensional data set. In practice, EDA and dynamic graphics are usually implemented together with the techniques of scientific visualization, which is the topic of the next section.

Scientific Visualization

The rapid development of computer graphics has created a number of new approaches for scientific research. One of these is scientific visualization, a set of tools that permits researchers to explore complex data sets interactively within environments that enhance and enrich the researchers' own intuition.

What is Scientific Visualization (SV)?

McCormick, et al. (1987) offered a computational view of scientific visualization (SV), and defined it as "a method of computing...(that) transforms the symbolic into geometric, enabling researchers to observe their simulations and computations." (p. 3) According to this view, SV generates comprehensible images from complex multi-dimensional data, thus helping researchers to interpret the data. MacEachren, et al. (1992) defined SV as "an act of cognition." Buttenfield and MacKaness (1991) introduced a more comprehensive definition which combines the computational view and cognitive view. MacEachren, et al.'s cognitive view is preferred by the author because cognition is close to the intuitive meaning of visualization. The computational view is somewhat confusing. If SV is "a method of computing," then how can we distinguish SV from scientific computation? For the same reason, the holistic definition by Buttenfield and MacKaness departs from what is intuitively felt about SV.
SV is often used in three different ways: tracking, steering, and postprocessing (Marshall, et al. 1990). Tracking means the real-time display of computer generated images. The user can monitor the computed results, but cannot interactively manipulate both computation and image display. This is sometimes referred as a "passive" form of real-time display. On the contrary, steering allows the user to interactively modify the computations and, hence, the graphic representation. This "active" form of real-time visualization requires intensive computation, which is often beyond the capacity of ordinary computers. For this reason, supercomputers are used in a steering visualization system. Postprocessing means displaying resultant data after computation is completed, which enables researchers to study their results in more detail.

This dissertation research loosely fits into the steering type of SV because the suggested prototype system incorporates capabilities for real-time display of graphics generated from a spatial flow data set and interactive manipulation of graphics and computation, i.e., setting a different initial direction may produce different informative projections in PP optimization.

**SV versus GIS and Cartography**

Although the concept of visualization has long been used in conventional cartography, it is different, in its nature, from current scientific visualization. The former deals with how to graphically represent what is known in an effective way; while the latter concerns how to reveal unknown structure or pattern through visual representation and visual analysis. As pointed out by Fisher, et al. (1993), the emphasis of SV is on "the development of ideas, not, as in traditional cartography, the presentation of an idea or view"
A number of geographers, mostly cartographers and GIS practitioners (Buttenfield, et al. 1990, 1991b; Monmonier 1989b, 1990, 1991, 1992a, 1992b; MacDougall 1992, 1993; MacEachren, et al. 1990, 1992a, 1992b; DiBiase, et al. 1992; Dorling 1991, 1992a, 1992b), have begun to introduce scientific visualization into geography. Many of their articles discussed the role of visualization in cartography and GIS and the potential of visualization in geography. For example, MacEachren and Monmonier (1992) propose the concept of geographic visualization, and argue that "if developed to its full potential, geographic visualization might not only promote richer and more thorough visual thinking but might also provide broader cartographic empowerment to groups and individuals addressing geographers." (p. 199)

**EDA, Dynamic Graphics and SV as Applied to Geography**

Dynamic visualization of geographic data is not new. Cartographic animation techniques have been used to demonstrate change in geographic phenomenon over time. Tobler (1970) used a computer to generate an animated, cartographic film to portray the growth of population in Detroit. Moellering (1973) created an animated, cartographic film to study the patterns of traffic crashes. But the integration of EDA, Dynamic graphics and SV for solving geographic problems started only a few years ago.

One of the early applications of EDA, dynamic graphics and SV to geography is the Planetary Data Visualization System (PDVS), which is a concept created by Duane Marble and implemented by Sandhu (1990) and Wentz (1990). PDVS aimed toward the visualization and analysis of very large global space-time data by using a dynamic, three dimensional, graphic representation. Although their implementation did not fully realize the
idea of PDVS, it offered a better tool for the study of space-time patterns based on a global earthquake data set.

Another early example is SPIDER (Spatial Interactive Data Explorer) (Haslett, et al. 1990, 1991). The idea behind SPIDER was to link the statistical graphics of EDA and the spatial map view of geographic data. SPIDER includes the following major features:

1) a multi-window interface;
2) the ability to overlay multiple spatial views of the data;
3) the control of the overlays through a graphic interface;
4) point, line, and region objects in 2-dimensional space;
5) graphic statistical displays;
6) dynamic linking of statistical and geographic views;
7) possibilities for generating (spatial) models based on views;
8) one-to-many linking (Haslett, et al. 1990).

Other systems include Polygon Explorer (MacDougall 1992, 1993), EXPLOREMAP (Egbert and Slocum 1992), and ViSDA/A (Tang 1993). All these systems have the properties of direct manipulation and instantaneous change, which allow the researcher to interact with the graphic elements on the computer screen in near real time. Geographic map and statistical plots of attribute data are dynamically linked; thus facilitating the exploration of geographic patterns in a data set.
Combining EDA, Dynamic Graphics, SV and Projection Pursuit Methods for Spatial Flow Pattern Visualization and Exploration

In the area of spatial flow visualization, some interesting studies were carried out and documented by Becker, et al. (1990a, 1990b, 1991a, 1991b). They developed two dynamic graphic techniques for flow visualization: dynamic link display and dynamic node display. In their dynamic link display, the length of link, and the lower and higher limits of flow on the link can be adjusted. Once an adjustment is made, the link flow map is changed instantly. This interactive control of link attributes allowed researchers to focus upon interesting data by discarding the others. Dynamic node display works in a similar way, the attributes of each node can be controlled and modified by the user in an interactive environment. Combination of the dynamic link display and node display offers an effective

Figure 16. Airline delays. This map shows hypothetical state to state airline delay data on a day where bad weather conditions are affecting most of the United States. There are so many line segments on the map that it is difficult to determine anything from the display. [source: Becker, et al. 1990a]
tool for the study of spatial flows. These techniques are applied to the study of overflowing (non-completed due to congestion) calls in the AT&T network following the October 17, 1989 earthquake in California (Becker, et al. 1991a) and hypothetical state to state airline traffic delay data in the United States (Becker, et al. 1990a). The raw data on hypothetical airline traffic delays are shown in Figure 16, from which one can hardly see anything because the figure is obscured by the mess of black lines. By using the dynamic link display technique, the most serious airline delays (those greater than 100 minutes) are extracted and shown in Figure 17, which clearly shows that serious delays are linked to cities in the Northeast. The node dynamic display (Figure 18) shows a similar pattern.

![Figure 17. The most serious airline delays (using dynamic link display technique). This map shows the same state to state airline data as in Figure 16, but the slider on the lower scale has restricted the map to showing only the routes on which delays are greater than 100 minutes. This map makes it obvious that the primary delays on this day are in the Northeast, perhaps as a result of a snowstorm. [source: Becker, et al. 1990a]](image-url)
The above studies suggest that dynamic graphics and SV can contribute to the improvement of spatial flow analysis. However, the proposed techniques are too primitive, and cannot fulfill the requirement of detecting spatial flow patterns posed by this research. One of the major problems is that their flow representation is weak. Lines are used to connect origins and destinations, and neither flow direction nor flow magnitude are presented on the flow map.


Figure 18. Node map (using dynamic node display technique). The rectangles encode the average air traffic delays at each location. The horizontal extent of the rectangle represents delays for inbound traffic, the vertical extent for outbound traffic. This map shows one particular point in time, hour 14:30. The dynamic node map tool displays maps like this in a rapidly varying time sequence, much like a movie, to allow the analyst to see time trends in the delay patterns.
spatial flows: lines with numbers, bands with width proportional to flow magnitude, and bands with shading gradient representing flow magnitude. In the first method, the magnitude of flow is written in numbers along the flow line. In the second, the magnitude of flow is proportional to the width of flow band. The wider the band, the larger the flow. In the third method, the shading of a flow corridor with constant bandwidth is used to represent flow. Intensity of the shading represents flow magnitude. These methods can be used to represent non-directional flows, and directional flows when an arrow is added at the end, or near the end of each flow line (band). Gou also implemented two-way brushing techniques to link the flow map and scatterplot matrix of O-D attribute data.

Compared to the work of Becker, et al. (1990a, 1990b, 1991a, 1991b), SEFLOW provides much better functionality and flow representation. But, it is still insufficient, especially when a flow data set contains multiple flow variables. The detection of flow patterns in SEFLOW relies on visual analysis of the flow map displayed on computer screen. Only one flow map can be displayed at a time. If a data set has K flow variables, one has to sequentially view K flow variables. Such sequential visualization is not only time-consuming, but it is also difficult to find overall patterns hidden in the K variables.

To improve studies of flow data sets containing multiple flow variables, this dissertation suggests a set of new methods. First, projection pursuit (PP) methods are used to reduce the complexity of flow data sets. A smaller number of new variables (usually one or two) generated by PP methods can substitute for a large number of original flow variables. Then, other methods based on SV, EDA and dynamic graphics are used to explore patterns in the new flow variables.
Summary

As illustrated above, studying spatial patterns in spatial flow data is a substantial challenge to geographers. Statistical methods and tradition cartography alone can not give satisfactory solution to this problem.

Fortunately, geographers are not alone in tackling spatial flows. Other scholars in transportation and telecommunication have also been studying spatial flows. Even though there are many problems in their studies, the ideas and lessons learnt from their research can be used to improve the current situation in spatial flow analysis. In addition, recent developments in projection pursuit (PP) methods, scientific visualization (SV), exploratory data analysis (EDA), and dynamic graphics, has brought us new tools which can be used in the study of spatial flows.
CHAPTER III

COMBINING PROJECTION PURSUIT METHODS, EXPLORATORY DATA ANALYSIS, DYNAMIC GRAPHICS, AND SCIENTIFIC VISUALIZATION FOR SPATIAL FLOW DATA ANALYSIS

The previous chapter reviewed the current status of spatial flow analysis, identified the existing problems and the areas needing improvement, and examined some new developments in exploratory data analysis (EDA), scientific visualization (SV) and dynamic graphics. These new developments provide additional tools that can be extended to the study of spatial flows. In addition, projection pursuit (PP) was introduced as a dimension reduction method that can reduce the complexity of a spatial flow data set. Based on the experience of previous studies and a synthesis of PP, EDA, dynamic graphics and SV, this chapter proposes a set of integrated methods to facilitate the detection of spatial flow patterns, which may lead the researcher to the generation of new hypotheses.

This chapter is organized as follows: the flow patterns used in this research are examined first; then, several new methods for detecting spatial flow patterns are proposed; PP, as a method for reducing the complexity of a spatial flow data set, is discussed next; a method for graphic representation of multiple flow variables on a flow map is suggested; and finally, details of how to detect spatial patterns that are hidden in a spatial flow data set containing multiple flow variables are addressed.
Spatial Flow Patterns in a Regional Flow Data Set Containing Multiple Flow Variables

In a complex flow data set containing multiple flow variables, there exist many possible flow patterns, which are difficult to identify. This section introduces the specific flow patterns that are the focus of this research, namely the spatial distribution of dominant flows, the spatial distribution of zero-flows, the similarity of flows, relationships between flow variables, relationships between flow variables and O-D attribute variables, relationships between flow variables and distance, and the change of flow patterns over time. Each of these is addressed individually below.

Spatial Distribution of Dominant Flows

The first important flow pattern is the spatial distribution of dominant flows. For example, the dominant commodity flow out of Iowa would be the flows of agriculture products. Spatial distribution of these flows reveals major economic ties between Iowa and other regions. In the case of migration, the distribution of dominant flows displays major movements of people. Detection of dominant flows is the first goal of this research.

Spatial Distribution of Zero-flows

In addition to dominant flows, zero-flows are also of interest to geographers. If the flow between a origin and destination is zero, no interaction exists between these two regions. Geographers are often interested in seeking the reasons underlying the lack of interaction between these regions, which may lead to other interesting discoveries. In this research, tools will be developed to facilitate the discovery of patterns associated with zero-flows.
Similarity of Flows

Another interesting flow pattern arises out of the similarity of flows. When only one flow variable is involved, the flows that have almost the same magnitude are considered similar. But, the definition of similarity becomes complicated when multiple flow variables are involved. How to find similar flows within a data set containing multiple flow variables will be discussed later in this chapter.

Relationships between Flow Variables

In a data set containing multiple flow variables, relationships may exist between the flow variables. In the case of interregional migration in the United States, migration flows during the 1960s are likely related to those during the 1970s and 1950s. It would be interesting to know which pair is better correlated: between the 60s and 70s or between the 60s and 50s. Such relationships will be displayed graphically by using tools implemented as part of this research.

Relationships between Flow Variables and O-D Attribute Variables

In addition to the relationships between flow variables, flow variables may also be related to attribute variables associated with the origin and destination regions. Again, in the case of interregional migration in the United States, migration flows may be related to the difference in job opportunities in the origin and destination states. Tools will be developed to reveal such relationships in this research.

Relationships between Flow Variables and Distance

Another relationship may exist between flow variables and distance. Based on the gravity model and other formulations, we see that flows have an inverse relationship with
distance. The larger the distance between the origin and destination, the smaller the magnitude of the flow; this is known as distance decay. This research attempts to develop tools that can be used to demonstrate and examine such relationships.

Change of Flow Patterns Over Time

When time series-based flow data are provided, the change over time in flow patterns can be examined. For example, the dominant migration flows during the 1980s may be different from those twenty years ago. Geographers are interested in what the differences are. Tools to be developed in this research should facilitate the detection of time-based changes in flow patterns.

Combining PP, EDA, Dynamic Graphics and SV for Studies of Regional Flow Data Sets Containing Multiple Flow Variables

This research presents an inventive combination of projection pursuit (PP) methods, exploratory data analysis (EDA), dynamic graphics, and scientific visualization (SV) to reveal the types of spatial flow patterns described in last section, and to aid the researcher in hypothesis generation. Among these flow patterns, dominant flows and similar flows actually represent outliers and clusters in a multi-dimensional space of multiple flow variables. Flows falling into the same clusters are considered similar. Unfortunately, the human eye-brain combination can not directly detect outliers in a space with more than three dimensions. Hence, a logical approach is to reduce the dimensions of the data set so that one can directly visualize outliers in a smaller (no more than three) dimensional space. PP is a dimension reduction method that reveals clusters and outliers. Projections generated by PP are used in the visual detection of outliers and clusters. The reason why PP has been chosen
over principal component analysis (PCA) and factor analysis (FA) as the dimension
reduction method for this research has been discussed in Chapter II and will be further
examined later in this chapter. Other flow patterns pertaining to relationships between
different aspects of a flow data set can be explored in an interactive, graphic environment
based upon a synthesis of EDA, dynamic graphics and SV.

A flow chart for applying the combined approaches to spatial flow data sets is given
in Figure 19. At the top of Figure 19 is a spatial flow data set, which contains three types of
data: multiple flow variables, multiple O-D attribute variables and locational data pertaining
to individual origins and destinations. First, projection pursuit methods are used to reduce
the number of flow variables and O-D attribute variables. Dimension reduction result in a
set of newly generated flow variables and O-D attribute variables, each preserving the
overall property of the original variables. Replacing the original variables by a smaller
number of PP-generated variables makes the process of spatial flow analysis less
complicated.

Second, the aspatial aspects of PP-generated variables and/or original variables are
displayed in statistical plots such as histograms or scatterplots. A researcher can visualize
the outliers and clusters on the statistical plots of flow variables, thus discovering dominant
flows and similarity of flows. Meanwhile, spatial aspects of PP-generated flow variables
and/or original flow variables are displayed on a flow map, and the spatial aspects of PP-
generated O-D attribute variables and/or original O-D attribute variables are displayed on
a choropleth map. At this point, both the aspatial and spatial aspects of the flow data and O-
D attribute data are represented graphically.
Figure 19. Combining PP, EDA, dynamic graphics, and SV for studies of regional flow data sets containing multiple flow variables.
Finally, the statistical plots of flow and O-D attribute variables, the spatial flow map, and the choropleth map are linked through dynamic brushing. Dynamic brushing was originally developed by Becker, et al. (1987) as a technique to explore relationships between variables displayed in the scatterplot matrix, which consists of multiple scatterplots. This research extends dynamic brushing to link the spatial and aspatial aspects of flow data and O-D attribute data.

Many spatial flow patterns can be revealed through dynamic brushing. Once a cluster of the statistical plots of flow variables is brushed, all flows in this cluster are displayed on the flow map. Since flows in the same cluster are considered similar, the flows displayed on the flow map also share similar characteristics. The spatial distribution of dominant flows can be revealed by brushing outliers in the statistical plots of flow variables. The distribution of zero-flows are displayed by brushing the flows that have a magnitude of zero in the statistical plot of flow variables. Relationships between flow variables, between flow variables and O-D attribute variables, and between flow variables and distance can be explored similarly in statistical plots of flow data and O-D attribute data. These flow patterns, when discovered, may lead to the generation of new hypotheses.

