THE TREATMENT OF NEGLECTED HETEROGENEITY

IN CROSS-SECTIONAL DATA SETS

BY
DI MENG

A DISSERTATION PRESENTED TO THE
GRADUATE FACULTY OF MIDDLE TENNESSEE STATE UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF ARTS

AUGUST, 2000

UMI Number: 9978698

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

THE TREATMENT OF NEGLECTED HETEROGENEITY

IN CROSS-SECTIONAL DATA SETS

APPROVED:

Joachim Zietz

Major Professor

EA Eff

Committee Member

James Huffman

Committee Member

John T. Lee

Chairman of the Department of Economics and Finance

Donald L Curry

Dean of the College of Graduate Studies

ABSTRACT

The Treatment of Neglected Heterogeneity In Cross-Sectional Data Sets

By Di Meng

This study is to show how one can apply classification analysis to pure cross-sectional data in economics to build up models that incorporate behavioral differences and that, therefore, allow for more accurate and efficient policy applications. The study looks into both the problems that typically arise when behavioral heterogeneity is neglected in the empirical estimation process and a possible solution to the problem of identifying viable groups of economic actors with homogeneous behavioral response patterns.

The study is not aimed at providing a general method that is applicable for every possible cross-sectional data set and/or every possible heterogeneity pattern. Rather, it demonstrates, for a particular economic example of interest in regional economics, how one can take advantage of a variety of multivariate procedures in the general area of classification analysis to help reveal the nature of heterogeneity and to explore possible avenues to detect homogeneity.

In this study, the testing methodology (Zietz, 2000) for the problem of neglected heterogeneity is successfully implemented in practice. The test results confirm the existence of the problem of neglected heterogeneity for the original unclassified cross-section data used in this study.

This study is able to develop an objective and effective classification methodology for a given specific data set to discover homogeneous subgroups with similar individual characteristics that are related to their economic behavior.

As demonstrated by this study, one can establish a close and economically meaningful relationship between group-specific economic behavior and the individual characteristics of the economic agents in each group. Such relationships should be of significant economic value. For example, economic policies that try to target specific groups need exactly this type of group-specific information. It is also valuable for such issues as deriving forecasts for the aggregate of all observations. Specifically, the results provide not only useful weights that can be used to aggregate the subgroups but also the different behavioral patterns that need to be aggregated.

The procedures developed in this study apply to many research areas outside of economics. Specifically, they apply whenever one is facing the problem of neglected heterogeneity. The application of the methodology in the area of economic education is illustrated as an example in the study.

The study can be used by the readers who are interested in homogeneity-heterogeneity topics as a learning tool because all steps are carefully laid out and discussed.

Readers interested in potential applications of the techniques in the field of educational research are provided with a number of suggestions on how the methodology can be of potential use.

ACKNOWLEDGMENTS

A study of this kind could not be accomplished without the help of many people. I would like to thank several of those who have contributed greatly.

First, I would like to thank my family. My wife, Jia Li, deserves great credit for the final product. She was very loving and supportive, and always encouraged me to finish the work whenever I was carried away from the study. My son, Yang, has also been very helpful with sacrificing his game time to give me access to the computer.

I would like to thank the three members of my outstanding dissertation committee. First, I am grateful to the supervisor, Dr. Joachim Zietz, for his original ideas, careful reading and in-detail comments and correction. Dr. Anthon Eff also made substantial contributions to this study, and was especially helpful in establishing the economic model and obtaining the data. Dr. James Huffman is much appreciated for his reading and comments.

I would also like to thank Dr. John Lee, Chairman of the Economic and Finance Department, and Dr. Reuben Kyle, Director of Business and Economic Research Center, for giving me the opportunities as a graduate assistant to obtain precious practical knowledge in economic research. Also, thanks to Dr. Duane Graddy for his professional training on microeconomic theories and practice.

Finally, I would like to thank my colleagues, Charlie Bastnagel and Ed Novak, for taking so much workload from my shoulders while I was doing this research. Also, thanks to Jerry Bearden for proofreading the paper and giving a professional polish to the final product.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

A modern economy presents a picture of millions of people, either as individuals or organized into economic units -- households, groups and firms, each pursuing their own disparate interests in a limited part of the environment. In this process, the varied individual activities lead to behavioral heterogeneity in the economy as a whole. There are no studies of disaggregated, micro level data that fail to find strong systematic evidence of individual differences in economic behavior (Stoker, 1993).

A typical example is the demographic differences of households and the differences in behavior associated with that. Firms spend considerable effort on strategic marketing, which means they are trying to take advantage of systematic differences among potential customer groups. One can even think of the differences in learning characteristics of different student groups as an example of behavioral heterogeneity. The existence of these kinds of differences has raised concerns over the treatment of behavioral heterogeneity in econometrics. In particular, there is a growing awareness that simply ignoring behavioral heterogeneity raises serious problems of inference. This is all the more true as it has become apparent that the problem of neglected heterogeneity is much more widespread in cross-sectional data sets than has previously been thought.

In his general survey study, Stoker (1993) relates the "problem of aggregation over individuals" to the issue that aggregate economic models are constructed on the basis of incomplete information about heterogeneous individual behavior patterns. In fact, individual heterogeneity and the composition of the population are left out of typical macro models altogether, therefore quite possibly leading to meaningless estimates and misleading prediction for economic policy making. Blundell, Pashardes, and Weber (1993) developed a complete consumer demand system based on a time series of individual household data and use it to measure the biases introduced into the study of consumer demand behavior when aggregate data are used in place of the appropriate microeconomic data. In their research, they find that the aggregate model neglects information on time-varying household characteristics and therefore has a much worse forecasting performance. Other than including certain distributional measures in an aggregate model they suggest that it may be worthwhile to develop microeconomic models that identify groups of economic agents, each with clearly identified common observable characteristics.

There are many reasons for being concerned with testing for and identifying heterogeneity in economic practice. One can target policy better to those groups in need. One can better predict what will happen to the aggregate economy if the relative weights of certain underlying groups are changing, for example, due to demographic effects. One is better able to predict the effect of certain policies on individual groups, rather than just on aggregate outcomes.

1. Statement of the Problem

The problem of neglected heterogeneity has long been troublesome for economic researchers. The problem would not exist if one had measurable variables on all relevant aspects of the behavior or economic agents. Most often, in econometric practice, some pertinent aspects of economic behavior are not captured by an econometric model, either because they are not observable or measurable, or because they are not captured by the underlying economic theory. As a consequence, neglected heterogeneity may arise. (Zietz, 2000a)

Traditionally, individual heterogeneity has been ignored and data have been analyzed on the assumption that all observations come from the same underlying population. This has been the typical procedure for forecasting and often also for obtaining insight into economic activities for the analysis of economic policy. Heteroskedasticity that shows up in the associated empirical models is treated as a specification problem of modeling aggregate data. The typical solution is to apply a mechanical statistical correction for heteroskedasticity in the hope that the observed heteroskedasticity is not the outgrowth of a deeper underlying problem, such as neglected heterogeneity or wrong functional form. With growing concerns over the potential severity of the problem, researchers have conducted numerous studies both theoretically and by empirical experiment on ways to identify neglected heterogeneity in economic models, to estimate its effect on policy making, and to search for a remedy for the problem.

1.1 Theoretical Considerations Related to Aggregation over Individuals

As a good example of many empirical studies that have been done in the past decades, the "representative agent" approach assumes that aggregate choice coincides with the one of a representative individual. This model ignores the problem of aggregation over individuals completely. Kirman (1992) argues that this position is untenable for the following reasons.

Whatever the objective of the modeler, there is no plausible formal justification for the assumption that the aggregate of individuals, even if a maximization calculus is applied to this aggregate, acts itself like an individual maximizer. There is simply no direct relation between individual and collective behavior.

Even under the assumption that such a representative agent exits, the reaction of the representative agent to some change in a parameter of the original model – a change in government policy, for example – may not be the same as the aggregate reaction of the individuals he/she "represents." This casts serious doubt on the validity of using such models to analyze the consequences of policy changes.

The sum of the behavior of economic agents with very simple behavior patterns can generate complicated aggregate behavior. When the data from the simple sum of individual agents are used in a model for empirical testing, one will find that it is very difficult to explain the complicated behavior and meaningless estimates that may arise from aggregate over heterogeneous agents with simple behavior.

Econometric modeling on the basis of a representative agent model amounts to a purely statistical approach to working with data series that is not well grounded in individual economic behavior. Because it misses a foundation for the practice of forcing aggregate data patterns to fit the restrictions of an individual optimization problem, the conclusion from the model may lead to biased estimates and hardly provides a convincing basis for decision-making and policy applications.

In his study, Kirman concludes that well-behaved individuals need not produce a well-behaved representative agent. The reaction of a representative agent to change need not reflect how the individuals of the economy would respond to change. The preferences of a representative agent over choices may be diametrically opposed to those of society as a whole. It is apparent from Kirman's study that the representative agent paradigm has severe limitations. Only if one is prepared to develop a paradigm in which individuals operate in a limited subset of the economy, are diverse both in their characteristics and the activities that they pursue, and interact directly with each other, will economics escape from the stultifying influence of the representative agent.

In his findings on the basis of a simple static demand example in his survey on the aggregation problem, Stoker (1993) points out that simplifying models down to simple averages and neglecting individual heterogeneity misses the structure inherent in existing behavioral reactions, which in turn, severely limits the usefulness of simplified models. In line with his findings, Stoker suggests that models that account for individual heterogeneity will typically not be estimable using data on economy-wide averages alone; additional data on distributional composition, or micro data on individual behavior, will need to be incorporated.

1.2. Practical Examples and Empirical Findings

Noticeable individual heterogeneity exists in many areas that are studied by economists and related disciplines. The inference problems that are introduced by ignoring such heterogeneity can be found in the empirical studies of many fields of specialization.

For example, Heckman and Sedlacek (1985) present an empirical equilibrium model of self-selection in the labor market that recognizes the existence of measured and unmeasured heterogeneous skills. They discuss the biases caused by such econometric analyses of aggregate labor market data that either ignore such heterogeneity entirely or assume homogeneous skills for workers classified by such criteria as age, race, education, and sex.

Demographic differences are very important individual characteristics that drive diversity in individual behavior, which, in turn, has a significant impact on aggregate economic behavior. In marketing research that targets the Hispanic population the Spanish language is commonly thought to be a unifying factor. But Ueltschy (1997), in her advertising effectiveness research, finds that the Hispanic audience is by no means a homogeneous market. When the importance of the Hispanic market was initially recognized, the view among many advertisers and advertising agencies was that the "best" way to reach that market was to advertise to them in Spanish. The fact that Hispanics speak different dialects and share distinct subcultural values and characteristics according to their heritage was neglected. Such a practice has failed to target differentiated markets among the Hispanic population and led to ineffective marketing outcomes.

There are many other practical examples and empirical studies that could be listed. They all make one common statement, which is that individual heterogeneity accounts for social and economic behavior, and neglecting such heterogeneity in economic research and empirical studies leads to biased results that are less than useful for the policy making process.

An increasing number of economists pay attention to the role of the composition of the economy. Empirical methodologies are being developed to capture and incorporate individual heterogeneity in econometric modeling. This study builds on the important recognition that neglected behavioral heterogeneity is an important aspect of reality and that it has a significant impact on cross-sectional data analysis. More specifically, this study is motivated by the findings of earlier studies (e.g., Chesher 1984, Zietz, 2000b) that focused on ways to identify behavioral heterogeneity in cross-sectional data sets. This study will go one step further than these studies and examine how statistical techniques centered around classification analysis can be used to identify groups of observations that display behavioral homogeneity, once statistical tests show that neglected heterogeneity may be a problem.

2. Contribution by the Study

Classification analysis is explored as a possible tool to go beyond the finding of behavioral heterogeneity in the data set and to identify groups of economic actors that are characterized by behavioral homogeneity.

The methodology of classification analysis is being applied – although it may be named differently - in a variety of scientific fields such as biology, psychology, and areas closely related to economics, such as marketing.

For example, in today's marketing education and the textbooks about marketing strategy and consumer behavior, one can easily identify the topic of market segmentation. Market segmentation is closely related to classification analysis. If all consumers were alike, if they all had the same needs, wants, and desires, as well as the same background, education, and experience, mass (undifferentiated) marketing would be a logical strategy. But market researchers claim that changes in social and economic environment, such as the growth of population, the growth in individualism, increased competition, technology development and so on, make mass marketing impossible. Having a good product or service is no longer sufficient. Companies must satisfy the discriminating customers with specific products or promotional appeals. To do so, segmenting the market into homogeneous subsets is necessary.

Unlike the mass marketing strategy, which treats all individuals in the market the same, segmentation research emphasizes different socioeconomic factors associated with individual activities to discover meaningful ways to divide them into distinct groups, and therefore provide a conceptual foundation for incorporating compositional heterogeneity in aggregate analysis.

In economic research, since traditional representative agent analysis fails to provide foundations to fit realistic conditions, many efforts, such as exact aggregation and micro simulation models (e.g. Lewbel, 1989, Stoker, 1993), have been constructed to try to find methods of incorporating individual differences into models. Classification analysis is an alternative methodology with a number of potential advantages that has been employed in different research areas of economics. In an empirical study of urban Los Angeles based on a Tiebout model, which implies the prediction that communities will be relatively homogeneous with respect to the preferences of their residents for the economic and non-economic local public goods, Heikkila (1995) uses factor analysis and cluster analysis to examine a broad range of socio-economic data taken from the 1990 census rather than focusing on a single variable such as income to give more insight into the outcome of the process of how communities end up homogeneous. The same statistical procedures have been employed by Berlage and Terweduwe (1988) to classify countries into categories in terms of different developing stages based on country characteristics. Compared with the classification based on subjective criteria by international agencies, their study provides clearer insight into each country's background and more suggestions for international policy making.

The key advantage of classification analysis is the fact that it has the potential to condense large amounts of seemingly disparate information at the level of the individual economic unit in ways that may be useful for the identification of behavioral patterns. Information about distinct behavioral patterns can be of significant value for policy makers that try to focus their efforts on one or more distinct groups of individuals. But it can also be of significant use for policy makers that focus on aggregate outcomes: information about the behavior patterns of large groups of individuals would make it possible to predict more accurately what is likely to happen at the aggregate level if the information about group behavior is combined with information about the demographic weights of these groups.

3. Purpose of the Study

The purpose of this paper is to show how one can apply classification analysis to pure cross-sectional data in economics to build up models that incorporate behavioral differences and that, therefore, allow for more accurate and efficient policy applications. The study will look into both the problems that typically arise when behavioral heterogeneity is neglected in the empirical estimation process and a possible solution to the problem of identifying viable groups of economic actors with homogeneous behavioral response patterns.

The study is not trying to provide a general method that is applicable for every possible cross-sectional data set and/or every possible heterogeneity pattern. Rather, it will demonstrate, for a particular economic example of interest in regional economics, how one can take advantage of a variety of multivariate procedures in the general area of classification analysis to help reveal the nature of heterogeneity and to explore possible avenues to detect homogeneity. Various methods and strategies are introduced in plain English, their advantages and disadvantages are examined, and their uses are discussed for identifying homogenous groups.

The study can be used by readers who are interested in homogeneity-heterogeneity topics as a learning tool because all steps are carefully laid out and discussed.

Readers interested in potential applications of the techniques in the field of educational research are provided with a number of suggestions on how the methodology can be of potential use.

4. Organization of the Study

This study develops several production functions based on a cross-section data set from the 1982 manufacturing census and 1980 population census of United States. The data set consists of all US counties and it is expected that significant differences exist among these counties in terms of socio-economic background and production behavior. The production functions are employed to explore the existence of neglected heterogeneity through econometric analysis and to measure the bias that is introduced into the study of production behavior when the whole data set is used in place of the appropriately classified data sets. The study assesses the suitability of the classified data sets versus the unclassified data set in terms of numerous statistical tests.

This paper is organized into six chapters, and follows the general pattern that is detailed below.

Chapter 1 is the introduction of this study.

Chapter 2 introduces the research methodology that is employed by this study. It explains the procedures that are developed in this study and discusses the economic model and the data that are being employed.

Chapter 3 reviews the methodology of cluster analysis and related issues. It is discussed how cluster variables are selected and how the data are pre-processed. As an important procedure to prepare for cluster analysis, the method of factor analysis is introduced and discussed in detail. Different clustering algorithms are described by reviewing a broad range of research applications found in the literature. It is discussed why the combination of two algorithms may be useful for this study.

Chapter 4 presents the results from the cluster analysis process and discusses how to carry out in practice the methodology that is described in Chapter 3. The detailed results of the cluster analysis are presented with figures and charts. This chapter also demonstrates how to determine and validate the final solution of clusters.

Chapter 5 discusses the empirical estimates for both the data set that neglects heterogeneity and the data set that is organized by subgroup. After obtaining the regression results for the undifferentiated data set and the one organized by subgroup, the study provides some insight into the issue of how the characteristics of production behavior vary with the characteristics of the counties in which the producing entities are located.

Finally, the study describes how neglected heterogeneity may be an issue for economic education studies that use large unclassified data sets. It is suggested that these types of studies may well be ideal testing grounds for the classification methodology developed in this study.

CHAPTER 2

RESEARCH METHODOLOGY

This study is comparative in nature. Estimates on unclassified data are compared with those on data classified by subgroup, where the subgroups are the outcome of the classification analysis. U.S. counties serve as the individual observation units in the unclassified data set. The data are subdivided into several homogeneous groups based upon such characteristics as county population, size, and educational level. Production function models are built and analyzed based on classification data analysis, standard econometric adequacy tests, and the economic implications for both the unclassified data set and the data sets that are classified into subgroups. The following specific procedures are used:

1. Estimation Procedures

1). An economic model is constructed and converted into a format that lends itself to estimation by regression techniques. Ignoring potential behavioral heterogeneity, this model is estimated by ordinary least squares (OLS) on all observations of the data set. Standard statistical adequacy tests are conducted to check whether the regressions are consistent with the statistical assumptions underlying OLS.

There are two principal tasks to be accomplished in this first step.

i). Particular attention is paid to those statistical tests that are potentially associated with neglected parameter heterogeneity or that can differentiate neglected parameter heterogeneity from other regression problems. Because neglected parameter heterogeneity tends to show up in significant test statistics for heteroskedasticity and wrong functional form (Zietz, 2000b), this study applies the Breusch-Pagan (1979) LM test to check for heteroskedasticity and Ramsey's (1969) Reset test to identify a potential problem with wrong functional form. The combination of these tests can be used to identify the problem of neglected parameter heterogeneity,[1] which is one of the important purposes of this study.

ii). To avoid multicollinearity and to achieve more efficient parameter estimates, the general-to-specific specification strategy (Gilbert, 1986, p.283-307) is employed to find the most parsimonious regression model that is consistent with the data at a common probability value (p-value).

2). A very important step in this study is classification analysis, which divides the population into several subgroups that have as many common characteristics as possible. The major premise is that a subgroup perspective leads to a more precise definition of behavioral characteristics of individual economic units within groups and the differences of behavioral characteristics across groups. It thus provides a better understanding of what shapes individual behavior patterns.