The inventive synthesis of PP, EDA, dynamic graphics, and SV offers an interactive, graphic environment and enriches the researcher's ability to explore spatial flow patterns in a complex flow data set and to generate new hypotheses. Once implemented, it will be a powerful system for geographers to study spatial flows.
The following sections are devoted to detailed discussions of three key issues: complexity reduction of spatial flow data sets, graphic representation of multiple flow variables, and detection of spatial flow patterns using dynamic brushing.

Reducing the Complexity of Flow Data Set Using PP Method

The fundamental goal of this research is to explore spatial flow patterns and to aid in hypothesis generation through visualization. Unfortunately, the human eye-brain combination can not directly visualize data in space that has more than three dimensions. To visualize a data set that has more than three variables, each of which represents a dimension, we must first reduce the number of variables.

There are a number of dimension reduction methods which can be used; the most commonly used are principal component analysis (PCA) and factor analysis (FA). Is PCA/FA the best dimension reduction method for the purpose of visualization in the present research? PCA produces orthogonal components based on variance-covariance of original variables. What is desired in this research are the most informative projections of the dimensions of multiple flow variables; the projections do not have to be orthogonal to each other. Are the principal components the most informative projections? Probably not, because the projections produced without the orthogonality constraint may be more informative. Thus, PCA/FA may not be the best dimension reduction method for the purpose of visualization in this research. A better alternative is needed to produce more informative projections of the multi-dimensional data set.

Projection pursuit (PP) is such an alternative. PP generates informative projections in a multi-dimensional data set through an optimization procedure (Friedman and Tukey
projections generated by the PP method of Friedman and Tukey (1974) tend to reveal clusters, while the projections generated by the PP methods of Jones and Sibson (1987) tend to reveal both outliers and clusters. As stated in Chapter II, both clusters and outliers are of interest to geographers in studies of spatial flows because outliers imply dominant flows while cluster represent similarity of flows.

There are two major reasons why PP was chosen over PCA. One is its superior capability for uncovering structures such as clusters and outliers. PCA is not designed for revealing clusters although the components derived from PCA are often used in subsequent cluster analysis. The present research focuses on revealing structures through the visualization of informative projections of a multi-dimensional data set. The ineffectiveness of PCA in generating such informative projections is clearly pointed out by Eslava-Gomez (1989, p.12). Although admitting that PCA is useful in "removing uninteresting directions of variation" (p. 2), Jones and Sibson (1987) regarded PCA as "something of blunt instrument" (p.2) and pointed out that PCA "relies for its success on the tendency for large variation also to be interestingly structured variation, a connection which is not logically necessary and often fails to hold in practice" (p. 2). Friedman and Tukey (1974), Jee (1985), and Jones and Sibson (1987) all demonstrated that PP methods are much better than PCA in revealing structures such as clusters. A second reason for the choice is that the projections generated by PP are invariant with changes in the scale of original variables (Jones and Sibson 1987). Certainly, a researcher does not want to see different results arise out of dimensional reduction simply because of changes in the scaling of the original variables. In
the case of commodity flows, one would expect to find similar flow patterns no matter if the
unit of flow is the ton or the kilo-ton. Unfortunately, the result of PCA varies as the scale of
original variables changes (Jee 1985, p. 10). Based on these two reasons, PP was chosen as
the dimension reduction method for reducing the complexity of spatial flow data sets in this
research.

To clearly show the difference in the results produced by PP methods and PCA, a
simple example is given in Table 1. Here, four points in a three-dimensional space are given.
Their coordinates are (1, 10, 10), (1, 1, 10), (10, 10, 1), and (10, 1, 1). The question is "can
we project the four points into a single dimension and still be able to separate them?" FA
(using PCA to extract factors) produces two factors, each of which gives two distinguishable
values: -0.866 and 0.866, representing two points. The result of PP methods is the PP-
generated variable, which has four values: -1.178, -0.336, 0.336, and 1.178, representing the
four points. It is obvious that the PP-generated variable yields better results than either of
the two factors.

When the above data are displayed graphically, the differences between FA and PP
method are even clearer. In Figure 20, (a) shows the four points in three dimensions. (b) is
the first factor, in which points 1 and 2 are at the same position, and points 3 and 4 are also
at the same position. (c) is the second factor, in which points 2 and 4 are overlaid; points 1
and 3 are also overlaid. Unlike (b) and (c), (d) clearly shows four points. Points 1 to 4 are
located from right to left. Assuming there is a cluster around each of the four points, the PP-
generated variable is able to reveal four clusters while FA could only detect two. Because
our objective is to visualize flow clusters and outliers in multiple flow variables, the PP method is clearly better than FA or PCA.

Table 1. Comparison of factor analysis and projection pursuit methods. Four points in a three-dimensional space are projected onto one dimension using factor analysis and projection pursuit methods. On each of the two factors, four points are projected to two points, while there are four separate points on the PP variable.

<table>
<thead>
<tr>
<th>Point No.</th>
<th>(x,y,z)</th>
<th>factor 1</th>
<th>factor 2</th>
<th>PP variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1,10,10)</td>
<td>-0.866</td>
<td>0.866</td>
<td>1.178</td>
</tr>
<tr>
<td>2</td>
<td>(1,1,10)</td>
<td>-0.866</td>
<td>-0.866</td>
<td>0.336</td>
</tr>
<tr>
<td>3</td>
<td>(10,1,10)</td>
<td>0.866</td>
<td>0.866</td>
<td>-0.336</td>
</tr>
<tr>
<td>4</td>
<td>(10,1,1)</td>
<td>0.866</td>
<td>-0.866</td>
<td>-1.178</td>
</tr>
</tbody>
</table>

Of course, PP methods are not without drawbacks. One is the requirement of heavy computation because optimization is involved. Fortunately, computer technology has developed rapidly and computation capacity has increased tremendously; therefore easing the difficulty of heavy computation. The other drawback is related to the accessibility of software. Unlike PCA, which is available in virtually all statistical packages, PP methods are not as widely available. Because of this, it was necessary to implement the PP method as part of the present research. Nonetheless, PP is an excellent dimension reduction method for the present research because the benefits of PP far out-weight its drawbacks.

Projection is located by optimizing Jones and Sibson's moment index
The key of the PP method is to determine the function that is to be optimized so that clusters and outliers can be detected. This function is referred as the PP index. Different PP methods have different PP indices. In Friedman and Turkey's PP, the PP index is defined as a product of local concentration and the overall spread of the data set; while a PP index based on Shannon's first order entropy is used by Jones and Sibson (1987). Both indices are very effective in revealing clusters, but the later is better at uncovering outliers (Jones and Sibson, 1987). Because we are interested in both clusters and outliers of spatial flow data, Jones and Sibson's PP index is utilized in our research.

![Figure 20. Graphical representation of the data in Table 1. PP result shows four points while FA results show only two points.](image)

**PP on Multiple Flow Variables**

Before moving into the use of the PP method for studying spatial flows, we need to explain certain mathematical notations relating to flow data. In this study, multiple flow variables are represented as a single dyadic flow matrix, in which the rows represent O-D
pairs and the columns are the different flow variables. The number of O-D pairs is equal to the number of origin and destination regions squared. For example, a 50x50 flow matrix has 2,500 different O-D pairs. $X_{NK}$ denotes a dyadic flow matrix with $N$ representing the number of O-D pairs and $K$ representing the number of different flows.

For better computation efficiency, Jones and Sibson's entropy index is simplified to a moment index, which is based on the third order and the fourth order moments of the projected data. The third order moment represents skewness while fourth order moment shows kurtosis (or peakedness). The moment PP index of $X_{NK}$ is represented as

$$pp(X_{NK}, p) = m^3_{3}(X_{NK}, p) + m^4_{4}(X_{NK}, p)$$

where $m_3(X_{NK}, p)$ and $m_4(X_{NK}, p)$ are the third and fourth order moments of $X_{NK}$ along projection $p$, and $p$ is the unit vector with $K$ elements which represent the direction of projection for each dimension (or each flow variable).

Jones and Sibson (1987) pointed out that the density with fixed mean and variance that minimizes index (4) is the normal density. Hence, maximization of (4) gives a $p$ that has the largest departure away from the normal distribution. Certainly, distributions with clusters and outliers have large PP values. Therefore, the detection of flow clusters and outliers becomes the following optimization process.

Maximize:

$$pp(X_{NK}, p) = m^3_{3}(X_{NK}, p) + m^4_{4}(X_{NK}, p) / 4$$

(6)
Subject to:

$$|p| = 1$$

(7)

The above optimization involves a non-linear objective function (6) with a non-linear constraint (7). This kind of optimization requires a sophisticated optimization algorithm, which will be addressed in chapter IV.

When a flow data set has more than three variables, it is very likely that the above optimization will have multiple solutions given different initial directions $$p$$. Most of these solutions represent local optima since there is only one global optimum. Both the global optimum and the various local optima are useful in the exploration of spatial flow patterns because they all reveal informative structures.

When structure-revealing projections are found, they can be stored for further analysis. The linear combination of original flow variables along each individual projection is named as a PP-generated flow variable, or a PP flow variable. The coefficient for the original flow variable $$k$$ in the linear combination is the $$k$$th element of the unit vector $$p$$, which represents the projection. Just like original flow variables, a PP flow variable can be mapped onto a two-dimensional flow map and displayed in a statistical plot such as a histogram or scatterplot.

When a PP-generated flow variable is displayed in a histogram, a researcher can visually identify clusters and outliers. Clusters represent similarity while outliers often represent dominant flows. It is possible that a cluster has some sub clusters, which may be of interest to us. To determine whether a cluster has sub clusters, the cluster can be isolated
and stored separately as a subset. Then we can apply PP to the subset to reveal sub-clusters. This process can be repeated recursively until no further sub-clusters are revealed.

Like the factors that are derived by factor analysis, PP flow variables cannot be explained in a straightforward fashion. Although a PP flow variable represents a linear combination of original flow variables, there is not a unique link between the value of the PP flow variable and the original flow variables via the linear combination. It is possible that two different sets of original flow variable observations may yield the same value of the PP flow variable. One can find exactly what the PP flow variable really represents by applying dynamic brushing to scatterplots containing both the PP-generated flow variable and the original flow variables.

**PP on a Flow Variable with a Time Series**

A single flow variable with a time series can be treated in the same way as multiple flow variables with each time period treated as a flow variable. In the dyadic flow matrix, columns represent flows in different time periods and rows represent O-D pairs. When the PP method is applied to such a dyadic flow matrix, PP flow variables are generated. The PP-generated flow variables uncover clusters and outliers. The clusters demonstrate which O-D pairs have similar flows during the whole period. The outliers exhibit extremely large flows over the entire time span. The flow patterns, e.g., clusters and outliers, can also be mapped onto a two-dimensional flow map.

**PP on Multiple O-D Attribute Variables**

Relationships between flow variables and O-D attribute variables are important to the study of flow patterns, and may contribute significantly to hypothesis generation. Spatial
flows may be related to many social-economic attributes of origins and destinations. There are two methods to explore the relationships between a flow variable and multiple O-D attribute variables. One method is to directly link the flow variable and each of the O-D attribute variables. This may be difficult, especially when the number of attribute variables is very large. The other option is to reduce the number of attribute variables by applying the PP method, and then to relate the flow variable to the PP-generated attribute variables. Since the number of PP-generated attribute variables is much smaller, a researcher can visually explore their relationship with the flow variable in the statistical plots.

**PP can be applied to multiple O-D attribute variables in a manner similar to that applied to multiple flow variables.** If there are \( N \) origins (or destinations) and \( K \) different O-D attribute variables, then a matrix \( X_{NK} \) with \( N \) rows and \( K \) columns can accurately represent the whole data set. Application of PP to \( X_{NK} \) generates new attribute variables. Each PP-generated attribute variable displays the overall characteristics of the original attribute variables from a certain optimized angle. Like original O-D attribute variables, a PP attribute variable can be mapped onto a choropleth map and statistical plots.

**The Detection of Spatial Flow Patterns Using Dynamic Brushing**

The original dynamic brushing developed by Becker, et al. (1987a) was used to explore relationships between multiple variables displayed in a scatterplot matrix. A number of geographers extended dynamic brushing to link scatterplots and a geographic map. Among those is Gou (1993) who implemented two-way brushing to link a flow map and a scatterplot matrix of O-D attribute variables. This research further extends dynamic brushing
to link a flow map, a choropleth map of O-D attributes, a statistical plot of flow variables, and a statistical plot of O-D attribute variables.

The following section discusses issues involved in mapping the flow of a single variable and multiple variables, choropleth mapping of O-D attributes and statistical plotting of both flow variables and O-D attribute variables. The use of dynamic brushing to reveal spatial flow patterns is demonstrated at the end of this section.

Flow Mapping

A flow map displays direction, magnitude and spatial distribution of flows. It is a vital component in an interactive environment for exploring spatial flow patterns because it is the only place where the spatial characteristics of flow data are displayed. Previous research focused on developing methods for graphic representation of a single flow variable. For example, Gou (1993) implemented three methods for flow representation. The first, called the line with number, uses a line to connect the origin and destination, and displays a number at the midpoint of the line showing the exact flow magnitude. Proportional band width is the second method, which draws a line between an O-D pair with a width that is proportional to the flow magnitude. The third method is graduated shading, in which origin and destination are connected by a corridor with constant width while the density of shading within the corridor is proportional to flow magnitude. Tobler (1987) believes that proportional band width method is interpreted more correctly. Gou (1993) agrees that the proportional band width method yields the best visual effect among the three.
Graphic Representation of a Single Flow Variable

The present research uses the line with number method and the proportional width method because the former displays exact magnitude information and the later has the best visual effect. In addition, a new method called the quantile width method is developed to distinguish small flows. The quantile width method is similar to the proportional width method. The only difference between the two lies in how the width is determined. In the proportional width method, the width is directly proportional to the flow magnitude. In the quantile width method, the width is proportional to an ordinal value which shows the relative magnitude of a flow compared with the rest of the flows. The details of how the width is determined are addressed in Chapter IV.

Once representation methods are chosen, they can be used to map a single flow variable onto a two-dimensional flow map. The contents of the flow map must be flexibly adjustable based on the upper and lower limits of flow magnitude, and also on the upper and lower thresholds of O-D distance. The latter is especially important, and it allows us to study the relationship between flows and the distance between O-D pairs. According to gravity models, spatial interaction normally decreases as O-D distance increases, which is often referred to as distance decay. The regularity of distance decay can be examined graphically using the approach described above.

Graphic Representation of Multiple Flow Variables

Mapping multiple flow variables onto a single flow map is not straightforward. In this research, two different methods are suggested to solve this problem. One is the mapping of PP-generated flow variables. A PP-generated flow variable can be displayed onto a flow
map using one of the three flow representation methods. The PP-generated flow map shows the overall flow patterns in a data set with multiple flow variables.

The second method is color composite flow mapping that uses a composite color to represent three flow variables. In image processing, three images representing different spectral bands can be combined into a color composite image by assigning each band to one of the three basic color components, i.e., blue, green and red. The resultant composite color is determined by the reflectances in each of the three spectral bands. For example, reflectance of vegetation might be high in band 7, and low in band 4 and 5. When blue, green and red are assigned to bands 4, 5 and 7 respectively, the color of vegetation in the composite image is red. The color of water in the composite image is blue because it has high reflectance in band 4. Detailed interpretation of the resulting color image requires special training.

Certainly, the color composite method could be used to study flow data with multiple flow variables. Three flow variables could be represented by blue, green and red with the magnitude of each flow variable represented by the intensity of each color. Therefore, a color composite flow map can indirectly display three flow variables. Lines with constant width could be drawn to connect O-D pairs, and filled with the composite color representing the values of three flow variables.

The color composite method is limited to the representation of only three variables at one time. When a flow data set has more than three flow variables, one can apply PP to reduce the number of flow variables to three or fewer, and then map the PP-generated flow variables using the color composite method. Many different colors may appear on the color
composite flow map, and it will be difficult to interpret, possibly more difficult than the interpretation of color satellite images.

No matter which of the two methods is used, interpretation of the flows on a flow map representing multiple flow variables is not straightforward because each flow is a combination of multiple flow variables. Such combinations can be displayed on scatterplots of flow variables. Hence, an intuitive method for interpretation is to link the flow map to scatterplots that represent the combinations, thus relating flows on the flow map to flow variables displayed on the scatterplots. When a flow is chosen on the flow map, the corresponding points, indicating values of the flow variables, are highlighted on the scatterplots.

Statistical Plot of Flow Variables

Statistical plots of flow variables are used to display aspatial aspects of flow data. A researcher can explore relationships between flow variables and structures such as clusters and outliers in statistical plots. Two kinds of statistical plots are used in this research: A histogram and scatterplots. A histogram is used when only one flow variable is selected while scatterplots are used when two or more flow variables are selected.