As a special statistical tool, cluster analysis is employed for the classification procedure. Cluster analysis, also known as classification analysis or numerical taxonomy (Everitt, 1974, p.1-5), is a broad field spanning many disciplines. The common element to cluster related issues is that a large number of observations are measured on a number of variables, and the observations are grouped. Based on the similarity of individual characteristics, cluster analysis will split the unclassified data set into groups that can be used for the next step in the statistical analysis.

[1] See Zietz (2000) for a detailed discussion of these points.

To successfully carry out this step, three requirements need to be met:

i) there must be clear differences between groups; they need to be internally homogeneous and externally heterogeneous;

ii) the groups must be reasonably large in size to generate statistically meaningful results;

iii) it must be possible to reach different groups; otherwise, the results have no policy implication.

It is hoped that the data set that is utilized for this study allows these requirements to be met in principle. The data are discussed in the following section.

3). Given the subgroup data sets that are obtained from the cluster analysis described in step 2, the empirical model that is initially fit to the unclassified data is re-estimated and analyzed in the same way on the new group-specific data sets. Standard statistical adequacy tests are applied to these group-specific models, including all those tests that have the potential to identify the existence of neglected parameter heterogeneity. Assuming that the classification analysis works as expected, the evidence for neglected behavioral heterogeneity should be much less for the group-specific data sets than for the unclassified data set.

4). Based on the estimation results, the study compares the results for the unclassified data set to those derived from the group-specific data in terms of statistical fit, modeling effectiveness, and policy applicability. The usefulness of the cluster analysis will also be discussed with the hope to identify a classification procedure for future studies of a similar type.

2. The Economic Model and Data

The empirical procedures in this study start with constructing and estimating a simple log-linear Cobb-Douglas function.

$$\ln(Q_i) = \alpha_0 + \sum_j \alpha_j * \ln(X_{ji}) + e_i$$

where Q_i is manufacturing value of shipments in county i; X_{ji} is amount of input j used in county i; α_0 is the constant term; $\Sigma\alpha_j$ is a 1xj vector of coefficients; and ε_i is an iid error term.

The estimate of this simple log-linear Cobb-Douglas production function serves an important purpose. When one tries to distinguish the problem of neglected parameter heterogeneity from that of wrong functional form, one may want to compare the statistical tests for heteroskedasticity and for wrong functional form for this simple form to those from an equation with higher powers added. If heteroskedasticity and wrong functional form remain a significant statistical problem, one should seriously consider that neglected parameter heterogeneity is the underlying problem and not wrong functional form (Zietz, 2000b).

The Cobb-Douglas production function is only estimated with the unclassified data set to illustrate this process of identifying neglected parameter heterogeneity. The economic model that is used in the subsequent analysis for both the unclassified data set and the subgroup data sets is a physical translog production function of the following form:

$$\ln(Q_i) = \alpha_0 + \sum_j \alpha_j * \ln(X_{ji}) + \sum_j \sum_{k \geq j} \alpha_{jk} * \ln(X_{ji}) * \ln(X_{ki}) + e_i$$

where Q_i is manufacturing value of shipments in county i; X_{ji} is amount of input j used in county i; X_{ki} is amount of input k used in county i; α_0 is the constant term; $\Sigma\Sigma\alpha_{jk}$ is a kxj vector of coefficients; and ε_i is an iid error term.

This empirical study is an extension of the research that has been done by Eff (1995). In his research, Eff employed the translog model and used data from the manufacturing census of 1982 and 1987 to study urban-rural differences in manufacturing technology among US counties.

The data used in this study are drawn from the manufacturing census of 1982 that is included in Eff's study and from the population census of 1980. The industry reports and State reports of the manufacturing census data include such statistics as number of establishments, employment, payroll, value added by manufacturer, cost of materials consumed, capital expenditures, product shipments, etc. The statistics are presented for each State and its important metropolitan statistical areas (MSA's), counties, and places. Observations in both data sets number over 3,100 individual U.S. counties. However, data suppression for small counties reduced the total to 1,897 counties with complete data (Eff, 1995).[2]

In Eff's study, the model is estimated using a translog production function, as given above, for four different sets of observations (Eff, 1995):

1) 618 MSA counties, 1982

2) 1279 non-MSA counties,1982

3) 618 MSA counties, 1987

4) 1279 non-MSA counties,1987

The study conducted by Eff (1995) raised several interesting points that make it well suited as a testing bed for exploring the issues surrounding neglected heterogeneity:

1) Apparent individual heterogeneity of the objects - counties in the study.
2) Some classification and subgrouping was done in the study, but in a subjective manner. Besides, only two subgroups were considered, geographical location and population.
3) The comparative analysis in Eff's research provides guidance and a tool for the analytical methodology employed in this study.

There are two separate categories of variables in the data set. Seven original variables (Table 2-01) from the manufacturing census are used to create the log transformed variables to estimate the model. Eight other variables that are derived from the two censuses are used to perform the cluster analysis. These classificatory variables shown in Table 2-03 describe the various demographic background of the counties such as average educational level, population and location as well as certain significant characteristics in the manufacturing sector such as production complexity and median firm size. These classification variables are discussed in a later chapter when cluster analysis is introduced. The basic statistics for the two categories of variables are presented in Table 2-02 and Table 2-04 respectively.

[2] Counties in which there are three or fewer establishments, or counties in which on establishment accounts for 80 percent or more of employment, income or payroll, are not reported. (U.S. Bureau of the Census, 1982)

Table 2-01 Variables Used In the Empirical Study

Variables	Descriptions
VS (ML$)	Value shipped by manufacturers.
VA (ML$)	Value added by manufacturers.
PAY (ML$)	Total Salary Paid to Production and Non-Production workers
HRS (ML)	Total Work Hours of Production workers
EMP (THO)	Total employment.
PEMP(THO)	Production employment.
CM (ML$)	Cost of materials used.

Notes: ML$ is in Million Dollars; ML is in Million; and THO is in Thousand.

Table 2-02 Basic Statistics for the Regression Variables of the Unclassified Data Set

Variable	No of Obs	Mean	Std Dev	Minimum	Maximum
VS	1897	947.08	3180.54	5.80	85763.10
VA	1897	408.01	1444.19	2.40	40260.40
PAY	1897	188.93	691.74	1.50	17897.80
EMP	1897	9.51	31.94	0.20	866.10
PEMP	1897	6.13	19.16	0.10	554.10
HRS	1897	11.64	36.78	0.20	1061.70
CM	1897	536.28	1798.09	1.40	45488.30

Table 2-03 Variables Used In the Classification Analysis

Variables	Descriptions
ZABR (0-8)	Measurement of county size and location.
EDA (Year)	Education Attainment
MFSIZ (THO/ABS)	Total Employment / Total Establishment in Manufacture sector
HWAGE (ML$/ML)	Total wages divided by total hours of production employment
LSHARE (ML$/ML$)	Payment for total employment / Value Added
COMPLEXI (ML$/ML$)	Valued added / Value Shipped
PRDTY (ML$/THO)	Value Added / Production Employment
KEMP (ML$/THO)	New Capital Expenditure / Total Employment in Manufacture Sector

Notes: ML$ is in Million Dollars; THO is in Thousand; and ABS is in absolute number.

Table 2-04 Basic Statistics for the Classification Variables of the Unclassified Data Set

Variable	No. of Obs	Mean	Std Dev	Minimum	Maximum
ZABR	1897	4.34	1.92	0.00	8.00
EDA	1897	12.07	0.76	7.70	15.60
HWAGE	1897	7.55	1.98	2.94	15.76
MFSIZ	1897	55.70	34.43	7.69	372.41
LSHARE	1897	45.08	12.12	10.54	98.98
COMPLEXI	1897	42.87	11.00	8.48	85.25
PRDTY	1897	53.94	27.38	9.83	297.75
KEMP	1897	3.00	3.61	0.00	48.63

The advantage of using this data set is that it meets the basic requirements of classification study, because it has a sufficient number of observations and a variety of individual characteristics as shown in Table 2-03 for classification analysis. In addition, it appears to be fairly representative of the type of socioeconomic cross-section data for which neglected heterogeneity may play a role. Thus the methodology that we will develop on the basis of these data may be helpful for future research on similar data sets and subjects.

3. Model Specification

The simple log-linear Cobb-Douglas production function model is obtained by unfolding the equation in section 2 of this chapter.

$$\ln(Q_i) = \alpha_0 + \alpha_K * \ln(K_i) + \alpha_H * \ln(H_i) + \alpha_N * \ln(N_i) + \alpha_M * \ln(M_i)$$

As was discussed before, to identify the possible problem of neglected parameter heterogeneity and distinguish it from the problem of wrong functional form, one may need to implement a dynamic process comparing the tests for heteroskedasticity and wrong functional form obtained from a simple function form and a more general function form.

The translog production function model (Christensen and Jorgenson 1969) with more higher power variables added is introduced as a more general function form.

$$\begin{aligned} \ln(Q_i) = \alpha_0 + \alpha_K * \ln(K_i) + \alpha_H * \ln(H_i) + \alpha_N * \ln(N_i) + \alpha_M * \ln(M_i) + \alpha_{KH} * \ln(K_i)\ln(H_i) \\ + \alpha_{KN} * \ln(K_i)\ln(N_i) + \alpha_{HN} * \ln(H_i)\ln(N_i) + \alpha_{MK} * \ln(M_i)\ln(K_i) \\ + \alpha_{MH} * \ln(M_i)\ln(H_i) + \alpha_{MN} * \ln(M_i)\ln(N_i) + \alpha_{MM} * \ln(M_i)\ln(M_i) \\ + \alpha_{KK} * \ln(K_i)\ln(K_i) + \alpha_{HH} * \ln(H_i)\ln(H_i) + \alpha_{NN} * \ln(N_i)\ln(N_i) \end{aligned}$$

The variables in these models are derived from the original variables shown in Table 2-01. Q is the total value of shipment; K is a measurement to capture the capital input in the production process; H is HRS which measures the total input by production workers; N is used to measure the non-production input, which is the employment figure derived by subtracting the production employment from the total employment; and M is the cost of materials input. The symbols for the coefficients follow the format that combines α and the subscripts of the corresponding variables. For example, α_{KN} is the coefficient for the joining product of K and N.