Histogram of a Flow Variable

An effective method to detect spatial flow patterns in a data set with a single flow variable is to draw a histogram. Figure 21. A histogram of a flow variable, showing clusters, outliers, and non-flows.
histogram (or density plot) of the flow variable. The horizontal axis of the histogram is the
classified magnitude of flow; while the number of O-D pairs that are in each flow class is
represented by the vertical axis. A visual inspection of the histogram can easily detect
clusters, non-flows and outliers. O-D pairs within the same cluster have similar flows, and
outliers have extremely high flow magnitude. Non-flows, if there are any, are located at the
far left side of the histogram. For example, Figure 21 shows clearly two large clusters in the
center, two outliers at the right hand side, and some non-flows at the left.

**Scatterplots of Flow Variables**

As reviewed in last chapter, Becker, et al. (1987a, 1987b) used a scatterplot matrix
to display multiple variables. In the case of K variables, there are K rows and K columns in
the scatterplot matrix. Each cell displays the correlation between two variables. The cells
in the upper half of the matrix are confusing and redundant because they are mirror images
of the lower portion.

In present research, a scatterplot triangle is used to represent multiple flow variables.
A scatterplot triangle is simply the lower portion of a scatterplot matrix, thus eliminating the
confusion and redundancy. The point of origin in each scatterplot is always at the lower-left
corner of the cell. Relationships between multiple flow variables can be explored in a
scatterplot triangle using dynamic graphics.

The role of the scatterplot triangle in exploring spatial flow patterns lies in two
aspects. One is for interpretation of flows on a flow map that represents multiple flow
variables using either color composite flow mapping or PP flow mapping. When a
scatterplot triangle is linked to a color composite flow map and a PP flow map, the exact
meaning of each flow on the flow map can be interpreted. The other role is the easy identification of structures such as clusters and outliers by visual inspection. Outliers indicate extremely large flows, while clusters imply the similarity of flows.

*Statistical Plot of O-D Attribute Variables*

Statistical plots of O-D attribute variables are used to display aspatial aspects of O-D attribute data. A researcher can explore relationships between O-D attribute variables and structures such as clusters and outliers in statistical plots. Comprehension of O-D attributes leads to a better understanding of spatial flow patterns because flows are often related to the social-economic conditions in the origin and destination regions. Two kinds of statistical plots are used in this research: a histogram and scatterplots. The histogram is used when only one variable is selected while scatterplots are used when two or more variables are selected. Because statistical plots of O-D attribute variables are similar to those of flow variables, the details concerning construction and function of the statistical plots are not repeated here.

*Choropleth Mapping of O-D Attributes*

A choropleth map is used to map the spatial aspects of O-D attribute data. One of the major roles that the choropleth map plays in this research is to display spatial distribution of regional characteristics. When the choropleth map is overlaid graphically by the flow map, one can visually study the relationships between spatial flows and regional characteristics. For example, by examining a migration flow map which is underlaid graphically by the choropleth map of unemployment rates, one may see clearly that people usually move from regions with high unemployment rates to regions where employment opportunity is better.
The flow map, choropleth map of O-D attributes, statistical plot of flow variables, and statistical plot of O-D attributes are the four basic views for studying spatial flows in this research. If these views stand alone, their contributions to the detection of spatial flow patterns are limited. But, when they are linked through dynamic brushing, the integration of the four offers an interactive graphic environment that enriches the researcher's ability to explore spatial flow patterns and to aid in hypothesis generation.

A hypothetical example, shown in Figure 22, is created to illustrate how dynamic brushing can be used to link the four views in this research. Figure 22 (a) shows flows between regions A, B and C. A choropleth map of regional characteristics can be on (a) and graphically overlaid by the flow map. Figure 22 (b) is a scatterplot triangle of three flow variables F1, F2 and F3. The scatterplot at row n and column m in the triangle is called panel (m, n). For example, panel (2, 1) is the scatterplot of F1 and F2. There are six data points in each panel, and each of them represents a flow between a pair of regions. Since there are three regions in this case, the number of interregional flows is six. Figure 22 (c) shows a scatterplot triangle of three O-D attribute variables. The number of data points in each scatterplot is three, which is the number of regions. At this point, (a), (b) and (c) stand alone, and the relationship between the three is unknown.

Anchor regions, which are the regions that are of interest to a researcher, are usually defined before dynamic brushing is activated from the flow map. In Figure 23, region B is defined as the only anchor region. When region A is brushed on the flow map, the flows
between A and the anchor region B are displayed on (a) and are also highlighted on all the scatterplots in (b). The attributes of the brushed region A are highlighted on panels (2, 1), (3, 1) and (3, 2) in (c). Thus, the spatial aspect (a) and aspatial aspect (b) of the flows are linked, and the flows are linked to the regional characteristics (c) of the origin and destination. This is specially useful when a PP-generated flow variable is mapped on the flow map. The highlighted points on (b) explain the meaning of flows displayed on (a).
Dynamic brushing may also be initiated from the scatterplot triangle of the flow variables. In Figure 24 shows such an example. When a data point is brushed in panel (2, 1) of (b), the flow from region A to B is displayed on the flow map, revealing that the brushed point represents the flow from region A to B. The data points representing flows from region A to B are also highlighted in panels (3, 1) and (3, 2). The attributes of regions A and B are
highlighted on panels (2, 1), (3, 1) and (3, 2) in (c). This kind of brushing is often used when interesting structures such as outliers and clusters are detected in the statistical plot of flow variables. Brushing the flows within a structure reveals the spatial distribution of these flows.

Figure 24. Dynamic brushing initiated from the scatterplot triangle of flow variables. When a data point representing the flow from region A to B is brushed in panel (2, 1) on (b), the corresponding flows from region A to B are displayed on the flow map (a) and other panels on (b). The attributes of regions A and B are highlighted in all panels on (c).

Furthermore, dynamic brushing may be initiated from the scatterplot triangle of O-D attribute variables. Figure 25 demonstrates such an example. When two data points in panel
(2, 1) on the triangle are brushed, the flows between the two regions are displayed. The flow map reveals that the two brushed points represent regions A and B. The flows between regions A and B are highlighted in all panels on the scatterplot triangle of flow variables. The attributes of regions A and B are also highlighted in other panels on (c).

Figure 25. Dynamic brushing initiated from the scatterplot triangle of O/D attribute variables. When two data points are brushed in panel (2, 1) on (c), the flows between the two regions represented by the two brushed points are displayed on the flow map (a) and all panels on (b). The attributes of the regions are also highlighted in other panels on (c).

The following examples demonstrate how dynamic brushing can be used to explore spatial flow patterns in an interactive graphic environment. The spatial flow patterns to be
explored are: spatial distribution of dominant flows, spatial distribution of zero-flows, similarity of flows, relationships between flow variables, relationships between flow variables and O-D attribute variables, relationships between flow variables and distance, and the change of flow patterns in time.

Detection of Dominant Flows

The first important flow pattern is the spatial distribution of dominant flows. Dominant flows are usually outliers in the statistical plot of the original flow variables and/or PP-generated flow variables. One can simply brush the outliers, then the dominant flows will be automatically displayed on the flow map. Another way to detect dominant flows is to display those flows that have a magnitude above a certain threshold.

Detection of Zero-flows

Zero-flows can be detected by brushing the flow with zero magnitude in the statistical plot of original flow variables and/or PP-generated flow variables. Once zero-flows are brushed, they are displayed on the flow map by using thin lines to connect the origin and the destination regions.

Detection of Similar Flows

Similar flows appear within the same cluster in the statistical plot of original flow variables and/or PP-generated flow variables. Brushing a cluster displays all flows within the cluster on the flow map.

Detection of Relationships between Flow Variables

Relationships between flow variables are displayed in scatterplot triangles of the original flow variables and/or PP-generated flow variables. Dynamic brushing of each
scatterplot in the scatterplot triangle may reveal relationships that can not be directly visualized without brushing (Becker, et al. 1987a, 1987b).

**Detection of Relationships between Flow Variables and O-D Attribute Variables**

Relationships between flow variables and O-D attribute variables can be revealed by linking the flow map, choropleth map, statistical plot of flow variables, and statistical plot of O-D attribute variables through dynamic brushing. In the case of state-level interregional migration in the United States, dynamic brushing may reveal that large migration flows are associated with states that have a large population. When large flows are brushed in the statistical plot of flow variables, the states with a large population will be highlighted in the statistical plot of O-D attribute variables. These large flows are also displayed on the flow map. When the flow map is graphically underlaid by the choropleth map of total population for each state, one can easily see the relationship between large migration flows and large population.

**Detection of Relationships between Flow Variables and Distance**

Relationships between flow variables and distance are revealed through interactive retrieval of flows within a given range of distances. The researcher can visualize the change of spatial flows as the distance range is adjusted.

**Detection of Changing Flow Patterns in Time**

Detection of changing flow patterns in time involves two steps. The first is to explore spatial flow patterns in each time period by using dynamic brushing and other tools. The second step is to compare flow patterns in each period.
Summary

To facilitate the exploration of spatial flow patterns and aid in hypothesis generation in a complex data set containing multiple flow variables, this research proposes an inventive combination of projection pursuit (PP) methods, exploratory data analysis (EDA), scientific visualization (SV), and dynamic graphics. PP is used to reduce the number of flow variables and O-D attribute variables to make the flow data set less complicated.

In an interactive graphic environment, data in a complex flow data set is displayed in four graphic views: a flow map of original flow variables and/or PP-generated flow variables, a choropleth map of original and/or PP-generated O-D attribute variables, a statistical plot of original and/or PP-generated flow variables, and a statistical plot of original and/or PP-generated O-D attribute variables. These four views are linked through dynamic brushing, which is a powerful tool for exploring relationships between spatial and aspatial aspects of flow data and O-D attribute data.

Two approaches for mapping multiple flow variables onto the same flow map is proposed. One is to map a PP-generated flow variable, which represents multiple flow variables. The other approach is color composite flow mapping, which uses composite color to represent three flow variables.

This chapter has also demonstrated how dynamic brushing can be used with other tools to explore spatial flow patterns such as spatial distribution of dominant flows, spatial distribution of zero-flows, similarity of flows, relationships between flow variables, relationships between flow variables and O-D attribute variables, relationships between flow
variables and spatial distance, and the change of flow patterns in time. The implementation of this new methodology is addressed in the next chapter.
CHAPTER IV
IMPLEMENTATION OF THE PROTOTYPE SOFTWARE: PPFLOW

This chapter focuses on the implementation of a prototype software package PPFLOW, which stands for Projection Pursuit FLOW Data Analysis System. PPFLOW is designed to be used to establish the proof-of-concept of the methods proposed in last chapter and not as a full-fledged operational system. The first section discusses structural and organizational issues related to PPFLOW, including system requirements, user interface, data organization, file structure, and hardware and software considerations. The modules and functionality of PPFLOW are addressed in the second section. The third section discusses the optimization of projection pursuit (PP), which is complex since it involves maximization of a non-linear PP index with a non-linear constraint. The details of the evaluation of the PP index and its partial derivative are discussed, and the optimization algorithm that is used is also addressed.

Structure and Organization of PPFLOW

Because of the experimental nature of this research, no formal design was conducted in the development of PPFLOW. Based on the experience of the author, it was established that PPFLOW should follow the following three guidelines:
1) it should be easy to use because potential users may not have strong computer background and they may not be able to operate complicated computer systems;

2) it must provide an interactive environment with effective techniques for spatial flow visualization, and

3) it must be easily exportable to other computer platforms;

The first guideline plays an important role in choosing the type of interface, which will be described later in this section. The second is a guideline for designing the visualization and analysis functionality of PPFLOW. The third is a major factor in determining the software and hardware environment of PPFLOW.

System Requirements

The focus of this research is to detect spatial flow patterns in a data set containing multiple flow variables. Based on the nature of the study, which has been addressed in Chapter III, PPFLOW should be able to:

1) represent multiple flow variables on a flow map using an effective graphical representation;

2) display one to one, one to many, and many to one directional flows;

3) display spatial flows based on distance;

4) display zero-flows;

5) adjust the content of the flow map based flow magnitude;

6) reduce the complexity of a data set containing multiple flow variables;
7) display regional characteristics using choropleth mapping approach;
8) link the flow map with statistical plots of the flow variables;
9) link the flow map with statistical plots of the O-D attribute variables;
10) link flow variables and O-D attribute variables; and
11) link elements of the choropleth map with O-D attribute variables.

These general requirements determine that PPFLOW should not only provide effective graphical representation of spatial flows and dynamic visualization and exploration, but also numeric methods such as PP methods to reduce the complexity of flow data sets containing multiple flow variables.

Data Organization and File Structure

The flow data sets studied in this dissertation research contain three types of data. One is the flow data itself, in which there are multiple flow variables. Each flow variable is represented as an O-D matrix. The second is the data giving the location and boundaries of the origin and destination regions. The third is social-economic data associated with these regions.

The three types of data are stored in three separate ASCII files. Thus, a spatial flow data set consists of three data files: the flow data file, the O-D attribute data file and the O-D location and boundary file. A flow data file contains names and descriptions of flow variables, the user identifiers of all the regions, and the flow magnitude between all O-D
pairs for all the flow variables. An O-D attribute data file contains names and descriptions of O-D attribute variables, the user identifiers of all the regions, and the attributes of all the regions for all the variables. An O-D location and boundary file contains the coordinates of the center point and boundaries, and user identifiers of all regions. The detailed content and structure of these files are described in Appendices A, B, and C.

Flow Chart and Modular Design of PPFLOW

To fulfill the general requirements outlined above, the major processes are proposed for PPFLOW, and the relationships between these processes are displayed in Figure 26. At the top of Figure 26 are three types of data that may exist in a spatial flow data set. Starting from flow variables, the first process is flow variable selection. The next process is to apply projection pursuit (PP) methods to the selected flow variables for the purpose of reducing them to a smaller number (usually one or two) of new flow variables, called PP-generated flow variables, each of which is a linear combination of the original flow variables. The next step splits into two directions. One implies that the original and/or PP-generated flow variables are selected and plotted onto scatterplots or a histogram, displaying the distributions and relationships between the selected flow variables. In the other direction, the original and/or PP-generated flow variables are selected and displayed onto a flow map with proper graphic representation. Flow mapping requires not only flow variables, but also the coordinates of the center point of each region.

Similar processes are applied to the O-D attribute variables. First, all or part of the original O-D attribute variables are selected for dimension reduction. Next, PP methods are applied to the selected O-D attribute variables for the purpose of reducing them to a smaller
Figure 26. Flow chart of PPFLOW. Major processes are displayed.
number (usually one or two) of new attribute variables, which are called PP-generated O-D attribute variables. The next step diverts into two directions. One direction is to select and plot the original and/or PP-generated O-D attribute variables onto scatterplots or a histogram, revealing the distributions and relationships between the selected attribute variables. The other alternate is to select and display the original and/or PP-generated O-D attribute variables onto a choropleth map. Choropleth mapping requires not only O-D attribute variables, but also the coordinates of boundaries for each region.

O-D location and boundary data are used as inputs to flow mapping and choropleth mapping. Also, the coordinates of center points are used to calculate the distance between O-D pairs, creating an O-D distance matrix.

Finally, the flow variable scatterplots, the flow map, the O-D distance matrix, the O-D attribute choropleth map, and the O-D attribute scatterplots are linked by using brushing techniques to allow the dynamic exploration of spatial flow patterns.

Based on the flow chart, the author divided PPFLOW into ten modules for the purpose of implementation. Each module performs a distinct function, and is integrated with the others. The ten modules and the relationships between them are displayed in Figure 27. The Data Input Module loads flow data sets into the system. The complexity of data may be reduced in the Projection Pursuit Module, in which PP-generated variables are produced to represent the original ones. The Flow Mapping Module maps a selected flow variable onto a flow map with proper spatial flow representation, which is selected in the Flow Representation Module. The content of the flow can be adjusted in the Flow Magnitude Adjusting Module and the O-D Distance Adjusting Module. The Choropleth Mapping
Module shows the distribution of selected social-economic characteristics of origin and destination regions. The Flow Variable Plotting Module draws statistical plots of selected flow variables. The O-D Attribute Variable Plotting Module draws statistical plots of selected O-D attribute variables that are associated with individual origin and destination regions. The Dynamic Brushing Module links the flow map, the choropleth map, the statistical plot of flow variables, and the statistical plot of O-D attribute variables in an animated fashion through dynamic brushing. The details of each module will be addressed later in this chapter.

User Interface

Currently, PPFLOW is designed as an experimental system. Because the author is currently the only user of the system, at first the user interface may not seem to be a critical issue. But, when it is considered that potential users of the system will be mostly geographers who may not necessarily have extensive computer knowledge, it is clear that the future operational version of PPFLOW should have an interface that guides users to make the operation of the system easy. The guideline used in designing the user interface of PPFLOW is that it must be as simple as possible to use and its operation must be intuitive.