The translog production function has been widely used in economic research because its more flexible form will help minimize any biases that might result from using the more restrictive Cobb-Douglas. The translog production function is an improvement over the Cobb-Douglas form since it allows the elasticity of substitution to vary by type of input, and the returns to scale and output elasticity to vary with the input specification (Brynjolfsson and Hitt 1995). This is a very important point for this study examines these statistics to evaluate the economic implications of the estimations. For these significant advantages, this study employs the translog production function form as the empirical model for all the data sets in question.

3.1. Identification of Neglected Heterogeneity

To identify whether neglected heterogeneity is a possibility for the given data set, this study follows the procedures suggested by Zietz (2000b) and implemented as follows. Some of the fundamental concepts have been discussed before and more detail will be given here.

Step 1: The log-linear Cobb-Douglas function is estimated first and all relevant statistical tests are conducted that relate to neglected parameter heterogeneity. In his study, Zietz (2000b) finds that when neglected heterogeneity or wrong functional form exist, the test statistics for heteroskedasticity and wrong functional form tend to be both highly significant when at least one regressor is correlated with the group-specific regression coefficients. The statistics from the estimate of the Cobb-Douglas function shown in Table 2-05 suggest the possibility of neglected heterogeneity or wrong functional form as both heteroskedasticity and wrong functional form are suggested by the test statistics at very high levels of statistical significance.

Step 2: To differentiate between neglected heterogeneity and wrong functional form, the more flexible translog function is estimated and the same statistical tests are conducted as for the Cobb-Douglas function. Zietz (2000b) points out that if neglected heterogeneity is the underlying problem then the Reset test for wrong functional form will continue to be statistically significant, even when higher powers of the regressors are added to the simple linear equation. On the other hand, if functional form is the underlying problem, the Breush-Pagan(1979) LM test is less likely to show significance than it does in the neglected heterogeneity case. Furthermore, in the wrong functional form case, adding higher powers of the regressors is most likely to eliminate the problems with heteroskedasticity and wrong functional form for any sample size. In this case, if the test statistics that are associated with the estimation of the Cobb-Douglas function are related to the use of a wrong functional form, in particular a functional form that is too simplistic, rather than to neglected behavioral heterogeneity, then one would expect that the test statistic for wrong functional form is much improved for the translog function compared to the Cobb-Douglas function. As shown in Table 2-05, the statistics on heteroskedasticity and wrong functional form remain significant at all common levels of statistical significance. We can now conclude that it is very likely that neglected parameter heterogeneity is the underlying problem and not wrong functional form.

Table 2-05. Tests for Heteroskedasticity and Wrong Functional Form

Regression	RESET Test	Breusch-Pagan LM
Cobb-Douglas	F[3, 1889] = 603.57	$X^2(4)$ = 1026.76
	P-value = 0.000	P-value = 0.000
Translog	F[3, 1879] = 18.74	$X^2(14)$ = 1475.04
	P-value = 0.000	P-value = 0.000

3.2. Model Specification

The results that are obtained from estimating the above translog model for the full data set are given in Table 2-06. The Breusch-Pagan LM Test for heteroskedasticity is significant for all common levels of statistical significance.

A simple general-to-specific procedure is used to test down this model to a more parsimonious form. The selection of which variables to drop is based on the rank of the t-values of each coefficient. In the first run, the variable with the lowest t-value is dropped and an F-test is executed with the rest of the variables. If the F-test's probability value is greater than 0.1, then the variable with the next lowest t-value is dropped, followed again by an F-test with the rest of the variables. This step is repeated until the F-test's probability value is less than 0.1. After the F-test, the last dropped variable is added back to the model.

Table 2-06. Regression with Unclassified Data Set

Variable	Coefficient	t-ratio	p-value
Constant	5.561	16.55	0.000
LK	0.227	6.17	0.000
LN	0.640	9.95	0.000
LH	0.406	6.88	0.000
LM	-0.008	-1.91	0.056
LKN	-0.008	-2.41	0.016
LKH	-0.004	-0.92	0.358
LKM	-0.097	31.23	0.000
LNH	0.018	2.84	0.005
LNM	-0.014	-9.98	0.000
LHM	-0.038	-8.36	0.000
LKK	0.055	45.71	0.000
LNN	0.024	7.43	0.000
LHH	0.054	1.10	0.271
LMM	0.084	37.18	0.000
RESET Test:	F[3, 1879]=18.74	p-value=0.000	
Breusch-Pagan LM:	X^2= 1475.04	p-value=0.000	

The restricted model is estimated with the same data set and the test statistics shown in Table 2-07 for heteroskedasticity and wrong functional form are obtained as well. Since the Reset test continues to identify a wrong functional form, neglected heterogeneity is indeed the likely reason for the equation's statistical problems.

Table 2-07 Regression with Unclassified Data Set After Specification

Variable	Coefficient	t-ratio	p-value
Constant	5.205	16.57	0.000
LK	0.332	51.67	0.000
LN	0.575	9.65	0.000
LH	0.392	8.73	0.000
LM	-0.114	-2.86	0.000
LKM	-0.104	-45.87	0.000
LNH	0.019	4.85	0.000
LNM	-0.044	-12.09	0.000
LHM	-0.035	-9.90	0.000
LKK	0.054	49.17	0.000
LNN	0.021	7.22	0.000
LMM	0.087	44.80	0.000
RESET Test:	F[3, 1882]=16.35	p-value=0.000	
Breusch-Pagan LM:	X^2= 1024.93	p-value=0.000	

4. Identifying Subgroups for the Data

Many different approaches exist to classify the observations of a data set into homogeneous subgroups in terms of meaningful characteristics. In this paper, the technique of cluster analysis is applied.

Applied to a data set, clustering is the grouping of observations into subsets on the basis of their similarity across a set of variables that describe the individual characteristics of the observations. The observations here are the objects one may want to classify; in this case, they are 1,897 U.S counties. The group of variables is developed according to meaningful characteristics that describe in what respect the observations in each subset are alike. The general objective of cluster analysis is to partition or subdivide a set of observations into homogeneous subgroups.

Generally, the steps for processing cluster analysis include the following:

1. Obtain the data matrix. This includes selecting a representative and adequately large sample of observations for study and selecting a representative set of variables that capture meaningful characteristics and that can also serve as domains on which the observations tend to show their similarity. For example, when one studies a group of people as observations for medical treatment research, common sense tells us that age and health condition of these people play important roles in classifying them into homogeneous groups, but hair color undoubtedly has nothing to do with it, even though it indeed is one significant characteristic of people.
2. Standardize the data matrix and/or conduct a dimensional analysis of the variables if that is necessary. When there are several variables that are used to

classify the observations, the cluster analysis is being done in a multiple dimensional space. Sometimes when there are too many variables that are involved in the analysis, it is important to reduce the dimensions by using statistical techniques in order to examine the results more clearly. Factor analysis is a well-known technique (Heikkila 1996) for this purpose and, therefore, it is employed in this study.

3. After the data are prepared and the dimensional space for the analysis is constructed, an appropriate clustering algorithm is selected. During the cluster analysis process, a similarity matrix is generated first. This matrix is the general layout of the observations in the given dimensional space. Similar observations tend to gather together to form homogenous groups in the matrix (Hartigan, 1985). The chosen clustering algorithm is applied to this matrix in order to determine the final solution of the clusters. There are two types of algorithms, hierarchical and non-hierarchical (Jain and Dubes, 1988, p.55-142). As discussed in later chapters, this study employs a combination of the algorithms.
4. Execute the clustering method. This step involves a fair amount of calculations and generates a number of statistics that will help to analyze the clustering process and to determine the final solution. Details on in-depth implementation of the procedure and examination of the output, plots and statistics will be presented in a later chapter.

The decision of what classification variables to include in the cluster analysis is very important. One needs to be very careful to define the dimensional space, because inclusion of a non-relevant variable provides no help but only noise to the final solution of the clusters. On the other hand, when relevant variables are left out of an analysis, some groups will remain undifferentiated from groups that are adequately defined. Cluster analysis is based upon the identification of the similarity of the observations in the data set. Unfortunately, similarity is not a general quality. When observations are compared in a given dimensional space of different variables, observations similar in one dimension need not be similar in another dimension, like the example of hair color in the above discussion. That is, there is no single way or given rule to categorize entities. Construction of the dimensional space must be based on the underlying theories for the study and the relevance of the data.

Once one has decided on the attribute space, one can usually find that it has more than three dimensions, which makes it almost impossible to locate clusters using two or three-dimensional plots. In many cases, researchers need other multivariate methods to make the clusters of observations apparent.

This brings up an important question: how should one choose the attributes to carry out the cluster analysis? Researchers usually find an answer in the application of a principle called the *factor asymptote*, which in practice is called factor analysis. The procedure can be described as follows.

First, one draws up a list of theoretically meaningful variables, that is, variables that enter into or reflect scientific theory. Second, one uses as many of the variables from this list as possible to describe the observations that are to be clustered. Third, a factor analysis is performed and used to find the factors that represent essential variables, that is, the ones that help to discriminate among clusters.

In the present application of cluster analysis, the observations of interest are 1,897 U. S. counties. The variable dimensional space to be constructed is the one that describes the behavior of those counties as it is captured by a production function.

In this study, eight initial variables are chosen to be the attributes that will be used in the cluster analysis. In addition, the residual that is obtained from the estimation with the unclassified data set is also included in order to cover any remaining information from the variables included in the production function.

- *ZABR*: county size and location. This is a nominal scale variable, which can assume integer value 0 through 8.

  0 - counties with population over one million;

  1 - fringe MSA counties with population over one million;

  2 - MSA counties with population of 250 thousand or more;

  3 - MSA counties with population less than 250 thousand;

  4 - for rural counties adjacent to MSA with population of 20 thousand or more;

  5 - for rural counties not adjacent to MSA with population of 20 thousand or more;

  6 - for rural counties adjacent to MSA with population less than 20 thousand;

  7 - for rural counties not adjacent to MSA with population of 7.5 thousand to 19,999;

  8 - for rural counties not adjacent to MSA with population less than 7.5 thousand.

County size and location are significant characteristics regarding the county's production behavior. The dense built environment of a large city is an excellent location for small-scale craft activities, but firms engaged in large-scale production find it easier to obtain economies of scale through the construction of a greenfield plant. (Eff, 1995)

Other than *ZABR*, the following variables are also relevant to production behavior of the US counties.

- *EDA*: Educational attainment in the county. It approximates the educational level of the local labor force.
- *MFSIZ*: Mean firm size, equal to total employment divided by the number of all establishments in the manufacture sector. Firm size of manufacturers is a meaningful measure for production efficiency and capability to reach economies of scale.
- *HWAGE*: Hourly wages, total wages divided by total hours of production employment in the manufacturing sector. It measures the average income status of production workers and the relationship of demand and supply in the local production labor market.
- *LSHARE*: Share of labor, defined as payment to total employment divided by value added in the manufacturing sector. It measures the total contribution from labor to the production process.
- *COMPLEXI*: Complexity of production procedures, defined as value added divided by the value of shipments of manufacturers.
- *PRDTY*: Labor productivity, defined as value added divided by production employment of manufacturers.

- *KEMP*: Per capita capital investment, obtained by dividing new capital expenditure of manufactures by total employment of manufactures and used to capture the relationship between labor input and capital input.

CHAPTER 3

METHODOLOGY, STRATEGY AND CRITICAL ISSUES IN THE USE OF CLUSTER ANALYSIS

1. Clustering Variables and Data Preparation

1.1. Standardization of Variables

Cluster Analysis calculates distances among elements and groups of elements. More specifically, it maximizes the distances among groups and minimizes the distances of elements within groups. In the multi-dimensional space that is defined by the selected variables, variables with large ranges are likely to separate the elements by larger distances than the ones with small ranges. Therefore, these variables will be given more weight in defining a cluster solution than those with smaller ranges during the process. For example, if one uses *height* and *weight* to construct a two dimensional space to apply cluster analysis on a group of people, the variable *weight* with range in hundreds has much more impact on the calculations than the variable *height* whose range is in ones. If a few variables in the clustering space have an unusually large range, they can dominate the clustering process and skew the final solution of clusters.

In this study, the values of variables *MFSIZE*, *PRDTY* and *LSHARE* have significantly larger ranges than those of the others. So they would have more impact on the selection of clusters. Since all the variables should be equally important in the study, one needs to find a way to equalize the statistical weight assigned to each variable. The solution to this problem is variable standardization. Standardized variables have a mean of zero and a standard deviation of one. The advantage of this transformation is that it allows variables to contribute equally to the definition of clusters.

However, standardization has also a potential downside: it may eliminate meaningful differences among elements. One must carefully study the variables in the clustering process and relationship among them to decide if a standardization procedure is necessary. For example, if the larger range and variation that a variable possesses contain meaningful information that needs to be included in the clustering process, standardization may not be appropriate.

1.2. Principal Component Analysis[3]

Principal Component Analysis is one type of *Common Factor Analysis*. Principal Component Analysis is a multivariate technique for examining relationships among several quantitative variables. It is widely used by researchers as a valuable tool in exploratory data analysis to summarize data and detect linear relationships. It also is popularly used to reduce the number of variables in regression, clustering, and other multivariate techniques, especially when multicollinearity exists among variables. In this study, since certain classification variables are derived from the same original variables, some of them are highly correlated as shown in Table 3-01. Therefore, factor analysis becomes a necessary step before conducting the cluster analysis.

[3] The mathematical background of factor analysis is explained in detail in *Factor Analysis* (Kim and Mueller, 1978) and other related literatures.

Table 3-01 Classification Variable Correlation Matrix

	RESIDU	COMPLEXI	EDA	HWAGE	MFSIZ	LSHARE	PRDTY	KEMP	ZABR
RESIDU	1.00	0.07	0.16	0.46	0.06	0.24	0.08	0.05	-.03
COMPLEXI	0.07	1.00	0.03	-.10	0.09	-.21	0.09	-.16	-.10
EDA	0.16	0.03	1.00	0.47	-.11	-.03	0.39	0.15	-.38
HWAGE	0.46	-.10	0.47	1.00	0.20	0.04	0.57	0.35	-.39
MFSIZ	0.06	0.09	-.11	0.20	1.00	-.01	0.14	0.10	-.12
LSHARE	0.24	-.21	-.03	0.04	-.01	1.00	-.52	-.13	0.02
PRDTY	0.08	0.09	0.39	0.57	0.14	-.52	1.00	0.37	-.36
KEMP	0.05	-.16	0.15	0.35	0.10	-.13	0.37	1.00	-.18
ZABR	-.03	-.10	-.38	-.39	-.12	0.02	-.36	-.18	1.00

For a data set with *p* numeric variables, *p* principal components are computed. Each principal component consists of linear combinations of the original variables, with coefficients equal to the eigenvectors of the correlation or covariance matrix. Table 3-02 provides the resulting eigenvalues of the principal components. In the table, the principal components are sorted in descending order of the eigenvalues. The eigenvalue is a measure of variance accounted for by a given dimension in factor analysis (Kim and Mueller, 1978). In other words, an eigenvalue is interpreted as the amount of variance explained by the factor in question, where each of the original variables is normalized to have unit variance.

Table 3-02 Eigenvalues of the Correlation Matrix from Factor Analysis

Factors	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	2.672	1.144	0.296	0.296
PRIN2	1.527	0.353	0.169	0.466
PRIN3	1.174	0.064	0.130	0.597
PRIN4	1.110	0.228	0.123	0.720
PRIN5	0.881	0.249	0.097	0.818
PRIN6	0.631	0.130	0.070	0.888
PRIN7	0.501	0.184	0.055	0.944
PRIN8	0.316	0.131	0.035	0.979
PRIN9	0.184	.	0.020	1.000

1.3. Number of Factors

In this study, factor analysis serves the core purpose of preparing the data for cluster analysis. The factors that are generated from the factor analysis represent the original variables as linear combinations of these variables in a redefined and reduced dimensional vector space. Therefore, while the results of factor analysis are being evaluated, researchers have to determine the number of factors to be retained.

The choice of the number of factors is always more or less arbitrary. However, there are certain rules of thumb that are applicable to the analysis.

First, a popular objective rule adopted by many researchers is that the eigenvalues of retained factors should be greater than 1 (Kim and Mueller, 1978). This is usually a necessary condition, but not a sufficient condition for determining the number of factors to be retained. One must apply the underlying economic theory to the process of evaluating the resulting statistics in order to obtain meaningful factors.

Second, the cumulative explanatory power of all retained factors should be over 50% of the explanatory power of all the original variables. However, it is not necessarily true that more cumulative explanatory power raises the likelihood that one has identified the correct number of factors. This is because irrelevant variables that are loaded into a factor may lead to spurious associations.

Finally, a researcher's subjective judgment plays an important role in determining the number of retained factors.

Table 3-03 presents positive and negative correlations among factors and original variables. In factor analysis, these correlations explain the factor loading for each variable. By examining how much of each variable has been loaded to a factor we can find interpretations for the factor, which in turn can be thought of as a composite variable.

Table 3-03 Correlation Matrix of Factors and Original Variables

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5
RESIDU	0.212	0.492	0.358	0.040	0.539
COMPLEXI	0.033	-.318	0.734	-.163	0.071
EDA	0.394	0.099	-.037	-.513	-.101
HWAGE	0.513	0.276	0.010	0.077	0.100
MFSIZ	0.139	-.020	0.344	0.745	-.313
LSHARE	-.155	0.689	0.050	-.015	-.306
PRDTY	0.498	-.308	-.056	0.022	0.177
KEMP	0.320	-.031	-.444	0.340	0.104
ZABR	-.376	0.004	-.105	0.178	0.673

Note: Variables are defined in Chapter 2 section 2. PRINs are names for factors.

Table 3-04 summarizes and interprets the results presented in Tables 3-02 and 3-03. Specifically, it expresses the retained factors in an economically meaningful way. Based on the discussed criteria of determining the number of factors, after evaluating the eigenvalues of each factor, factors 1 through 4 meet the first criterion – the eigenvalue is greater than 1. By retaining these four factors, their cumulative explanatory power is 72%.

The fourth column of Table 3-04 reports all variables that have comparatively larger loadings (>0.3 in absolute value) for each factor. The third column provides a meaningful interpretation of the factor on the basis of the variables in the fourth column.

From Table 3-04 it is apparent that the first three factors cover all the original variables and have over 50% explanatory power. Considering the serious multicollinearity among the variables, the fourth factor is left out of the final solution.

Factors 1 through 3 will be used as new variables to construct a reduced dimensional space for further cluster analysis.