Among the many types of user interfaces, menu-type interfaces and command line interfaces are the most popular. In command line interfaces, the user types in a command to have the computer undertake the desired action. Many GIS packages, such as Arc/Info and OSU-MAP-for-the-PC, use a command line interface. One of the disadvantages of the command line interface is that it requires a long learning period. A user has to learn the meaning and syntax of many commands before he/she can operate the system. A menu-type
Figure 27. The ten modules of PPFLOW.
and the typing effort is minimal. Because potential users of PPFLOW may not have extensive computer knowledge, a menu-type interface was chosen for PPFLOW to make it easy to use.

Figure 28 shows the user interface of PPFLOW. At the first level, the menu bar has nine items: file, flow mapping, flow representations, choropleth mapping, plot flow variables, plot O/D attribute variables, projection pursuit, brushing, and help. Each menu item, except for the last one, is associated with one of the ten modules described above. Clicking a menu item reveals a pulldown menu, which consists of a number of push-buttons. When a push-button is pushed, the specific function attached to it will be activated.

The middle-left of Figure 28 is the drawing area for the flow map and the choropleth map, the upper-right is the drawing area for statistical plot of flow variables, and the lower-right is the drawing area for statistical plot of O-D attribute variables. The lower-left contains the adjustment bars for flow magnitude and distance, which are associated with the Flow Magnitude Adjusting Module and the O-D Distance Adjusting Module.

Hardware and Software Consideration

Speed and portability are the two major concerns of PPFLOW. High speed computation and graphics are required for an interactive, real-time flow visualization system. In addition, PPFLOW should be easily portable to any other computer platforms.

Graphics workstations achieve higher speed graphics and computation than personal computers or main-frame computers. Personal computers such as the PC and Apple Macintosh are too slow in terms of both graphics and computation. Main-frame computers are faster, but their graphic capabilities are often severely limited. Among graphics
Figure 28. Interface of PPFLOW.
workstations, Sun workstations and Silicon Graphics workstations are the most popular. The latter has higher speed, but is slightly more expensive than the former. Since both are available in the GIS Lab of Geography Department, a Silicon Graphics workstation was selected as the platform on which to implement PPFLOW.

For better portability, X Window is the best choice. X Window is a window system that is supported on virtually all workstations. One of the major advantages of X Window is its hardware independence. A similar environment is also available on personal computers, such as MS Windows on the PC. X Window offers a powerful Xlib library, which provides the lowest level programming interface to the X Window system. In addition, there are many tools such as Motif and Xview built upon Xlib. These tools facilitate the development of user interfaces. For example, Motif allows a programmer to build the interface without having to learn many of the details of the underlying window system. Motif has been chosen as the interface development tool for this research because it is the de facto industry standard and it is far more powerful than Xview, which is easier to use, but will soon be completely replaced by Motif. Currently, Release 5 of X11 (version 11) and Motif 1.2 are being used in the present research.

The C language was chosen because of its portability across different platforms and its powerful functions such as dynamic memory allocation. In addition, X Window and Motif are written in C, and have a natural interface with C.

In summary, PPFLOW has been developed on a Silicon Graphics workstation, written in C, with the Xlib and Motif supporting libraries. The software environment of PPFLOW is illustrated in Figure 29.
Modules and Functionality of PPFLOW

As described in last section, PPFLOW is divided into ten modules for the purpose of implementation. The integration of the ten modules enables the user to explore spatial flow patterns in an interactive, real-time environment. The function of each module and its relation to other modules are discussed in detail in the following paragraphs.

Data Input Module

The Data Input Module is designed to read flow data sets into PPFLOW. As described in the Data Organization and File Structure section, a flow data set may contain three types of data: flow data, O-D attribute data, and the boundary data of origin and destination regions. Each type of data is organized in a separate file. For example, a data set of state-level, interregional migration in the United State may contain:
1) a number of migration matrices, each representing state-to-state migration during one period of time;
2) a number of attribute variables such as unemployment rate, crime rate, per capita GNP, and total population for each state; and
3) coordinates of a center point and the boundaries of each state.

When the above data are organized following the format specified in Appendices A, B and C, they can be loaded into PPFLOW through the \textit{Data Input Module}. PPFLOW can accept any flow data sets as long as they are organized in the specified format.

\textit{Flow Mapping Module}

This module allows flexible flow mapping of a selected flow variable in a flow data set. As described in last chapter, there are two approaches for mapping multiple flow variables: PP flow mapping and color composite flow mapping. When a PP-generated flow variable is selected, multiple original flow variables are mapped onto the flow map because the PP variable represents multiple original variables. Currently, the color composite flow mapping approach has not been implemented due to time limitations.

The user can specify a set of origin and destination regions, and map the flows between the selected regions. Multiple sets can be specified and displayed on the same map. For the purpose of distinction, each set may be displayed in a different color, which is chosen by the user at any stage of the operation. The graphic representation of spatial flows is determined in the following module.
**Flow Representation Module**

Three graphic representations are implemented in PPFLOW. In the proportional width method, a line with an arrow at the end connects an origin with a destination. The width of the line is drawn proportional to the magnitude of the flow using the following formula:

\[ W_i = \frac{(F_i - F_{\text{min}})}{(F_{\text{max}} - F_{\text{min}})} \times W_{\text{max}} \]  

(8)

where

- \( W_i \): width of the line representing flow \( i \);
- \( F_i \): magnitude of flow \( i \);
- \( F_{\text{max}}, F_{\text{min}} \): the largest flow and smallest flow in a flow data set;
- \( W_{\text{max}} \): the largest width applied to a flow data set.

For example, if the smallest and largest flows are 0 and 100 respectively and the largest width is 10, a flow of 100 has a width of 10, and a flow of 50 has a width of 5.

The second is the quantile width method, in which the width of a line is determined by the relative magnitude of a flow compared to the rest in a data set. To determine the quantile width of a flow, all the flows in a data set must be sorted from low to high based on flow magnitude, and then the quantile width can be calculated as:

\[ W_i = \frac{F_{p_i}}{F_{\text{max}}} \times W_{\text{max}} \]  

(9)

where

- \( W_i \): width of the line representing flow \( i \);
- \( F_{p_i} \): the position of flow \( i \) in the sorted list;
\( F_{\text{max}} \): the number of flows, which is the product of the number of origins and the number of destinations in a flow data set;

\( W_{\text{max}} \): the largest width applied to a flow data set.

In a data set consisting of 100 flows, the quantile width of the tenth (from the smallest) flow will be 2 when the largest width \( W_{\text{max}} \) is set at 20, no matter what the magnitude of this flow is.

It is simple to tell the difference between the proportional width method and quantile width method. For example, a hypothetical flow data set contains 100 flows, in which the largest is 1000 and the smallest is 0. Flow i has a magnitude of 100 and is the eleventh largest flow in the data set, so its position in the sorted list will be 90. When the maximum width is set at 20, proportional width method assigns flow i a width of 2 while quantile width method yields a width of 18.

In a migration data set, most migration flows are small and only a few are very large (Tobler, 1987). When the proportional width method is used, all small migration flows have similar, narrow width; while the quantile width method will assign different widths to these flows. Hence, the quantile width method is better for distinguishing the small flows in a migration data set.

The third flow representation method is called the line with number. A line connects the origin and destination of a flow with an arrow at end pointing the direction of the flow. The magnitude of the flow is displayed near the midpoint of the line. This method is useful when the user is interested in the exact amount of flow, but its overall visual effect is not as good as the previous two.
The three representation methods can also be used to represent two-directional flows. Given origin i and destination j, the line connecting i and j is divided into two equal-length segments. The segment closer to j is used to represent the flow from i to j, while the one closer to i represents the flow from j to i.

This module is designed so that the user can switch from one representation method to another at any stage of the operation. Switching representation methods in this module will change the appearance of flow representation on the flow map. The proportional width method is the default representation.

Flow Magnitude Adjusting Module

This module has been developed to display flows based on an adjustment of flow magnitude. Often, the user is interested in spatial flows in a certain range of magnitudes. For example, one may want to concentrate on large flows for to examine the distribution of dominant flows, or one may want to see only the distribution of zero-flows.

Two slide bars are implemented to control the range of flow magnitude. One sets the upper-limit of flow magnitude, and the other sets the lower-limit. Whenever either of the two limits is adjusted, the content of the flow map will change instantly. The two limits can be set at the same value; thus allowing the display of flows at a single magnitude. This can be used effectively to display zero-flows.

Flow magnitude adjustment can be applied either to the whole data set, or to a set of pre-selected regions. In the former case, all flows that are within the specified magnitude range will be displayed. In the latter, only those flows that are connected with the pre-selected regions and are within the range will be displayed. In both cases, the flows
displayed on the flow map are also highlighted in the histogram or scatterplots of flow variables, and the attributes of regions that are associated with these flows are highlighted in the histogram or scatterplots of O-D attribute variables. All these actions are instantly executed at the same time as the adjustment is made on the slide bar.

**O-D Distance Adjusting Module**

This module has been developed to display flows based on the adjustment of the distance between the origins and destinations. Based on gravity models, distance is a major factor that affects a flow between two regions. The longer the distance, the smaller the flow tends to be. This is often referred as distance decay. This module is used to visualize and examine the phenomenon of distance decay in spatial flows in a two-dimensional geographic space. To achieve this goal, two slide bars have been implemented to control the range of O-D distance. One sets the upper-limit of O-D distance, and the other sets the lower-limit. Whenever one of the two limits is adjusted, the content of the flow map will change instantly.

The O-D distance adjustment can be applied to either the whole data set, or a set of pre-selected regions. In the former case, all flows that are within the specified distance range will be displayed. In the latter, only those flows that are connected with the pre-selected regions and are within the distance range will be displayed. In both cases, the flows displayed on the flow map are also highlighted in the histogram or scatterplots of flow variables, and the attributes of regions that are associated with these flows are highlighted in the histogram or scatterplots of O-D attribute variables. All these actions are instantly executed at the same time as the adjustment is made on the slide bar.
Choropleth Mapping Module

In PPFLOW, choropleth mapping is used to map O-D attribute variables of the data set for the purpose of revealing the relationships between spatial flows and regional characteristics. According to gravity models, spatial flows are determined by not only O-D distance, but also the attractiveness and other characteristics of the regions. In the case of migration, the attractiveness is often related to economic conditions, living environment, etc. In PPFLOW, the regional characteristics are represented by a selected O-D attribute variable or a PP-generated attribute variable, which represents a set of original attribute variables.

Once a variable is chosen, the entire range of variable value is divided into a number of equal intervals (currently, the number of intervals is set to 5). Regions within the same interval are classified into the same group. Each group is then represented by a color shade on the choropleth map, in which the darker the color is, the higher the value of the variable. Thus, the spatial variation of the selected variable is clearly displayed in a two-dimensional geographic space.

PPFLOW graphically overlays the choropleth map and flow map, with the latter on the top. This allows the user to visually analyze the relationships between spatial flows and regional characteristics.

Flow Variable Plotting Module

When a single flow variable is selected, a histogram (or density plot) of the variable will be displayed in PPFLOW. The horizontal axis of the histogram is the flow magnitude divided into a number of intervals; while the number of O/D pairs that lie in each interval
is represented by the vertical axis. A visual inspection of the histogram can easily detect clusters, non-flows and outliers.

When two flow variables are selected, a scatterplot showing the distribution and relationships between the two variables will be displayed. When more than two variables are involved, a scatterplot triangle (the lower portion of the scatterplot matrix) will be displayed in PPFLOW. The point of origin in each scatterplot is always at the lower-left corner of the cell. The variable names are displayed in the diagonal cells of the matrix. For each scatterplot, the variable represented by the vertical axis is always the one at the right side, and the variable on the horizontal axis is the one at the top. The location of a point as referenced to the point of origin in each scatterplot represents the flow magnitudes of the two flow variables. The number of points in each scatterplot is the number of O-D pairs, which is the product of the number of origins and the number of destinations. A visual inspection of the scatterplot triangle can easily identify outliers and clusters. Outliers indicate extremely large flows, while clusters imply that flows are similar.

Like original flow variables, PP generated flow variables can be included in the histogram and the scatterplot triangle, in which the user can visualize the relation between the PP flow variables and the original flow variables.

**O-D Attribute Variable Plotting Module**

Like flow variables, O-D attribute variables can also be plotted in a histogram, a scatterplot, or a scatterplot triangle. A histogram shows the distribution of the values of a selected attribute variable, while a scatterplot displays the relationship between two attribute variables. PP generated attribute variables can also be displayed in a histogram or
scatterplots. The number of points in a scatterplot is the number of unique origin and destination regions involved in the flow data set.

**Projection Pursuit Module**

In PPFLOW, a projection pursuit (PP) method has been implemented to reduce the dimensions of flow variables and O-D attribute variables. The PP-produced new variables are treated the same way as the original variables, and can be used for visualization. For example, a PP-generated flow variable can be mapped onto a flow map, demonstrating the spatial distribution of the PP-generated flow. A PP-generated O-D attribute variable can be displayed on a choropleth map, showing the variation of overall regional characteristics. Because optimization of PP is a complicated issue, it will be discussed separately in a later section.

**Dynamic Brushing Module**

Dynamic brushing is designed to link multiple views such as the flow map, the O-D attribute choropleth map, the statistical plots of flow variables and attribute variables in a real-time fashion. It can help the user examine the relationships among these views; thus better revealing the relationships between spatial flows and regional characteristics. Based on these relationships, new spatial flow patterns may be generated.

Dynamic brushing can be initiated from any of the above described views. When brushing is started from a scatterplot in a scatterplot triangle of flow variables, a rectangular brush appears and the data points covered by the brush are highlighted. The corresponding data points in other scatterplots of the scatterplot triangle are also highlighted. In addition, the corresponding flows are displayed on the flow map. The attributes of regions that are
associated with the highlighted flows are also highlighted on the histogram or the scatterplot triangle of O-D attribute variables. When brushing a histogram of a flow variable, the user should move the brush along horizontal axis to select flows by flow magnitude.

When brushing is started from a scatterplot in a scatterplot triangle of O-D attribute variables, a rectangular brush appears and the data points covered by the brush are highlighted. The corresponding data points in other scatterplots of the scatterplot triangle are also highlighted. In addition, flows that connect highlighted regions are displayed on the flow map, and highlighted on the histogram or the scatterplot triangle of flow variables. When brushing a histogram of an O-D attribute variable, the user should move the brush along the horizontal axis to select regions by the O-D attribute value.

Brushing flow map/choropleth map can be carried out in two different ways: one with pre-selected regions, known as anchor regions; the other without anchor regions. In the presence of anchor regions, flows between a brushed region and anchor regions are displayed on the flow map, and these flows are highlighted on the histogram or the scatterplot triangle of flow variables. The attributes of the brushed region will be highlighted on the histogram or scatterplot triangle of O-D attribute variables. Without anchor regions, brushing will not draw any new flows on the flow map, and the histogram or the scatterplot triangle of flow variables, only the attributes of the brushed region will be highlighted on the histogram or the scatterplot triangle of O-D attribute variables.
Projection Pursuit Optimization

In Chapter III, a moment index was chosen as the PP index because it can reveal structures such as clusters and outliers that depart from the normal distribution. Optimizing such an index is very challenging because the index is not linear and it has a non-linear constraint.

This section is organized in three parts: the calculation of the moment index and its partial derivative, the optimization algorithm, and the overall optimization procedure.

Calculation of the PP Index and its Partial Derivative

Given a dyadic flow matrix and projection p, the third order and fourth order moments in Equation (6) are defined as (following Jones and Sibson 1987):

\[ m_3(X_{nk,p}) = \sum_{n=1}^{N} \left( \sum_{k=1}^{K} p_k x_{nk} \right)^3 \]

\[ = \sum_{n=1}^{N} \sum_{k=1}^{K} p_k x_{nk} \sum_{j=1}^{K} p_j x_{nj} \sum_{l=1}^{K} p_l x_{nl} \]

\[ = \sum_{n=1}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} p_j p_k x_{nj} x_{nk} x_{nk} \]

\[ = \sum_{n=1}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} p_j p_k x_{nj} x_{nk} x_{nk} \sum_{l=1}^{N} \sum_{k=1}^{K} p_l x_{nl} \]

\[ = \sum_{n=1}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} T_{nk} x_{nj} x_{nk} \]

\[ = \sum_{n=1}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} T_{nk} \]

\[ = \sum_{n=1}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} T_{nk} \]
where $T_{ik}$ is the third order tensor of $X_{NK}$

$$T_{ik} = \sum_{n=1}^{N} X_{nk} X_{ik} X_{nk}$$

(12)

and $T_{ijkl}$ are the fourth order tensor of $X_{NK}$

$$T_{ijkl} = \sum_{n=1}^{N} X_{nk} X_{ik} X_{mr} X_{ml}$$

(13)

Notice that $T_{ik}$ and $T_{ijkl}$ need calculated only once for a given $X_{NK}$. The computational complexity is in the order of $N$. In addition, the evaluation of $m_{i}(X_{NK}, p)$ and $m_{j}(X_{NK}, p)$ require computation in the order of $K^3$ and $K^4$ respectively. Together, the evaluation of the PP index in (6) requires computational complexity is in the order of $NK^4$. That means the number of flow variables, not the number of O-D pairs, has the larger impact on the speed.
of optimization. Thus, this PP index is effective for a data set with large number of O-D pairs.