Table 3-04 Factor Interpretation

Factor	Eigenvalue	Factor Label	Associated Variable

One	2.67	Capital-
		Intensive
			EDA(0.39);
			HWAGE(0.51);
			PRDTY(0.50);
			KEMP(0.32)

Two	1.52	In City
		Labor-
		Intensive
		General
		Machinery
			ZABR(-0.38)
			LSHARE(0.69);
			RESIDUE(0.49)
			COMPLEXI(-0.32);
			PRDTY(-0.31)

Three	1.17	Labor-
		Intensive
		Craft Shop
			COMPLEXI(0.73);
			MFSIZ(0.34);
			RESIDUE(0.36)
			KEMP(-0.44);

Four	1.11	Physical
		Capital-
		Intensive
			MFSIZ(0.74);
			KEMP(0.33)
			EDA(-0.51)

The first factor listed in Table 3-04 points to a large group of highly intercorrelated variables. The variables that are positively correlated with factor one include *Education Level* (*EDA*), *Hourly Wages* (*HWAGE*), *Productivity* (*PRDTY*), and *Per Capita Investment* (*KEMP*). The factor loading is approximately equal on all variables. None of them dominates. The variables point to profitable enterprises in capital intensive industries with highly paid employees, who are most likely well educated white collar professionals managing and producing efficiently.

Factor one points also to variable *ZABR*, which is negatively correlated with it. In Chapter 2, variable *ZABR* was discussed in detail. It is a measure of location and population of US counties. A lower value of this variable indicates closer location to the MSA area and a larger population. Hence, this variable has a positive correlation with the other three variables that are highly related to first factor, because MSA areas are more likely to have population with higher educational level and hourly wages and to have more investment capital.

Factor two points to two variables with the variable *Labor Share* (*LSHARE*) dominating the loading. A large amount of positive loading for *Labor Share* (*LSHARE*) would point toward a labor-intensive industry. From Table 3-01, one can find that *RESIDUE* has a positive correlation with *Hourly Wages (HWAGE)* and *Labor Share* (*LSHARE*). One may interpret this correlation to mean that *RESIDUE* is a supplemental factor to labor resources.

The negative loading of *Productivity (PRDTY)* on factor two points toward lower productivity in labor intensive industries. Furthermore, the negative loading of *Complexity of Production (COMPLEXI)* gives a hint that large amounts of simple labor are used in the production process that lowers productivity even more.

The positive and high loading of *Complexity of Production (COMPLEXI)* on factor 3 along with a positive *RESIDUE* loading relates this factor to labor intensive and highly complex activities, which most likely are those of hand-making crafts. The negative loading of *Per Capita Investment(KEMP)* also supports the viewpoint that the high value added comes from the intensive use of labor resources. In addition, the minor negative loading of *ZABR* suggests that the location is near or in a city where craft shops usually operate their business. Factor 3 also shows a positive correlation with *Size of Firm (MFSIZ)*, which does not have a tight correlation with either of the variables above.

At the end of the discussion of factor loading, one should keep in mind that the loading of a combination of variables describes the characteristics of the factor that represents them in the new dimensional space, when these variables are correlated with each other mathematically and economically. However, this does not prevent non-correlated variables from loading on the same factors. It is the researcher's task to interpret the representation of a factor in accordance with the underlying theory and previous empirical evidence for any specific study.

From the above results on factor loadings and variable correlations, it is apparent that the three retained factors cover all the variables in the original data set. Each variable has been loaded to different factors with different weight. Factor analysis does not try to retain 100 percent of he explanatory power of the original data set. Instead, it decomposes the variables and concentrates on the meaningful factors. It generates a reduced dimensional space. Given this purpose of factor loading, it is reasonable to conclude that the three factors that are retained in this study do represent the variables of the original data set.

2. Cluster Analysis

2.1. Choice of Clustering Algorithms

The common basis for all clustering algorithms is to cluster together similar or neighboring points in a multidimensional space. The major differences among the algorithms are the methods and criteria used for establishing similarity and the rationale according to which clusters are joined together.

In general, there are two prominent types of algorithms: hierarchical and nonhierarchical clustering methods. The mathematical background for these two types is the same. They both start the process by calculating a similarity matrix. The similarity, which is a measure of the proximity of the pairs of observations in multidimensional space, is the Euclidean distance between any given two points. During the clustering process, this matrix is being updated whenever two clusters are joined. Once the similarity matrix has been established, the various clustering methods can be applied to it.[4]

Hierarchical Algorithms

Hierarchical Clustering has basically two different approaches: agglomerative and divisive procedures. In agglomerative clustering, each observation in a data set is initially considered an individual cluster of its own. The hierarchical classification is built up by a series of linkages in which the most similar pairs of clusters are joined until all of the compounds[5] are in a single cluster. The agglomerative clustering process is a process of maximizing the similarity among the observations. Conversely, the divisive algorithm begins by placing all the observations into one cluster, which is then progressively subdivided into smaller ones until each observation is again in a cluster of its own. The divisive clustering process this is a process of minimizing the dissimilarity among the observations. The most popular agglomerative algorithms are single linkage, complete linkage, average linkage, the centroid method and Ward's method (Jain and Dubes, 1988,

[4] The mathematical background of cluster analysis is explained in detail in *Cluster Analysis* (Everitt, 1974) and other related literatures.

p.55-142). Once an algorithm is selected, the researcher must specify the number of groups required. The use of divisive methods in the social sciences has been limited to very few research fields.[6] None of these methods are applied to this study either.

Hierarchical cluster analysis has certain advantages over non-hierarchical method. First, it provides a history of the clustering process. Researchers can examine the history of the clustering process to gain some insight into the underlying structure and characteristics of clusters. Second, many of the statistics that are produced during the clustering process are very useful for determining the number of clusters in the population. Researchers usually do not know how many subgroups are hidden in the population that is being studied. Most of the time, identifying the number of clusters is one of the intermediate tasks for researchers. Finally, by the nature of hierarchical cluster analysis (agglomerative), it reaches the final cluster from "bottom up" and produces the trace of merging the clusters along the way. The output from this has usually been used by researchers to draw a tree diagram of the cluster hierarchy, which provides researchers a visual insight of the clustering process. Researchers often apply their field-specific knowledge to make the decision to "cut the tree" at the desired level.

A number of disadvantages of hierarchical algorithm have been discovered by researchers in a variety of studies. First, "it is essential to realize that most methods are biased toward finding clusters possessing certain characteristics related to size (number of members), shape, or dispersion." (SAS Institute, 1990, p. 56). None of the methods can provide researchers with any guarantee that the results of the clustering process are

[5] A cluster Compound represents a joining unit in a dynamic clustering process. At any given time or stage, it is a cluster. But when the clustering process moves forward in a sequential hierarchy, it will be a component of bigger clusters.

'correct'. Second, because there are so many clustering methods that have been developed in several different fields, with different definitions of clusters and similarity among observations, and because researchers often do not know the underlying structure of the data in advance, selection of a correct clustering method can be a very lengthy process. Third, hierarchical clustering is a one step "non-stop" process. Once it is carried out, there is no intermediate stage where researchers can modify the joining of the clusters. Finally, it is not practical for very large data sets because of the very heavy computational burden at each stage of cluster joining. Many researchers also find that, when the data set is very small, solutions often become unstable when observations are dropped.

Nonhierarchical Algorithms

The nonhierarchical methods are very useful tools for disjoint clustering of large data sets and one can find good clusters with multiple passes over the data. Different nonhierarchical methods function in essentially the same manner. The most popular algorithm is the k-means algorithm (MacQueen, 1967). The *FASTCLUS* computing procedure provided by the SAS package (SAS/STAT, 1990, p. 823-850), which uses the nearest centroid sorting methodology (Anderberg, 1973, p.160-173), gives an insight into the standard clustering process of nonhierarchical algorithm.

The *FASTCLUS* procedure operates in four steps:

[6] The mathematical background and empirical evidence for not using the divisive method are well explained by Ketchen and Shook (1996)

1). It starts by selecting observations as cluster seeds (centroid), either randomly or based on a given data set that contains pre-determined cluster seeds.[7]

2). Observations are being assigned to the cluster with the nearest seed. These clusters are temporary. Each time a new assignment is passed through all of the observations, the cluster seeds will be recomputed and updated as the current mean of the cluster.

3). Step 2 is repeated until the changes in the cluster seeds become small or zero based on the number of iterations that the researcher chooses.

4). When no more observations need to be assigned, the final clusters are formed.

Compared with hierarchical algorithms, nonhierarchical methods have several advantages. First, nonhierarchical methods allow observations to switch to different clusters multiple times and cluster seeds to be updated. This process, therefore, optimizes the observation assignment and minimizes the disturbances from possible outliers in the data set. Second, since its final clusters are formed by assigning each observation to the nearest seed, it optimizes within-cluster homogeneity and between-cluster heterogeneity. Finally, the nonhierarchical clustering process can be easily applied to large data sets. Most computer software package can generate, from a nonhierarchical process, an output file that can be used as an input data set for hierarchical clustering.

The biggest disadvantage of a nonhierarchical clustering process is that it requires the number of clusters to be specified *a priori*. In many research fields, especially the social sciences, cluster analyses are often exploratory. Researchers do not necessarily have good priors on what kind of clusters to expect and how many of them there may be.

[7] Cluster seeds are those observations with which the mathematical calculation of the cluster analysis starts. These seeds serve as attraction centers that other observations can join based on their distance from the

Another restriction on the usefulness of nonhierarchical clustering methods relates to small data sets: the clustering results may be highly sensitive to the order of the observations in the data set.

This brief discussion summarizes the dilemma faced by researchers using cluster analysis. Based on their reported experience and many simulation studies, many researchers (e.g., Milligan, 1980; Punj and Stewart, 1983) suggest an approach of a two-stage procedure of clustering. The two-stage procedure combines the two types of algorithms of hierarchical clustering and nonhierarchical clustering and is also used in this study.

There are two different approaches of two-stage procedures to select from.

First, one starts with hierarchical clustering and then follows up with nonhierarchical clustering. The hierarchical algorithm is used to define the number of clusters and cluster seeds. These, in turn, serve as starting points for the subsequent nonhierarchical clustering.