The partial derivative of (6) to \( p_s \) element of vector \( p \), is:

\[
\frac{dpp(X_{NRD})}{dp_s} = 2m_3(X_{NRD}) \frac{dm_3(X_{NRD})}{dp_s} + \frac{1}{2}m_4(X_{NRD}) \frac{dm_4(X_{NRD})}{dp_s}
\]

(14)

where

\[
\frac{dm_3(X_{NRD})}{dp_s} = 3 \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{t=1}^{K} pp_t T_{ij} - pp_{ij} m_3(X_{NRD})
\]

(15)

and

\[
\frac{dm_4(X_{NRD})}{dp_s} = 4 \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{k=1}^{K} pp_{ik} T_{ij} - pp_{ij} m_4(X_{NRD})
\]

(16)

The details on how to derive the first partial derivative of the moment index are addressed in Appendix E. Each partial derivative requires \( O(K^3) \) computation, so that the calculation of \( K \) partial derivatives require \( O(K^4) \) computation, given that \( T_{ik} \) and \( T_{ij} \) are already calculated.

The evaluation of the objective function (6) and its gradient at each iteration has a computational complexity of \( O(NK^4) \).
Optimization Algorithm for the Moment PP Index

Since (6) is a non-linear objective function and has a non-linear constraint (7), the solution for optimizing (6) can not be analytically derived. By normalizing $p$ at each iteration, we can change this constrained optimization problem to a unconstrained optimization. There are many numeric algorithms for non-linear, multi-variate ($K$ variables in this case) unconstrained optimization. Some require the calculation of the gradient, and others do not, but the algorithms which use the gradient are usually more efficient.

Among the algorithms that use the gradient, steepest descending method is probably the one most commonly used. In the steepest descending method, the direction of movement at each iteration is always along the direction opposite to the gradient. Because the distance moved at each step is very short, many iterations are required to maximize or minimize the objective function. This is time-consuming, and not suitable for an interactive system like PPFLOW.

Conjugate gradient method is another gradient-dependent algorithm. Unlike steepest descending method, the moving direction in the conjugate gradient method is determined by the gradient and the direction of last step. The directions in two consecutive iterations are conjugate each other. The direction of iteration $i+1$ is defined as

$$h_{i+1} = g_{i+1} + \gamma_i h_0$$

(17)

where $h_0$ is the direction at iteration $i$, $g_{i+1}$ is of gradient at $p_{i+1}$, and $\gamma_i$ is

$$\gamma_i = \frac{(g_{i+1} - g_i) \cdot g_{i+1}}{g_i \cdot g_{i+1}}$$

(18)
Once a moving direction is determined, the moving size along that direction can be easily solved by maximizing the PP index value.

The conjugate gradient method ensures that the number of iterations required to reach a local optimal is at most the number of variables K (Jacobs, 1977). This is considerably more efficient than the steepest descending method. Based on the efficiency criteria, conjugate gradient method has been adopted to optimize the moment-based PP index in PPFLOW.

*Optimization Procedure of the Moment-based PP index*

The procedure for optimizing the moment-based PP index is summarized in Figure 30. An explanation of the procedure is given below:

1) A data set with multiple flow variables are given and represented in a dyadic flow matrix $X_{NK}$ where $N$ is the number of O-D pairs while $K$ is the number of flow variables;

2) Then, the third order and fourth order tensors of $X_{NK}$ are calculated based on equations (7) and (8);

Now, we are ready to optimize the PP index.

3) A initial projection $p_0$ (a K-element vector) is given. $p_0$ can be any of the K axes in the original data, or any direction in the K-dimensional space;

4) $p_0$ is set to $p$;

5) $p$ is normalized so that it becomes a unit vector;
Figure 30. Flow Chart of PP optimization using conjugate gradient method.
6) The gradient at $p$ is calculated based on equations (9), (10) and (11), and stored in $g$, which is also normalized;

7) The moving direction $a$ is calculated based on equations (11) and (12), and is normalized;

8) The moving distance $d$ along $a$ is determined using a one-dimensional optimization method;

9) Move to the new projection $q$: $q = p + d\, a$. Normalize $q$;

10) If $q$ equals $p$, then $q$ is optimal and stop; if not, $p$ is set to $q$ and go back to step 5).

Steps 5 to 10 are repeated until an optimal solution is located. Because the conjugate gradient method is used, the number of iterations will not be larger than $K$, the number of variables. This high efficiency fits well to the interactive, real-time system PPFLOW.

Summary

Practical issues concerning the implementation of the methods that are proposed in Chapter III have been addressed in this chapter. According to the criteria of easy-of-use, a menu-type interface has been chosen for PPFLOW. PPFLOW is divided into ten modules, each of which performs a distinct function in the system. Among the ten modules, six deal with the extracting, mapping, plotting and linking of flow data and O-D attribute data in various ways, one with the data input, and one with projection pursuit, in which the conjugate gradient method is used to optimize the moment-based PP index. Integration of the ten modules enables the user to effectively explore a spatial flow data set by examining the distribution of spatial flows, the relationships between spatial flows and regional
characteristics, and the relationships between spatial flows and distance; thereby revealing spatial flow patterns which lead to new hypotheses.
CHAPTER V
RESULTS AND AN ASSESSMENT OF THE
PROOF-OF-CONCEPT IMPLEMENTATION

The implementation of PPFLOW (Projection Pursuit FLOW Data Analysis System) was discussed in the last chapter. This chapter demonstrates the operation of this trial implementation and addresses its utility. The demonstration focuses on the following aspects: the layout of the PPFLOW interface, generation of new variables by projection pursuit, overlay of the spatial flow map on a choropleth map, spatial flow query by O-D distance, spatial flow query by flow magnitude, brushing the spatial flow map, brushing the histogram and scatterplot triangle of flow variables, and brushing the histogram and scatterplot triangle of O-D attribute variables.

A data set containing state-level interregional migration in the forty-eight contiguous states of the United States is used in this chapter. This data set is used to test the system because it is easily available in digital format and is familiar to most geographers. Since the objective is not a study of migration, this data set is used only to demonstrate the functionalities of PPFLOW, and is not intended to reveal new migration patterns. However, it is hoped that migration researchers will be encouraged to try the new tools developed in this research.
The data set contains three types of data. One is migration flow data, which come from the population census of 1960, 1970 and 1980.

MIGRAT60: inter-state migration during the period from 1955 to 1960, represented by the number of persons in all ages;

MIGRAT70: inter-state migration during the period from 1965 to 1970, represented by the number of persons in all ages; and

MIGRAT80: inter-state migration during the period from 1975 to 1980, represented by the number of persons in all ages.

Each of the three flow variables is represented by a 48x48 O-D flow matrix. The second type of data is attribute data for each state from the Statistical Abstract of the United States. The attribute variables are selected from those that are available in digital format.

UNEMPLOY: unemployment rate of each state in 1980, measured by the number of unemployed per hundred people;

CRIME: crime rate of each state in 1980, measured by the number of offenses per million population;

GNP/PC: per capita GNP (in current dollars) of each state in 1980;

POPU75: total population of each state in 1975;

POPU80: total population of each state in 1980;

AREA: area of each state, measured by square miles.

The third type of data is the boundary and the center point of each state.
The Layout of the PPFLOW Interface

The Menu Bar and Slide bar of PPFLOW

Figure 31 shows the layout of the PPFLOW interface. At the top is the menu bar, which has nine items: File, Flow Mapping, Flow Representation, Choropleth Mapping, Plot Flow Variables, Plot O-D Attributes, Projection Pursuit, Brushing, and Help. The first eight items correspond to eight of the ten modules described in last chapter. The lower-left area

Figure 31. The layout of PPFLOW interface. Flow map, choropleth map, histogram of a flow variable, and scatterplot triangle of O-D attribute variables are displayed.
contains two sets of slide bars: the one at left is for flow magnitude adjustment, corresponding to the flow magnitude adjusting module; the one at right is for O-D distance adjustment, corresponding to the O-D distance adjusting module.

**Drawing Area for the Flow Map and the Choropleth Map**

At the left-center of Figure 31 is the drawing area for the flow map and the choropleth map. The variable displayed on the choropleth map is a PP-synthesized variable PP_ATT1, generated from UNEMPLOY, CRIME, GNP/PC. The depicted flow variable is PP_FLOW1, a synthesized flow variable that has been generated by the projection pursuit module from the original flow variables MIGRAT60, MIGRAT70 and MIGRAT80. Synthesized spatial flows between Ohio, California and Florida are displayed on the flow map.

**Drawing Area for the Statistical Plot of Flow Variables**

At the top-right is the drawing area for the statistical plot of flow variables. A histogram of the PP-generated flow variable PP_FLOW1 is displayed in this example. The horizontal axis shows the value of PP_FLOW1, and the minimum and maximum values are displayed below the axis. The values are classified into 20 equal intervals (the number of intervals is set to 20 after experimentation by the author because it produces the best visual effect). The occurrence (or frequency) of each interval is depicted on the vertical axis. It is obvious that there are a large number of small flows and only a few large ones; which follows a Pareto distribution as suggested by Tobler (1987).

When multiple flow variables are selected, a scatterplot triangle (see Figure 40 and Figure 41 later in this chapter) of the variables will replace the histogram. A point in a
scatterplot represents a flow between an O-D pair. A detailed description of a scatterplot triangle is given next.

*Drawing Area for the Statistical Plot of O-D Attribute Variables*

At the bottom-right of Figure 31 is the statistical plot of O-D attribute variables. In this case, five variables are displayed in a scatterplot triangle. PP_ATT1 and PP_ATT2 are two variables generated by the projection pursuit module from original attribute variables UNEMPLOY, CRIME and GNP/PC. Notice that each point in the scatterplots represents one state. Each scatterplot in the scatterplot triangle is called a panel. For example, the scatterplot at row 2 and column 1 is referred to as panel (2, 1), and the one at row 4 and column 3 is panel (4, 3). The panels located at diagonal cells in the triangle are used to display the variable names and the range of their values. The minimum value is displayed at the lower-left corner while the maximum value of a variable is displayed at the upper-right corner of a panel. For example, panel (4, 4) is PP_ATT1, whose minimum and maximum values are -2.6 and 2.8. For the panels with data points, the vertical axis represents the variable at the right side while horizontal axis represents the variable at the top. In panel (3, 1), the vertical axis represents GNP/PC while the horizontal axis represents UNEMPLOY.

When only one variable is selected, a histogram of the variable will replace the scatterplot triangle. The histogram of an O-D attribute variable is similar in structure to that of a flow variable.
Generating New Variables by Projection Pursuit

Three new variables are generated by maximizing the moment-based PP index, as described in Chapters III and IV. One of them is a flow variable, and the other two are O-D attribute variables. Before continuing, the reader may want to go back to the Projection Pursuit Optimization section of Chapter IV to review the meaning of initial direction (or starting position), unit vector, moving direction, moving distance (or moving size), and iteration.

PP_FLOW1 is generated from three flow variables MIGRAT60, MIGRAT70 and MIGRAT80. The initial direction is given as (1, 2, 3), which is a directional vector in the three-dimensional space constructed by MIGRAT60, MIGRAT70 and MIGRAT80. Normalization of this directional vector yields a unit vector (.27, .53, .80), which is used in optimization. An optimal solution is reached after one iteration. The details of each iteration are displayed in Table 2. The relationship between PP_FLOW1 and original flow variables are shown in Figure 41 later in this chapter. Panel (4, 2) of Figure 41 shows a nearly perfect linear-correlation between PP_FLOW1 and MIGRAT70. Panels (4, 1) and (4, 3) also display good linear relationships between PP_FLOW1 and MIGRAT60, and between PP_FLOW1 and MIGRAT80. Thus, this single, new flow variable can effectively represent the original three flow variables.
Table 2. PP_FLOW1 produced by PP from MIGRAT60, MIGRAT70 and MIGRAT80. Initial position is set at (1,2,3)

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Starting position</th>
<th>PP index value at starting position</th>
<th>Moving direction</th>
<th>Moving size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(.27, .53, .80)</td>
<td>183,198,547,968.00</td>
<td>(.90, .15, -.40)</td>
<td>.50</td>
</tr>
<tr>
<td>1</td>
<td>(.64, .54, .54)</td>
<td>255,314,165,760.00</td>
<td>(.89, .44, -.09)</td>
<td>.00</td>
</tr>
</tbody>
</table>

PP_ATT1 and PP_ATT2 are both generated from UNEMPLOY, CRIME and GNP/PC. As has pointed out in previous chapters, there may exist multiple informative projections in a multi-dimensional data set, and these projections are produced by assigning different initial directions. The initial direction for PP_ATT1 is (1, 2, 15), and PP_ATT1 is reached after two iterations (Table 3). PP_ATT2 is reached after three iterations (Table 4) given an initial direction of (3, 2, 3). These initial directions are assigned arbitrarily by the user. The results from these initial directions depend upon the original data set. For example, the applications of PP to MIGRAT60, MIGRAT70 and MIGRAT80 always produce the same projection PP_FLOW1 no matter what the initial direction is, while the applications of PP to UNEMPLOY, CRIME and GNP/PC produce two different projections PP_ATT1 and PP_ATT2 depending upon the initial direction.

Table 3. PP_ATT1 generated by PP from UNEMPLOY, CRIME and GNP/PC. Initial direction is (1,2,15)

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Starting position</th>
<th>PP index value at starting position</th>
<th>Moving direction</th>
<th>Moving size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(.07, .13, .99)</td>
<td>3,789.25</td>
<td>(-.39, .91, -.10)</td>
<td>1.01</td>
</tr>
<tr>
<td>1</td>
<td>(-.23, .74, .63)</td>
<td>18,062.52</td>
<td>(-.92, .38, -.06)</td>
<td>.02</td>
</tr>
<tr>
<td>2</td>
<td>(-.25, .75, .63)</td>
<td>18,063.80</td>
<td>(-.92, .38, -.06)</td>
<td>.00</td>
</tr>
</tbody>
</table>
Table 4. PP_ATT2 generated by PP from UNEMPLOY, CRIME and GNP/PC. Initial direction is (3,2,3)

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Starting position</th>
<th>PP index value at starting position</th>
<th>Moving direction</th>
<th>Moving size</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(.64, .43, .64)</td>
<td>10,397.71</td>
<td>(.54, .34, -.77)</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>(.94, -.04, -.33)</td>
<td>11,931.86</td>
<td>(.94, -.04, -.33)</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>(.94, .26, -.22)</td>
<td>12,814.74</td>
<td>(.43, -.14, .89)</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>(.90, .08, .59)</td>
<td>13,939.91</td>
<td>(.75, .30, .59)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

The relationships between the two PP-generated attribute variables and the original attribute variable are displayed in Figure 31. Panel (4, 2) shows a good linear relation between PP_ATT1 and CRIME. PP_ATT1 also has a good linear correlation with GNP/PC, which is depicted in panel (4, 3). Thus, we can use PP_ATT1 to represent both CRIME and GNP/PC. PP_ATT2, on the other hand, has very good linear relationships with UNEMPLOY (see panel (5, 1)).

All three new variables are generated from the three original variables in no more than three iterations. This supports the notion that conjugate gradient method can optimize the PP index in no more than K iterations in K dimensional data.

**Overlay of the Spatial Flow Map on the Choropleth Map**

Three flow representation methods, i.e., the line with number, the proportional width and the quantile width as described in chapter IV, are implemented in PPFLOW. Figure 32 shows the two-directional migration from 1975 to 1980 between Ohio, California and Florida by using the line and number method. The magnitude of each flow is displayed beside the line connecting the respective origins and destinations.
Figure 32. Flow map using the line and number method. Two directional flows between Ohio, California and Florida are displayed.

The same flows are displayed on Figure 31 by using the proportional width method. It is clear that the Ohio-to-Florida migration is much larger than the Florida-to-Ohio migration, and that the Ohio-to-California migration is larger than the California-to-Ohio migration. The proportional width method clearly has a better visual effect than the line and number method, which is consistent with the results of Tobler's experiment (1987).

Figure 33 and Figure 34 are used to compare the proportional width method and quantile width method. Both display the two-directional migration flows between Ohio,
Wyoming and Louisiana. Figure 33 results from application of the proportional width method, and displays no visual difference among any of the flows because all of them are relatively small, compared to large flows. Figure 34, resulting from application of the quantile width method, on the other hand, clearly reveals differences among the flows. This suggests that the quantile width representation is a better method when flow magnitudes are small.
A choropleth map of O-D attributes is displayed in Figure 31. The displayed variable is a PP-generated variable PP_ATT1. The values of PP_ATT1 are divided into five equal intervals (equal intervals are used due to time constraints in implementation process, and uneven intervals and quantile methods could be used in the future), each of which is represented by a color shade. The darker the color is, the higher the value of PP_ATT1. Because PP_ATT1 displays a good linear relationship with both CRIME and GNP/PC, the
Figure 35. Choropleth map of unemployment rate of 1980, overlaid by the PP-synthesized migration flows between Ohio, California and Florida.

darker areas such as California and Florida are among the states with high crime rate and high per capita GNP. The differences in regional characteristics in terms of crime rate and per capita income are clearly displayed in this choropleth map, a result of combining projection pursuit approaches and choropleth mapping.