Second, one starts with nonhierarchical clustering and continues with hierarchical clustering. The idea of this approach is to apply nonhierarchical clustering to a large input data set to produce a smaller summary file, which can be used by hierarchical analysis algorithms.

The first approach is applicable to smaller data sets. Researchers often get to know their data well because of the smaller size and know that the existing clusters are well separated. The only thing not known to the researcher is how many clusters reside in the data set.

seeds in a multidimensional space.

The second approach is very useful for large input data sets. This approach can benefit researchers by eliminating possible outliers in the data set, for most hierarchical algorithms are sensitive to disturbances from outliers.

The second approach, with Ward's minimum variance method being the hierarchical algorithm, is employed by this study because of following conditions.

1). Ward's minimum variance method is chosen because it has been shown to be best among the hierarchical algorithms by a variety of simulation studies and it is often used in social science fields.

2). The large size of the data set that is being used in this study requires the nonhierarchical procedure to predefine subgroups for further analysis.

3). As has been discussed before, all hierarchical methods are more or less biased on finding clusters. Ward's minimum variance method tends to find clusters with roughly the same number of observations in each cluster (SAS/STAT, 1990, p. 56). Empirical studies (e.g., Milligan, 1980) have also found that the solutions that this method provides tend to be heavily distorted by outliers. Both of the problems can be solved by using a nonhierarchical method first to process the data. In this way, the outliers can be identified and removed from the data set. Because the nonhierarchical method will generate the 'preliminary clustering' output data set, which will be used as the input data set by the subsequent hierarchical method, Ward's method will start with pre-seeded clusters that will reduce the chance that the method averages the final clusters.

2.2. Determining the Number of Clusters

In most real world applications, researchers are often faced with the dilemma of selecting the number of clusters for the final solution. Along with the development of a variety of clustering methods, numerous (over 30) procedures and criteria for determining the number of clusters have been proposed (e.g. Dubes and Jain, 1979; Milligan and Cooper, 1985). These techniques, which usually are applied to hierarchical algorithms, are called stopping rules.

There are two types of decision errors that can happen when researchers try to apply these rules in contrast to a "correct" solution of the true clusters. In their simulation study on procedures for determining the number of clusters in a data set, Milligan and Cooper (1985) introduce two errors. The first type of error arises when the stopping rule indicates a final solution point when the number of clusters is larger than the number of true clusters in the data. By contrast, the second type of error happens when the final solution contains fewer clusters than the number of true clusters in the data. The second type of error might be considered more serious in most applied analyses because information is lost by merging distinct clusters. (Milligan and Cooper, 1985)

In their simulation study, which focused on estimating the capability of finding the true underlying cluster structure and performing consistently across different samples of data for all stopping rules, Milligan and Cooper (1985) found that three criteria – *pseudo F* statistic developed by Calinski and Harabasz (1974), a statistic *Je(2)/Je(1) (pseudo T Square)* by Duda and Hart (1973, p.217-224) and *Cubic Clustering Criterion*

(CCC) by Sarle (1983) – performed best among 30 criteria in the study. This study employs these three statistics and looks for consensus among them as a criterion to determine the number of clusters in the original unclassified data set.[8]

More and more multivariate techniques, sophisticated computer software and various statistical procedures developed by researchers make the objective judgment on determining the number of clusters more realistic than ever before. However, the researcher's knowledge of the subject matter, admittedly a subjective factor, still plays a key role in making the 'correct' decision. *A priori* theory continues to serve as a key non-statistical tool to provide a benchmark for assessing the results.

[8] Determining the number of clusters is a very complex process in cluster analysis. One should develop the method of using different criteria based on the ongoing research, chosen cluster algorithm and references of similar previous studies.

CHAPTER 4

THE CLUSTER ANALYSIS PROCESS AND ITS RESULTS

The clustering process of this study can be divided into several steps. These steps will be discussed in turn.

1. Pre-Process with Original Data Set for Pre-Seeded Clusters

As has been discussed in the previous section, three factors are retained in the clustering data set, along with the observation identifier, *STCO* the state-county code. These three factors are the clustering variables. A *FASTCLUS* procedure, which is a nonhierarchical method in SAS, is carried out first. This step produces 100 preliminary clusters, which one can summarize and visually examine to find clues for the formation of final clusters. This procedure also creates an output data set that contains the preliminary cluster means and other statistics for each cluster. This preliminary cluster data set is used for the hierarchical cluster analysis that follows.

The following chart gives one a good idea how this process can be done.

Figure 4-01 Flowchart of Pre-Process of Cluster Analysis

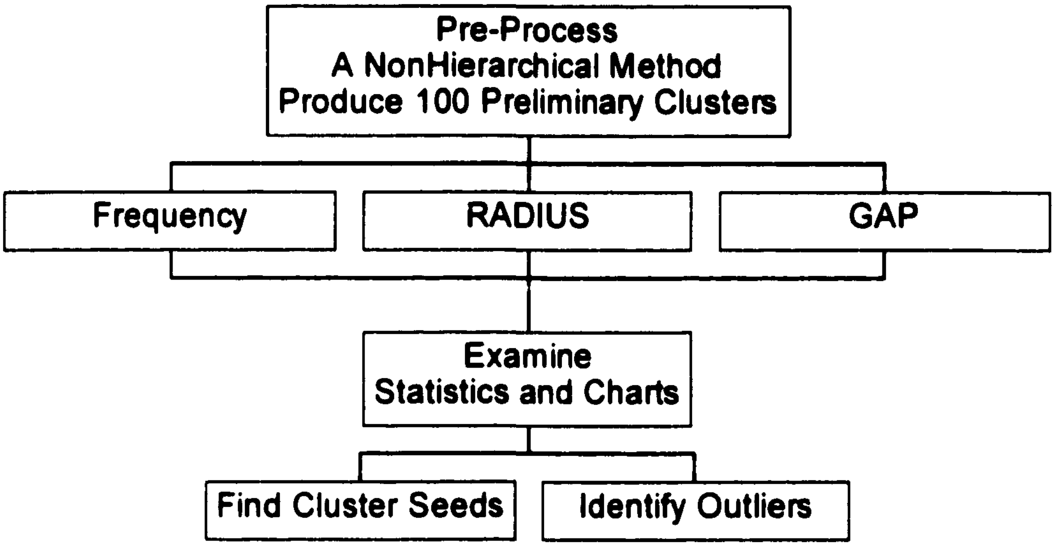


Table 4-01 lists all clusters with more than 50 observations out of 100 preliminary clusters. There are 10 of them. Among these, six clusters have over 90 observations.

Table 4-01 Clusters with More than 50 Observations after 1st Round Selection by *FASTCLUS* - a Nonhierarchical Clustering Procedure

Cluster Position	1	3	28	40	42	48	50	75	92	97
Frequency	165	95	183	75	59	158	114	72	56	91

Figure 4-02 is a scatter plot of the *Distance Between Cluster Centroids (GAP)* by *Frequency (FREQ)* and an overlay of *Maximum Distance From Seed To Observation Within A Cluster (RADIUS)* by *Frequency*, which researchers can use as a baseline to compare against the values of *GAP*.

Visual examination of such graphical output provides insightful information on the distribution of observations, any tendency for cluster formation and possible outliers.

In Figure 4-02, one can split the plot area into two, left and right areas. In this figure, one should look for those clusters with high *Frequency*, large *GAP* and relatively small *RADIUS*. High *Frequency* means that this cluster attracts more similar observations, large *GAP* and small *RADIUS* means that observations in this cluster are closer to each other and farther from any other observations outside of the cluster. One can see in the right side area of the plot that good clusters appear with large values of both *GAP* and *FREQ*. One can also notice that the 'G's are always above the 'R's for these clusters. That means that they are well-separated clusters with gaps between them greater than the radius within each of them. These are good potential cluster seeds for further analysis.

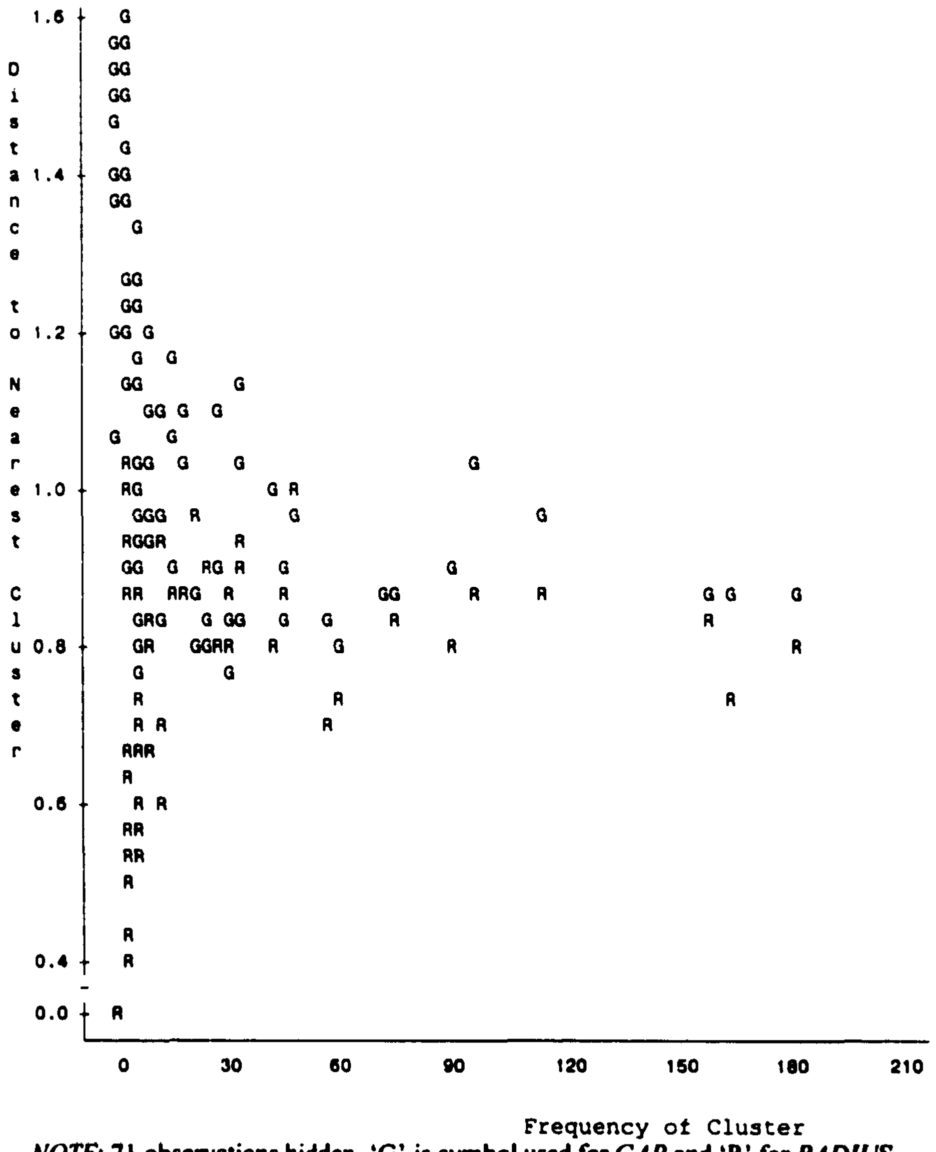
Outliers often appear as clusters with only one member. They are located in the left side area in the plot where the clusters are low in *Frequency*. Some of them with extremely high *GAP* and low *RADIUS* in the far-left side area are most likely those clusters with single or few observations that are far away from others.