Graphic overlay of the flow map on a choropleth map illustrates relationships between spatial flows and regional characteristics. In Figure 31, the flow map, underlaid by the choropleth map of PP_ATT1, shows the PP-synthesized flows (represented by
Figure 36. A flow map containing all migration flows from or to Ohio in the period from 1975 to 1980.

PP_FLOW1) between Ohio, California and Florida. An visual inspection easily identifies that the flows from the light color area (Ohio) to darker areas (California and Florida) are larger than the reverse flows.

Figure 35 shows a similar example, in which the flow map displays the PP-synthesized migration flows between Ohio, California and Florida, and the choropleth map displays the unemployment rate in 1980. The graphic overlay of the two maps indicates that the flows from states with a high unemployment rate (i.e., Ohio) to states with a low
unemployment rate (i.e., California and Florida) are larger than those in the reverse direction. This might suggest a hypothesis that the migration is driven by better employment opportunity.

**Spatial Flow Query by O-D Distance**

Since spatial flows are often related to distance based on gravity models, it would be useful to display only flows within a certain distance. The function of query by distance is implemented in the O-D distance adjusting module. Figure 36 shows all the flows that are

---

Figure 37. A flow map containing all migration flows from or to Ohio, and that are within 502 miles of Ohio.
Figure 38. A flow map containing all migration flows that are connected with Ohio, California and Florida, and that are within 528 miles from each of these three states.

connected to Ohio. Notice that these flows are also highlighted in the flow variable histogram. The attributes of states that are associated with these flows are highlighted in the scatterplot triangle of O-D attribute variables. All 48 states are included in this example.

When the upper-limit of the O-D distance slide bar is adjusted to 502 miles, the number of flows on the flow map is largely reduced (see Figure 37). Only those flows that are connected to nearby states remain. As a result, not all the states are highlighted in the
Figure 39. A flow map showing all migration flows that are within 425 miles from each of the 48 states.

scatterplot triangle of O-D attribute variables. From the remaining flows, it is not very difficult to see that larger the distance is, the smaller the migration; this supports the regularity of distance decay.

The distance adjustment can also be applied to multiple states. Figure 38 displays the flows that are associated with Ohio, California and Florida, and that are within 528 miles from each of the three states.
Figure 40. A flow map containing migration flows that are connected to Ohio and are larger than 31377 persons.

When no states are selected before the adjustment of the O-D distance bar, the distance adjustment is applied to all states. Figure 39 shows such an example, in which all flows that are within 425 miles from each state are displayed.

Once the distance value of the O-D distance slide bars is adjusted, the content of flow map changes instantly, and the corresponding data are simultaneously highlighted in the statistical plots of flow variables and O-D attribute variables. This provides a good tool
Figure 41. Dynamic brushing initiated from flow map/choropleth map. The anchor states are California and Florida. The brushed state is Ohio. The PP-synthesized migration flows between the anchor states and the brushed state are displayed.

for geographers to visually examine the relationships between spatial flows and distance, and between spatial flows and selected regional characteristics.

**Spatial Flow Query by Flow Magnitude**

Another way to reduce the complexity of a flow map is to narrow the range of flow magnitudes. This is implemented in the flow magnitude adjusting module of PPFLOW. In an earlier example, all flows that are connected to Ohio are displayed in Figure 36, which
Figure 42. Dynamic brushing initiated from the histogram of the flow variable PPFLOW1. Three bars of the histogram, and the flows within the brushed bars are displayed on the flow map.

is cluttered because too many flows are displayed at the same time. When the lower-limit is raised to 31377, the content of the flow map is reduced to the one displayed in Figure 40. The remaining flows are the large ones, such as the flows from Ohio to California and Florida. This reflects the general patterns of north to south migration and east to west migration.
Figure 43. Dynamic brushing initiated from scatterplot triangle of flow variables. Two outliers in panel (4, 2) are brushed, and the flows represented by the brushed outliers are displayed on the flow map.

**Brushing the Spatial Flow Map and the Choropleth Map**

Brushing the spatial flow map/choropleth map provides a tool to link spatial data, i.e., the spatial flows displayed on the flow map and the spatial distribution of regional characteristics displayed on choropleth map, to aspatial data, i.e., the statistical plots of flow variables and O-D attribute variables. Brushing can be carried out in two different ways. One is brushing with pre-selected regions, known as anchor regions, which are of interest to the
Figure 44. Dynamic brushing initiated from scatterplot triangle of O-D attribute variables. Five outliers in panel (2, 1) are brushed, and the flows between the states represented by the brushed outliers are displayed on the flow map.

researcher. When a region is brushed, the flows between the brushed region and the anchor regions are displayed. Figure 41 shows the result of such brushing. In this example, California and Florida are selected as anchor regions. When Ohio is brushed, the flows between Ohio and the anchor regions instantly appear on the flow map. Meanwhile, the attributes of Ohio are highlighted in the scatterplot triangle of O-D attribute variables.
UNEMPLOY, CRIME and GNP/PC. Also, the flows are highlighted in the scatterplot triangle of flow variables MIGRAT60, MIGRAT70, MIGRAT80, and PP_FLOW1.

Another way of brushing the flow map/choropleth map is to brush without anchor regions. Because the brushed region is the only region selected in the brushing process, there will be no flows displayed on the flow map and the statistical plot of flow variables. Only the attributes of the brushed region are highlighted on the statistical plot of O-D attribute variables.

Brushing the Histogram and the Scatterplot Triangle of Flow Variables

Brushing the statistical plot of flow variables offers a mechanism to link the aspatial properties (as displayed on the statistical plot) to spatial properties (as displayed on spatial flow map) of the flow data. Figure 42 shows an example of brushing the histogram of the flow variable PP_FLOW1. When some of the large flows on the histogram are brushed, they instantly pop up on the flow map. We notice that the flows mainly move from east to west, which is an important spatial flow pattern that can not be discovered on the statistical plot of flow variables without brushing. The attributes of the states associated with the brushed flows are also highlighted in the scatterplot triangle of the O-D attribute variables UNEMPLOY, CRIME and GNP/PC.

An example of brushing the scatterplot triangle of flow variables is displayed in Figure 43. When two outliers on panel (2, 1) are brushed, two flows appear instantly on the flow map. One is from New York to New Jersey, and the other is from New York to Florida. These north to south migration flows represent important spatial flow patterns that are not revealed in the statistical plot of flow variables without brushing. Combining the flow
patterns displayed in Figure 42 and Figure 43, we now have two major spatial flow patterns: one is from east to west, and the other is from north to south.

The role that statistical plots play in this pattern detection process is to reveal structures such as clusters and outliers, which are linked to important flow patterns. The spatial flow patterns are uncovered on the spatial flow map by brushing the structures in the statistical plots.
Brushing the Histogram and the Scatterplot Triangle of O-D Attribute Variables

The aspatial to spatial brushing can also be initiated from the statistical plots of O-D attribute variables. An example of brushing the scatterplot triangle of O-D attribute variables is shown in Figure 44. As the states with the highest crime rate are brushed in panel (2, 1), the flows among California, Nevada, Colorado, Arizona, and Florida pop up on the flow map. This kind of brushing can be used as a tool to reveal the spatial flows between states.
that are of interest to the researcher. The histogram of an O-D attribute variable can also be brushed in a similar way.

**Studying Temporal Migration Data Using PPFLOW**

As described in the beginning of this chapter, the migration data set has state to state migration data for three time periods: 1955 to 1960, 1965 to 1970, and 1975 to 1980. This section illustrates how PPFLOW can be used to detect flow patterns in this temporal flow data set.

One of patterns that can be revealed by PPFLOW is the distribution of large migration flows during the three periods. Because there are only three time periods, the researcher can examine the large flows in each time period, and then summarize a general conclusion to the entire time span. Application of the flow magnitude adjusting module reveals large migration flows from 1955 to 1960, as displayed in Figure 45. Among the seven large flows, four are directed to California from New York, Illinois, Texas, and Washington. The other three are from New York to New Jersey and Florida, and from Pennsylvania to New Jersey.

In a similar way, large flows from 1965 to 1970 are revealed and depicted in Figure 46. Five of the eight large flows are related to California. Three of the five flows are directed to California from New York, Illinois and Texas while the other two are from California to Washington and Texas. Notice that the Texas-to-California migration is still larger than the California-to-Texas migration. The remaining three large flows are from New York to New Jersey and Florida, and from Pennsylvania to New Jersey.
Figure 47. Large migration flows during the periods from 1975 to 1980.

Figure 47 displays six large migration flows from 1975 to 1980. The largest one is from New York to Florida, followed by the migration flows from New York to New Jersey and California. The other three are from California to Texas, Oregon and Washington state.

In these three time periods, there are three consistent flows, which are from New York to New Jersey, Florida and California. The flows connected with California experience

major changes. From 1955 to 1960, large migration flows are directed to California from New York, Illinois, Texas, and Washington State. From 1965 to 1970, while large flows are still directed to California from New York, Illinois and Texas, large flows are also directed away from California to Texas and Washington, which were the major sources of migration to California during the period from 1955 to 1960. From 1975 to 1980, the only large flow
to California is from New York, while other large flows move out of California to Texas, Oregon and Washington.

Another approach to the study of temporal flow data is to explore PP-synthesized flows. When the flow magnitude adjusting module is applied to PP_FLOW1, which is generated by using the PP module and is a combination of the flows in the three time periods, major migration flows are revealed in the flow map (Figure 48). These flows are most similar to the ones from 1965 to 1970 because PP_FLOW1 is most strongly correlated to MIGRAT70. The patterns of east to west migration and north to south migration are well represented in Figure 48. When the number of time periods is large, this approach should be more efficient than the first one that explores the flows in every time period sequentially.

Preliminary Assessment of PPFLOW

The above is a demonstration of results derived from various functions in PPFLOW. Based on these practical results, it can be fairly concluded that the original objective of integrating projection pursuit (PP), scientific visualization (SV), exploratory data analysis (EDA) and dynamic graphics to detect spatial flow patterns and to aid in hypothesis generation is achieved. Of the many innovative tools that have been presented, several are worthy of special emphasis. First, PP is an effective tool to reduce the number of variables in spatial flow analysis. The application of the conjugate gradient method to PP optimization has ensured that PP variables are produced quickly enough to meet the requirement of an interactive visualization and exploration system. Second, the representation of multiple flow variables is achieved by mapping a PP-generated flow variable. Third, the inventive quantile width method provides an effective flow representation to visually distinguish small flows,
which was a problem that was not solved by any of the previous flow representation methods. Fourth, the interactive, graphic environment provides dynamic views of spatial and aspatial properties of a spatial flow data set. Dynamic brushing effectively clarifies relationships between the flow map, the choropleth map, the statistical plot of flow variables, and the statistical plot of O-D attribute variables. Finally, the O-D distance adjusting module provides an innovative tool to reveal the relationships between spatial flows and distance, which may lead to interesting spatial flow patterns and new hypotheses.

However, a complete assessment of a visual exploration system like PPFLOW is difficult at present time. As argued by Tang (1993) and Gou (1993), there is no generally accepted criteria and effective procedures for testing visual exploration systems. The goal of a visual exploration system is to discover new findings. Whether this goal can be achieved depends upon both the knowledge of the user and the quality of the data sets used for testing, of which neither can be judged objectively.

The second difficulty in assessing PPFLOW is the lack of test data sets. There are not many flow data matrices available in digital format. In fact, the migration data set used in this research is not complicated enough to fully demonstrate the power of PPFLOW because it contains only three flow variables MIGRAT60, MIGRAT70 and MIGRAT80, and the PP method is designed to visualize a much larger number of flow variables. Currently, finding data sets with enough complexity to fully test PPFLOW is quite difficult.

The third difficulty in evaluating PPFLOW is lack of expert users. PPFLOW is designed for expert users to explore flow patterns and aid in hypothesis generation. To achieve this objective, the user must have expertise in the area related to the flow data set.
For example, if the data set represents migration, the user should be specialist in migration research. Finding a number of expert users who have time and are willing to test PPFLOW is not an easy task.

Given the above difficulties, a complete assessment of PPFLOW at present time is extremely difficult, if not impossible. Nonetheless, the author has demonstrated the system a number of times to members of the faculty and graduate students in Geography Department at The Ohio State University. Their response to the demonstrations has been very positive. Of course, PPFLOW is not flawless. Further improvements to PPFLOW are suggested in next chapter.
CHAPTER VI
CONCLUSIONS

The objective of this research was to develop and implement new methodology to facilitate the exploration of spatial flow patterns, leading to new hypotheses. The results demonstrated in here strongly suggest that this objective has been achieved in this research. The proof-of-concept system on Silicon Graphics workstation, PPFLOW (Projection Pursuit FLOW Data Analysis System) represents the first system that integrates projection pursuit (PP), scientific visualization (SV), exploratory data analysis (EDA), and dynamic graphics to enable researchers to address hypothesis generation problems in the area of spatial flow analysis.

Although the methodology presented here has been developed within the context of spatial flow analysis, its basic principle of complexity reduction, visualization and exploration can be extended to general geographic research directed toward hypothesis generation. Dynamic linking between graphic representations of both spatial data and aspatial data is an effective approach for revealing the relationships between and within spatial and aspatial data; such relationships may lead to the generation of hypotheses.

This chapter summarizes the results of the research in terms of what has been achieved and what still needs to be accomplished. Future research directions will also be
discussed. Finally, the implication of this research for general geographic studies will be addressed.

Results of the Research

The most tangible result of this research is PPFLOW, a prototype system intended for implementing the two-stage methodology: first complexity reduction and then exploration by SV, EDA and dynamic graphics. The selected hardware and software yielded very good results. The graphics of the Silicon Graphics workstation are very fast, which yields real-time, interactive change in dynamic brushing. Even though a significant amount of time was invested learning Motif, sophisticated GUI (Graphic User Interface) software, it was well spent because of Motif's flexibility and enormous functionality in user interface design, and its good programming interface with the C language. Motif produced a menu interface for PPFLOW that is straightforward and intuitive. Even those without computer experience require only approximately 30 minutes to learn how to use this click-and-go system.

The Contributions of PPFLOW

The contributions of PPFLOW to studies of spatial flows lie in the following areas. First, the implementation of the moment-based PP provides an effective approach to reduce the complexity of a spatial flow data set; thus easing the detection of spatial flow patterns and the generation of new hypotheses. The projections generated by PP depart from the normal distribution, revealing structures such as clusters and outliers. PP can be used to reduce the number of flow variables and O-D attribute variables. PP-generated flow variables can be used in flow mapping and statistical plots of flow variables while PP-
generated attribute variables can be used in choropleth mapping and statistical plots of O-D attribute variables.

A second contribution of PPFLOW is to the display of multiple flow variables. Two approaches are proposed to display multiple flow variables. One is the color composite approach, in which red, green and blue in the composite color on the flow map are used to represent three individual flow variables. The other approach is PP flow mapping, which first synthesizes the multiple flow variables to form a new flow variable by using the PP method and then maps this PP-generated variable on a flow map. The relationships between the PP variable and original variables are displayed in scatterplots. Because of time limitations, only the second approach is implemented in PPFLOW.

To improve the graphic representation of spatial flows, the quantile width method was invented to visually distinguish small flows on the flow map. Differentiating small flows on graphics in a flow visualization system is very important because small flows count for the majority of the flows in some flow data sets such as migration data sets (Tobler, 1987). Unfortunately, previous methods, such as the proportional width method, the constant band with shading-gradient method, and the line and number method, can not visually differentiate small flows. The effectiveness of the quantile width method is clearly displayed in Chapter V.

Another contribution is that PPFLOW enables graphic overlay of the spatial flow map on a choropleth map, thus visually revealing the relationships between the spatial aspect of the flow data and the spatial aspect of O-D attribute data. This overlay mechanism offers
a rich environment that helps the user to generate spatial flow patterns that relate to regional characteristics.

The next contribution of PPFLOW lies in the multi-way linking through dynamic brushing. Brushing is not new, but never before has dynamic brushing been implemented to link both the spatial and aspatial aspects of flow data and regional attribute data. In PPFLOW, the aspatial aspects of both types of data are displayed on histograms or scatterplots, while the spatial aspects are displayed on a flow map and a choropleth map. The dynamic brushing in PPFLOW is able to link these graphics and to reveal relationships between multiple flow variables, between multiple O-D attribute variables, and between flow variables and O-D attribute variables. These relationships within and between the spatial and aspatial aspects of flow data and O-D attribute data are the basis for generating spatial flow patterns and new hypotheses.

Finally, the implementation of the O-D distance slide bars in PPFLOW enables the revelation of relationships between spatial flows and distance. For the first time, such relationships can be examined graphically and interactively on a computer screen. When a range of distance is assigned through the adjustment on the O-D distance bars, the flows falling into the selected range will be instantly brought up onto the screen. In addition to the distance slide bars, the flow magnitude slide bars are implemented to reveal large flows and zero-flows.

Future Improvements

Although PPFLOW is a successful proof-of-concept system, there are some areas which can be improved. First, map scale needs to be added on the flow map and choropleth
map. The format of the boundary data file in PPFLOW follows the format specified by the UNGENERATE command in ARC/INFO; this ARC/INFO format does not provide map scale. Currently, the map scale in PPFLOW is hard-coded for the U.S. interregional migration data set. To solve this problem, a record containing the map scale should be added to the boundary file. Second, the function of regional aggregation should be added to PPFLOW in the future. With the aggregation function, the user can study the same data set from different spatial resolutions. Third, the color composition method for representing multiple flow variables on the spatial flow map needs to be implemented. These have not been done due to time limitations.

Summary

The synthesis of complexity reduction, visualization and exploration offers an effective approach for studying complex spatial flow data. As a proof-of-concept system, PPFLOW has successfully integrated projection pursuit (PP), scientific visualization (SV), exploratory data analysis (EDA) and dynamic graphics. The test of the system using the U.S. interregional migration data set suggests that PPFLOW is a powerful system with effective tools for detecting spatial flow patterns, and aiding the researcher in the generation of new hypotheses.

Future Research Direction

This research has focused only on interregional flow data at regional-level. It has not solved, and was not intended to solve, all the problems of spatial flow analysis. There are many questions left unsolved. Along the line of this research, some directions for future research are proposed in the following sections.
Visualization and Exploration of Network-based Flows

One of the future research directions is the visualization and exploration of network-based flow data. A network contains nodes, links, and routes, which consist of multiple connected links. A route is a spatial entity that does not exist in a area-based or point-based flow data set. So, the representation of flows along routes is a new problem that is inevitable in the study of network-based flows. Because a link may be shared by many routes, displaying the flows of many routes on the same link will be a challenging problem. Nonetheless, the principle of complexity reduction by projection pursuit and integration of SV, EDA and dynamic graphics can be applied to studies of network-based flows.

Studies of Global Flows

A second future research direction is the exploration of global flow patterns. The two-dimensional visualization techniques used in this research are sufficient for regional flows, but global flows require three-dimensional visualization. There exist techniques to portray a globe using three-dimensional graphics, and the existing methods for the graphic representations of flows can be extended to three dimensions; therefore the representation of global flows using three-dimensional graphics is technically possible. Performance would be a concern because three-dimensional graphics requires significantly heavier computation than two-dimensional graphics. Could a near real-time performance still be achieved? Other than that problem, the methods and techniques adopted and developed in this research can certainly be extended to the study of global flows.
Integrating Existing Models with PPFLOW

Another direction of future research is to integrate existing models of spatial flows into PPFLOW. With modeling capability, PPFLOW could be used to forecast future flows. For example, one could build a gravity model for the state-level interregional migration using the existing migration data. The techniques of linear programming or linear regression could be applied to the model building process, with logarithms taken on both sides of the gravity model in order to transform it into a linear equation. If it is assumed that there is a population forecasting model that can predict future population for each state, one can use the future population in the gravity model to estimate future migration between the states. Then, one can visually analyze future migration using PPFLOW.

Implication to General Geographic Analysis

Although the methodology is developed for the study of spatial flows, its basic principle of complexity reduction, visualization, exploration and dynamic linking can be extended to general geographic research. The methodology consists of two parts. One is to reduce the complexity of a spatial flow data set by using projection pursuit methods. Geographers have been using factor analysis (FA) and principal component analysis (PCA) to reduce the number of variables in a complex multi-variate data set. Projection pursuit is another complexity reduction method and can be used in other geographic research together with PCA and FA. The other part of the methodology is to explore spatial flow patterns by using EDA, SV and dynamic graphics. Both the spatial and aspatial aspects of a flow data set are displayed graphically on computer screen, and the relationships between the two are
revealed by dynamic brushing. This, again, can be applied to the analysis of spatial and aspatial data involved in generic geographic studies.

Figure 49 displays a methodology for detecting spatial patterns in general geographic research. At the top is the spatial data set, followed by complexity reduction, which may be skipped if the data set is very simple. PCA, FA and PP can be used for complexity reduction. Then, the data is visually, and interactively analyzed in an environment which integrates SV, EDA and dynamic graphics. Such an environment, as implemented in PPFLOW, enriches the intuition and ability of a researcher to explore spatial patterns. As a result, new hypotheses may be generated. Confirmation of hypotheses leads to a different type of research and is not addressed here.

Figure 49. A methodology for general geographic research
With the success of experiments in dealing with spatial flow data in this research, the author has reason to believe that the methodology outlined in Figure 49 will be widely adopted by other geographers in the future. Though significant efforts will need be made to make the methodology more comprehensible and easily accessible, the day will come when user-friendly tools integrating SV, EDA and dynamic graphics are available to all geographers.
APPENDIX A

FORMAT OF FLOW DATA FILE

This appendix describes format of flow data file specified for PPFLOW. Courier New font is used for the contents of the data file while Italic Times New Roman font is used for explanation.

The following is a flow data file that contains state-level interregional migration among 48 contiguous states in the U.S. Migration in three time periods 1955-60, 1965-70 and 1975-80 are included in this file. Flow data file name must have extension ".flow".

Notice that columns represent origins while rows represent destinations, which is format used in the original census report, but is contradictory to conventional notation in geography.

```
area flow /*flow type: area flow, point flow are the two options*/
48 /*the number of places or origins/destinations*/
This is population migration between 48 states in US in 1960,70,80 /*description
of the data file, no longer than 80 characters*/
3 /*the number of variables*/
MIGRAT60 /*name of variable 1*/
Inter-regional Migration between 48 States in U.S. 1955-60 /*description of variable
1, no longer than 80 characters*/
MIGRAT70 /*name of variable 2*/
Inter-regional Migration between 48 States in U.S. in 1965-70 /*description of
variable 2, no longer than 80 characters*/
MIGRAT80 /*name of variable 3*/
Inter-regional Migration between 48 States in U.S. in 1975-80 /*description of
variable 3, no longer than 80 characters*/
47 46 44 43 42 37 41 38 30 26 23 25 21 16 17 18 11 12 13 14 40 39 32 31
33 34 35 36 27 28 29 24 19 20 15 22 10 48 9 8 7 5 6 4 1 2 3 /*user identifier of each
origin/destination in the order from column 1 to column 48 (or row 1 to row 48), numbers are separated by
```
The identifier must match the ones used in O-D attribute data file and boundary data file.

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>4746</td>
<td>1137</td>
<td>13517</td>
<td>1497</td>
<td>5515</td>
<td>5946</td>
</tr>
<tr>
<td>2488</td>
<td>1429</td>
<td>506</td>
<td>1268</td>
<td>1037</td>
<td>344</td>
<td>318</td>
</tr>
<tr>
<td>566</td>
<td>81</td>
<td>174</td>
<td>362</td>
<td>756</td>
<td>162</td>
<td>971</td>
</tr>
<tr>
<td>1987</td>
<td>380</td>
<td>1080</td>
<td>647</td>
<td>717</td>
<td>2559</td>
<td>383</td>
</tr>
<tr>
<td>385</td>
<td>380</td>
<td>216</td>
<td>427</td>
<td>517</td>
<td>2301</td>
<td>122</td>
</tr>
<tr>
<td>101</td>
<td>406</td>
<td>583</td>
<td>401</td>
<td>122</td>
<td>85</td>
<td>1012</td>
</tr>
</tbody>
</table>

/* row 1 of the O-D matrix for variable 1, numbers are separated by spaces, not commas */

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>6195</td>
<td>1270</td>
<td>17155</td>
<td>1488</td>
<td>9369</td>
<td>6521</td>
</tr>
<tr>
<td>2484</td>
<td>1642</td>
<td>540</td>
<td>1201</td>
<td>1047</td>
<td>553</td>
<td>588</td>
</tr>
<tr>
<td>386</td>
<td>137</td>
<td>162</td>
<td>271</td>
<td>459</td>
<td>234</td>
<td>1627</td>
</tr>
<tr>
<td>2591</td>
<td>130</td>
<td>1045</td>
<td>547</td>
<td>848</td>
<td>2185</td>
<td>331</td>
</tr>
<tr>
<td>259</td>
<td>476</td>
<td>174</td>
<td>326</td>
<td>440</td>
<td>1690</td>
<td>188</td>
</tr>
<tr>
<td>58</td>
<td>509</td>
<td>283</td>
<td>648</td>
<td>53</td>
<td>232</td>
<td>891</td>
</tr>
</tbody>
</table>

/* row 1 of the O-D matrix for variable 2, numbers are separated by spaces, not commas */

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>11489</td>
<td>2233</td>
<td>25600</td>
<td>2143</td>
<td>11234</td>
<td>9699</td>
</tr>
<tr>
<td>3978</td>
<td>2624</td>
<td>761</td>
<td>1736</td>
<td>1634</td>
<td>405</td>
<td>643</td>
</tr>
<tr>
<td>487</td>
<td>90</td>
<td>136</td>
<td>580</td>
<td>543</td>
<td>287</td>
<td>2774</td>
</tr>
<tr>
<td>3149</td>
<td>278</td>
<td>1259</td>
<td>852</td>
<td>498</td>
<td>5022</td>
<td>704</td>
</tr>
<tr>
<td>297</td>
<td>371</td>
<td>591</td>
<td>508</td>
<td>580</td>
<td>2296</td>
<td>74</td>
</tr>
<tr>
<td>113</td>
<td>1447</td>
<td>487</td>
<td>591</td>
<td>240</td>
<td>260</td>
<td>849</td>
</tr>
</tbody>
</table>

/* row 1 of the O-D matrix for variable 3, numbers are separated by spaces, not commas */

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
</table>
| 5497 | /* row 1 of the O-D matrix for variable 3, numbers are separated by spaces, not commas */
|   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |

... */ rows 2 to 48 of the O-D matrix for variable 1 omitted */
APPENDIX B

FORMAT OF O-D ATTRIBUTE DATA FILE

This appendix describes format of O-D attribute data file specified for PPFLOW.

Courier New font is used for the contents of the data file while Italic Times New Roman font is used for explanation.

The following is a O-D attribute data file that contains attribute data of 48 contiguous states in the U.S. The name of O-D attribute data file must have extension ".att".

area att /* attribute type: area att, point att are the two options */
48 /* the number of places or origins/destinations */
Social Economic Data of 48 States in US in 1980 /* description of the data file, no longer than 80 characters */
3 /* the number of variables */
UNEMPLOY /* name of variable 1 */
Unemployment Rate in 1980 /* description of variable 1, no longer than 80 characters */
CRIME /* name of variable 2 */
Crime Rate in 1980: The Number of Offenses per 100,000 Population /* description of variable 2, no longer than 80 characters */
GNP/PC /* name of variable 3 */
Per Capita GNP in 1980 /* description of variable 3, no longer than 80 characters */

The identifiers must match the ones used in inflow data file and boundary data file */

7.7 4.7 6.4 5.6 7.2 5.9 7.5 7.2 7.8 8.4
9.6 8.3 12.6 7.0 5.7 5.7 7.0 4.9 4.7 4.0
4.4 7.7 6.4 5.1 9.4 6.5 6.9 6.4 6.0 8.1
7.2 8.8 7.5 7.6 6.7 4.8 5.2 6.0 7.9 3.9
5.6 7.4 6.6 6.2 6.2 7.5 8.2 6.8 /* data of variable 1 */
4368 4680 4988 6079 5933 5882 6912 6401 3736 5431
930 5275 6676 4799 4799 4747 5433 2964 3243 4305
5379 6777 6630 4620 2552 4640 5439 5604 8402 3434
4498 4934 3417 3811 5454 5053 6143 5024 4782 4986
7333 5979 8171 5881 8854 6915 6687 7833 /* data of variable 2 */
7925 9131 7827 10125 9444 11720 10260 10924 9434 9462
| 8936 | 10521 | 9950 | 9348 | 9724 | 9358 | 8982 | 8747 | 7806 | 9365 |
| 9383 | 10339 | 10460 | 9392 | 7800 | 7819 | 7266 | 8073 | 8996 | 7613 |
| 7720 | 7488 | 6580 | 7268 | 8458 | 9116 | 9545 | 8536 | 8056 | 10898 |
| 10025 | 7841 | 8791 | 7649 | 10727 | 10309 | 9317 | 10938 | /* data of variable */
APPENDIX C

FORMAT OF BOUNDARY DATA FILE

This appendix describes format of boundary data file specified for PPFLOW. Courier New font is used for the contents of the data file while Italic Times New Roman font is used for explanation.

The following is a boundary data file that contains the coordinates of boundary and center point of 48 contiguous states plus Lake Michigan in the U.S. The name of boundary data file must have extension ".bnd".

You may notice that the format described here is consistent to the ASCII file format generated from ARC/INFO by using UNGENERATE command. The only difference is that the former uses "0 0" to separate two polygons while the later uses "END" to separate two polygons.

11 11.395360 14.203790 /* User identifier and the coordinates of the center point of a polygon. The identifiers must match the ones used in flow data file and O-D attribute data file */
12.180719 14.686713
12.236459 14.360172
12.271564 14.121241
12.302679 13.868314
12.346729 13.731348
12.351813 13.562401
12.351813 13.562401
11.421217 13.570602
10.448273 13.635513
10.448273 13.635513
10.509013 14.760424
10.509013 14.760424
11.246706 14.727272
11.744502 14.700171
coordinates of vertices on the boundary of a polygon.

notice that the first point is the same as the last point */
0 0 /* terminator of a polygon */
... /* repeat the above format for other polygons */
APPENDIX D
USAGE OF PPFLOW

PPFLOW (Projection Pursuit FLOW Data Analysis System) is developed to establish the proof-of-concept of the newly proposed methods that integrates projection pursuit, scientific visualization, exploratory data analysis, and dynamic graphics. It is developed on SiliconGraphics workstation, written in C, with the supporting libraries Xlib, Motif. The user interface of PPFLOW has the look-and-feel of Motif.

The user should type "ppflow" to start PPFLOW. After PPFLOW is started, the interface (Shown in Figure 28 and Figure 31) appears on screen. At the top is a menu bar, which has nine menu items: file, flow mapping, flow representations, choropleth mapping, plot flow variables, plot O/D attribute variables, projection pursuit, brushing, and help. A menu item has a pulldown menu, which in turn consists of a number of push-buttons. The middle-left is the drawing area for flow map and choropleth map, upper-right is the drawing area for statistical plot of flow variables, and the lower-right is the drawing area for statistical plot of O-D attribute variables. The lower-left are the adjustment bars for flow magnitude and O-D distance. The operation of PPFLOW is simply click-and-go, and does not require the user to input any data from keyboard. The following sections illustrated how to operate PPFLOW.
Data Input

Data files must be loaded before any other functions can be executed. If no data files are loaded, attempting to use other functions will bring up an error message which informs the user to load data first.

After the "File" button in the menu bar is clicked by pushing any buttons on the mouse, the file selection menu (Figure 50) pops up. There are four push-buttons in this menu. The "Open Flow Data File (.flow)" button is used to load flow data file, which must have the extension ".flow" and follow the format specified in Appendix A. Currently, an interregional migration data set is available and is stored in a file named "OD607080.flow".

The "Open O/D Attribute Data File (.att)" button is used to load O-D attribute data file, which must have the extension ".att" and follow the format specified in Appendix B. Currently, a file named "us.att" is available and it contains unemployment rate, crime rate, per capita GNP, total population, area of each state in the U.S.

The "Open Boundary Data File (.bnd)" button is used to load boundary data file, which must have the extension ".bnd" and follow the format specified in Appendix C. Currently, a file named "us.bnd" is available and it contains boundaries and a center point for 48 contiguous states of the U.S. When a boundary file is selected, the boundaries are

Figure 50. File selection menu
instantly displayed on screen in black color, and the title "Flow Map/Choropleth Map" is also displayed in black.

The "Quit" button is used to exit PPFLOW.

**Flow Mapping**

The "Flow Mapping" button in the menu bar is for flow mapping. The user must load flow data file and boundary file before any attempts to map flows, otherwise, a message that reminds the user to load data files will pop up.

After the "Flow Mapping" button is clicked, the flow map menu (Figure 51) appears on screen. There are file push-buttons in this menu. The "O/D Selection" button is for selecting origin and destination regions. A region is selected by clicking any point within the boundary of the region. The number of regions that can be selected is not limited. Double clicking finishes a O-D selection process. If the user wants to add more regions to the selection later, he/she should not terminate (double-clicking) the selection process. To unselect the regions, the user shall move the cursor to a place that is not within any of the regions on the map and then double click any button on the mouse.

The "Flow Variable Selection" button selects a flow variables to be mapped onto the flow map. After this
button is clicked, the flow variable selection menu (Figure 52) appears on screen. This menu contains all the flow variable existed in the flow data file, plus two PP-generated flow variables PP_FLOW1 and PP_FLOW2. When a variable is clicked, it will be highlighted. After the desired variable is highlighted, the user clicks the "OK" button to finish the variable selection process. Only one flow variable can be selected for flow mapping. If multiple variables are highlighted, the one closer to the top of the menu is selected.

The "color" button is used to change the color for flow mapping. When this button is clicked, the color selection menu (Figure 53) pops up on screen. An array of colors are available for selection. The user simply click the one that he/she wants. The default color is set to black. The selected color is used to display not only flows, but also the title, which is the description of the selected flow variable.

When the "Go" button is clicked, the flows of the selected flow variable between the selected origin and destination regions are displayed on the flow map in the selected color. Two-directional flows are displayed.

The "Clear" button is designed to clear the flow map. After the flow map is cleared, the color is reset to the default color, which is black. The flow map can be displayed again by clicking the "Go" button.

Figure 53. Color selection menu.
Flow Representation

The "Flow Representation" button is the menu bar is for setting flow representation methods. There are three methods implemented in PPFLOW. One is the proportional width method. A line with an arrow at end connects an origin and a destination, and the width of the line is proportional to the magnitude of the flow. The second is the quantile width method, in which the width of a line is determined by relative magnitude of a flow compared to the other flows. The third is called the line and number. A line connect the origin and destination of a flow with an arrow at end pointing the direction of the flow. The magnitude of the flow is displayed near the middle point of the line.

When the "Flow Representation" button is clicked, the flow representation selection menu (Figure 54) appears on screen. The "Proportional Width" button is for the proportional width method, the "Quantile Width" button is for the quantile width method, and the "Line and Number" button is for the line and number method. The flow representation method can be changed at any time during the operation of PPFLOW. To apply the newly selected flow representation to the flows already displayed on the flow map, the user simply click the "Go" button in the flow map menu (Figure 51).

Choropleth Mapping

When the "Choropleth Mapping" button is clicked on the menu bar, the choropleth map menu (Figure 55) appears on the screen. This menu has five buttons. The first one "O/D Selection" is not functional at this moment.
The "O/D Attribute Selection" button is for selecting an O-D attribute variable to be mapped onto the choropleth map. When this button is bushed, the O-D variable selection menu (Figure 56) pops up on screen. This menu contains all the O-D attribute variables existed in the O-D attribute data file, plus two PP-generated attribute variables. When a variable is clicked, it will be highlighted. After the desired variable is highlighted, the user clicks the "OK" button to finish the variable selection process. Only one attribute variable can be selected for choropleth mapping. If multiple variables are highlighted, the one closer to the top of the menu is selected.

The "color" button is used to select the color for displaying the legend and title, which is the description of the selected attribute variable. When this button is clicked, the color selection menu (Figure 53) pops up on screen. An array of colors are available for selection. The user simply click the one that he/she wants. The default color is set to black.

When the "Go" button is clicked, the choropleth map of the selected flow variable are displayed. Currently, the choropleth map is displayed in five classes, each is represented by a red color with varying darkness. The darker the red color is, the higher the value of the selected variable.
The interval of each class is displayed in the legend which is located at lower-right corner of the drawing area for flow map/choropleth map. If the flow map already exists on the screen, the choropleth map will overwrite the flow map. To bring the flow map back on screen, the user shall push the "Flow Mapping" button on the menu bar and then click the "Go" button in the flow map menu (Figure 51).

The "Clear" button is designed to clear the choropleth map, as well as the flow map if it is displayed. After the map(s) is cleared, the color is reset to the default color, which is black. The choropleth map can be displayed again by clicking the "Go" button on the choropleth map menu (Figure 55), and the flow map can be brought back by clicking the "Go" button on the flow map menu (Figure 51).

**Statistical Plot of Flow Variables**

The "Plot Flow Variables" button on the menu bar is designed to draw statistical plots such as scatterplots and histograms of flow variables. Clicking this button brings up the flow variable plot menu (Figure 57). The first button "O/D Selection" is not functional at this moment.

Clicking "Flow Variable Selection" on the flow variable plot menu brings up the flow variable selection menu (Figure 52). This menu contains all the flow variables existed in the flow data file, plus two PP-generated flow variables PP_FLOW1 and

![plot_flow_menu](image)
PP_FLOW2. When a variable is clicked, it will be highlighted. After the desired variables are highlighted, the user clicks the "OK" button to finish the variable selection process.

The "color" button is used to select the color for drawing statistical plots. When this button is clicked, the color selection menu (Figure 53) pops up on screen. An array of colors are available for selection. The user simply click the one that he/she wants. The default color is set to black.

The "Go" is designed to draw statistical plot. When only one flow variable is selected, the histogram of the variable is displayed. If multiple variables are selected, a scatterplot triangle of the selected variables is displayed.

Clicking the "Clear" button clears the statistical plot, and set the color back to the default color, which is black. The statistical plot can be brought back by clicking the "Go" button on the flow variable plot menu (Figure 57).

**Statistical Plot of O-D Attribute Variables**

The "Plot O/D Attributes" button on the menu bar is designed to draw statistical plots such as scatterplots and histograms of O-D attribute variables. Clicking this button brings up the attribute variable plot menu (Figure 58). The first button "O/D Selection" is not functional at this moment.

Clicking "O/D Attribute Selection" on the attribute variable plot menu brings up the attribute variable selection menu (Figure 56). This menu
contains all the attribute variables existed in the flow data file, plus two PP-generated attribute variables PP_ATT1 and PP_ATT2. When a variable is clicked, it will be highlighted. After the desired variables are highlighted, the user clicks the "OK" button to finish the variable selection process.

The "color" button is used to select the color for drawing statistical plots. When this button is clicked, the color selection menu (Figure 53) pops up on screen. An array of colors are available for selection. The user simply click the one that he/she wants. The default color is set to black.

The "Go" is designed to draw statistical plot. When only one variable is selected, the histogram of the variable is displayed. If multiple variables are selected, a scatterplot triangle of the selected variables is displayed.

Clicking the "Clear" button clears the statistical plot, and set the color back to the default color, which is black. The statistical plot can be brought back by clicking the "Go" button on the attribute variable plot menu (Figure 58).

**Projection Pursuit**

The projection pursuit (PP) method is implemented in PPFLOW to reduce the number of variables (both flow variables and O-D attribute variables). When the "Projection Pursuit" button is pushed on the menu bar, the projection pursuit menu (Figure 59) appears on screen. This menu contains eight push-buttons. The "O/D Selection" button, the "Single Flow: R-Mode" button and the "Single Flow: Q-Mode" button are not functional at this moment.
**PP on Flow Variables**

To reduce the number of flow variables, the user should select flow variables by clicking the "Flow Variable Selection" button, which brings up the flow variable selection menu (Figure 52). This menu contains all the flow variables existed in the flow data file, plus two PP-generated flow variables PP_FLOW1 and PP_FLOW2. When a variable is clicked, it will be highlighted. After the desired variables are highlighted, the user clicks the "OK" button to finish the variable selection process. At least two variables must be selected because PP cannot be performed on a single flow variable.

The next step is to select a variable name to save PP result. The user shall click the "Save PP Result" button, which pops up the flow PP save selection menu (Figure 60). This menu contains two variables PP_FLOW1 and PP_FLOW2. When a variable is clicked, it will be highlighted. After the desired variable is highlighted, the user clicks the "OK" button to finish the variable selection process. The user can choose any one of the two variables to save the result of PP.
Then, the user shall select initial direction for PP optimization. After the "Assign Initial PP Direction" on the projection pursuit menu (Figure 59) is pushed, the flow PP direction selection menu (Figure 61) appears on screen. In Figure 61, there are three flow variables MIGRAT60, MIGRAT70 and MIGRAT80. A slide bar is created to adjust the directional value for each variable. The directional value can be adjusted between 1 and 100. When the directional values are set for all the variables, clicking the "OK" button will end the direction selection process.

The final step is to click the "Go" button on the projection pursuit menu (Figure 59) to start the PP optimization. The details of each iteration of the optimization are displayed in the text window, from which the PPFLOW was started.

**PP on O-D Attribute Variables**

To reduce the number of O-D attribute variables. The user should select attribute variables by clicking the "O/D Attribute Selection" button, which brings up the attribute variable selection menu (Figure 56). This menu contains all the attribute variables existed in the attribute data file, plus two PP-generated attribute variables PP_ATT1 and PP_ATT2. When a variable is clicked, it will be highlighted. After the desired variables are highlighted, the user clicks the "OK" button to finish the variable selection process. At least two variables must be selected because PP can not be performed on a single variable.
The next step is to select a variable name to save PP result. The user shall click the "Save PP Result" button, which pops up the attribute PP save selection menu (Figure 62). This menu contains two variables PP_ATT1 and PP_ATT2. When a variable is clicked, it will be highlighted. After the desired variable is highlighted, the user clicks the "OK" button to finish the variable selection process. The user can choose any one of the two variables to save the result of PP.

Then, the user shall select initial direction for PP optimization. After the "Assign Initial PP Direction" on the projection pursuit menu (Figure 59) is pushed, the attribute PP direction selection menu (Figure 63) appears on screen. In Figure 63, there are three flow variables UNEMPLOY, CRIME and GNP/PC. A slide bar is created to adjust the directional value for each variable. The directional value can be adjusted between 1 and 100. When the directional values are set for all the variables, clicking the "OK" button will end the direction selection process.

The final step is to click the "Go" button on the projection pursuit menu (Figure 59) to start the PP optimization. The details of each iteration of the optimization are displayed in the text window, from which the PPFLOW was started.
Dynamic Brushing

Dynamic brushing is designed to link multiple views such as flow map, O-D attribute choropleth map, statistical plots of flow variables and attribute variables in a real-time fashion. It can facilitate the user to examine the relationships among these views; thus better revealing the relationships between spatial flows and regional characteristics. Based on these relationship, new spatial flow patterns may be generated.

Clicking the "Brush" button in the menu bar brings up the brushing menu (Figure 64). This menu contains four push-buttons. The first two "Brush Shape" and "Brush Size" are not functional at present time. The brush is set to a square by default. The "Go" button initiate the dynamic brushing, while the "Stop Brushing" button terminates the brushing.

Dynamic brushing can be initiated from any of the above described views. When brushing is started from a scatterplot in a scatterplot triangle of flow variables, a rectangular brush appears and the data points covered by the brush are highlighted. The corresponding data points in other scatterplots of the scatterplot triangle are also highlighted. In addition, the corresponding flows are displayed on flow map. The attributes of regions that are associated with the highlighted flows are also highlighted on histogram or scatterplot triangle of O-D attribute variables. When brushing a histogram of a flow variable, the user should move the brush along horizontal axis.
When brushing is started from a scatterplot in a scatterplot triangle of O-D attribute variables, a rectangular brush appears and the data points covered by the brush are highlighted. The corresponding data points in other scatterplots of the scatterplot triangle are also highlighted. In addition, flows that connect highlighted regions are displayed on flow map, and highlighted on histogram or scatterplot triangle of flow variables. When brushing a histogram of an O-D attribute variable, the user should move the brush along horizontal axis.

Brushing flow map/choropleth map can be carried out in two different ways: one with pre-selected regions, known as anchor regions; the other without anchor regions. In the presence of anchor regions, flows between a brushed region and anchor regions are displayed on flow map, and these flows are highlighted on histogram or scatterplot triangle of flow variables. The attributes of the brushed region will be highlighted on histogram or scatterplot triangle of O-D attribute variables. Without anchor regions, brushing will not bring anything new on flow map, and histogram or scatterplot triangle of flow variables, only the attributes of the brushed region will be highlighted on histogram or scatterplot triangle of O-D attribute variables.

**Query by O-D Distance and Flow Magnitude**

In PPFLOW, the content of the flow map can be changed by adjusting distance and flow magnitude. On the lower-left corner of the interface of PPFLOW (Figure 28 and Figure 30) are four slide bars, which are enlarged and displayed in Figure 65. The two slide bars on the left are for flow magnitude adjustment, and the two on the right are for distance adjustment. The minimum and maximum distances on the slide bars are set when the
boundary data file is loaded. The minimum and maximum flow magnitudes on the slide bars are set when a flow variable and at least two regions are selected and the "Go" button is clicked to display flows. After the minimum and maximum values are set on the slide bars, the user can change the content of the flow map by adjusting the values on the slide bars.

The O-D distance adjustment can be applied to either the whole data set, or a set of pre-selected regions, known as anchor regions. The anchor regions are selected by following the procedure applied to the O/D selection in flow mapping. In the former case, all flows that are within the specified distance range will be displayed. In the later, only those flows that are connected with the anchor regions and are within the distance range will be displayed. In both cases, the flows displayed on flow map are also highlighted in histogram or scatterplots of flow variables, and the attributes of regions that are associated with these flows are highlighted in histogram or scatterplots of O-D attribute variables. All these actions are instantly executed at the same time as the adjustment is made on the slide bar.

Flow magnitude adjustment can also be applied to either the whole data set, or the pre-selected anchor regions. The anchor regions are selected by following the procedure applied to the O/D selection in flow mapping. In the former case, all flows that are within the specified magnitude range will be displayed. In the later, only those flows that are...
connected with the anchor regions and are within the range will be displayed. In both cases, the flows displayed on flow map are also highlighted in histogram or scatterplots of flow variables, and the attributes of regions that are associated with these flows are highlighted in histogram or scatterplots of O-D attribute variables. All these actions are instantly executed at the same time as the adjustment is made on the slide bar.
This appendix illustrates how the first derivative of the moment index is derived. Because the third and fourth order moments are similar, only the derivative of the third moment is derived here. The derivative of the fourth order moment can be worked out in a similar way.

Following Equation (10), we have the third order moment represented as

\[ m_3(X_{XK}, \mathbf{p}) = \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{k=1}^{K} pp_i p_j T_{ik} \]  

(19)

The first partial derivative of \( m_3(X_{XK}, \mathbf{p}) \) to \( p_s \) the \( s \)th (1 \( \leq \) \( s \leq K \)) element of vector \( \mathbf{p} \), is

\[
\frac{d}{dp_s} m_3(X_{XK}, \mathbf{p}) = \frac{d}{dp_s} \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{k=1}^{K} pp_i p_j T_{ik} p_s T_{ik} \\
= \sum_{j=1}^{K} \sum_{k=1}^{K} pp_j \frac{d}{dp_s} p_i T_{ik} \\
+ \sum_{i=1}^{K} \sum_{k=1}^{K} pp_i \frac{d}{dp_s} p_j T_{ik} \\
+ \sum_{i=1}^{K} \sum_{k=1}^{K} pp_{k} \frac{d}{dp_s} p_i T_{ik} \\
(20)
\]
Recall that the vector $p$ is required to be an unit vector in projection pursuit optimization; hence, $p_i$ should be normalized as

$$p_i = \frac{p_i}{\sqrt{\sum_{n=1}^{K} p_n^2}}$$

(21)

So,

$$\frac{d}{dp_i} \sum_{i=1}^{K} p_i T_{ik} = \frac{d}{dp_i} \left( \frac{p_i}{\sqrt{\sum_{n=1}^{K} p_n^2}} \right) T_{ik}$$

$$= \frac{K}{\sum_{n=1}^{K} p_n^2} \frac{d}{dp_i} \left( \frac{p_i}{\sqrt{\sum_{n=1}^{K} p_n^2}} \right) T_{ik}$$

$$= \frac{K}{\sum_{n=1}^{K} p_n^2} \frac{d}{dp_i} \frac{p_i}{\sqrt{\sum_{n=1}^{K} p_n^2}} T_{ik} + \frac{\frac{d}{dp_i} p_i}{\sqrt{\sum_{n=1}^{K} p_n^2}} T_{ik}$$

$$= \sum_{i=1}^{K} \frac{-p_i \sum_{n=1}^{K} p_n^2}{2 \sum_{n=1}^{K} p_n^2} + \frac{1-p_i^2}{\sqrt{\sum_{n=1}^{K} p_n^2}} T_{ik}$$

(22)

In the above equation, the length of the vector $p$ is assumed to be 1 because $p$ is normalized in every iteration during the optimization process. Similarly, we have
Substitute Equations (22), (23) and (24) back to Equation (20), we have

\[ \frac{d}{dp} m_s(X_{km\alpha}) = \sum_{i=1}^{K} \sum_{k=1}^{K} p_{i,k} \left( \sum_{j=1}^{K} \left( -p_{j,k} T_{j,k} \right) + T_{h,k} \right) \]

The above equation is exactly the same as Equation (15).
BIBLIOGRAPHY


187


Wentz, E. 1990. M.A. research in Geography Department at The Ohio State University.