In a typical classification study for which the final goal is classification itself, outliers and sometimes even clusters with small frequencies are removed from the original data set to improve cluster separation. This strategy is not applied in this study for two reasons.

First, the purpose of this research differs from that of the typical classification study. Classification analysis is used not as an end in itself but as a tool to explore the underlying structure and characteristics of observations in order to explain their economic behavior. Applying restrictions at an early stage may cause too much loss of information.

Second, this preliminary cluster analysis provides an output data set that contains all the good potential clusters as cluster seeds. These are the starting points for the consequent hierarchical clustering process. Since these good potential clusters, which have more observations such as those over 50 observations, tend to attract other clusters or observations to form new clusters in the next clustering process, disturbance of the outliers that are far from these clusters will be limited.

Figure 4-02 Plot of *Distance Between Cluster Centroids (GAP)* and *Maximum Distance From Seed To Observation Within A Cluster (RADIUS)* against *Frequency* of clusters



NOTE: 71 observations hidden. 'G' is symbol used for *GAP* and 'R' for *RADIUS*

2. Determination of Cluster Solutions

Ward's method of hierarchical clustering is applied to the data set created in the preliminary process.[9]

Before one can make any decision on a final solution of what and how many of the clusters to retain, a series of statistics has to be studied and explained carefully.

In Table 4-02, statistics for the first 15 cluster solutions are summarized. To determine the number of clusters, a practical way to proceed is to look for agreement among the three statistics (*pseudo F, pseudo T Square,* and *CCC*) that tend to be most useful in determining the number of clusters. It is the combination of local peaks of the *CCC* and *pseudo F* along with a small value of the *pseudo T Square* and a larger *pseudo T Square* for the next cluster fusion that suggests that an optimal number of clusters may have been found.[10]

Figures 4-03 and 4-04 provide plots of these statistics against different final cluster solutions. From Figure 4-03, one can see that the local peaks of the values of *CCC* are not significant, but that there are turning points at 5, 7, 11 and 16 clusters. The only peak for the values of *Pseudo F* in Figure 4-04 is at 7 clusters. The values of *pseudo T Square* have several clear indications of a drop in value followed by an upward jump. These points occur at 5, 7, 10, 14 and 18 clusters.

[9] Ward's method of hierarchical clustering algorithm discussed in this chapter is implemented with the *CLUSTER* procedure provided by SAS (SAS/STAT, 1990 p. 519-614)

[10] Mathematical backgrounds and applications of these three statistics are discussed in detail by Calinski and Harabasz (1974), Duda and Hart (1973), Sarle (1983) and Milligan and Cooper (1985). Also, SAS (1990) provides much guidance and advice on their practical usage.

Table 4-02 Statistics for First 15 Solutions of Clusters

Number of Clusters	CCC	Pseudo F	Pseudo T^2
15	-35.1	352.4	194.7
14	-35.7	354.2	70.5
13	-36.1	357.9	68.7
12	-36.1	364.7	127.5
11	-36.8	367.1	145.0
10	-37.2	372.3	114.5
9	-37.7	377.2	233.9
8	-37.8	386.8	201.4
7	-37.8	398.6	159.7
6	-39.6	395.7	268.2
5	-40.3	403.6	179.3
4	-41.2	413.4	213.3
3	-26.1	413.6	288.6
2	-13.8	429.0	373.2
1	0.0	.	429.0

As described above, too few clusters in the final solution will force unrelated clusters to merge together, which leads to a loss of information. However, too many subgroups are not useful either because it makes economic policy applications rather difficult. Yet providing useful information for policy applications is a key intent of this study. Therefore, those solutions with clusters in excess of 10 are not under consideration. One is left with two solutions, one having 7 clusters and one with 10 clusters. These are chosen for an extended comparison study. The details of the comparison study are discussed later. At this point, it will only be noted that the solution with 10 clusters is preferred on the basis of the comparison study and that this solution is used for the remainder of the study.

Figure 4-03 Plot of *Cubic Clustering Criterion (CCC)* against *Number of Clusters*

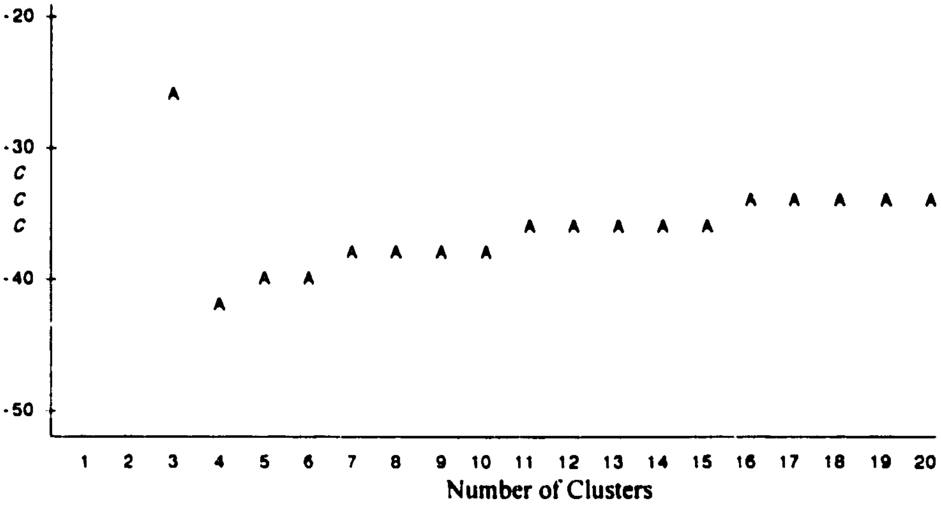
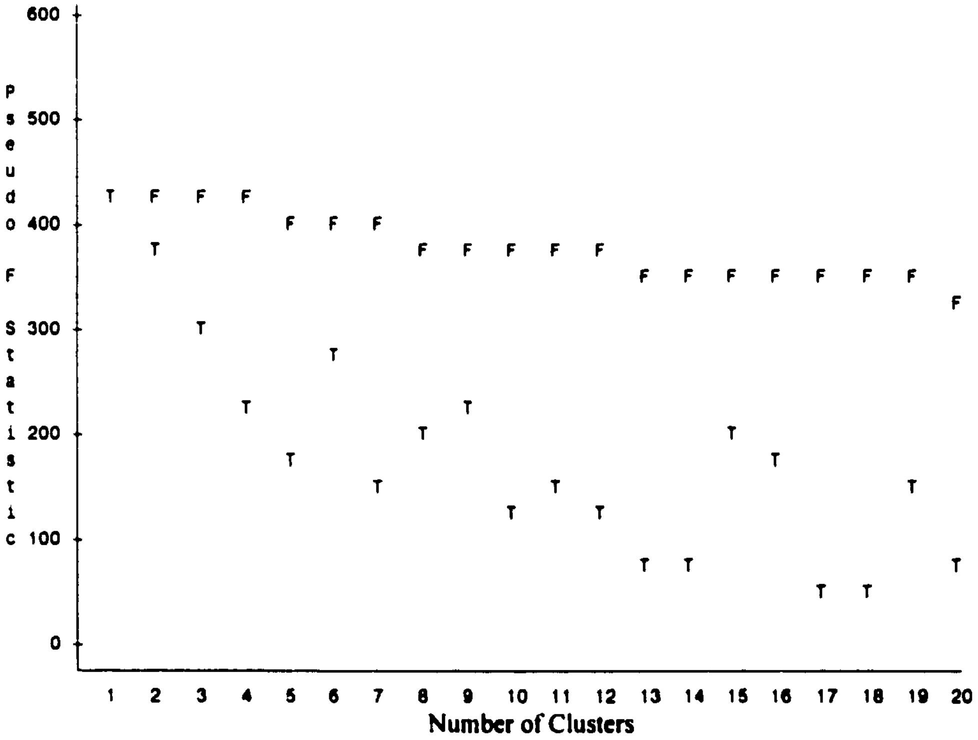


Figure 4-04 Plot of *pseudo F (PSF)* and *pseudo T Square (PST2)* against *Number of Cluster*



Note: F is the symbol for *pseudo F (PSF)* and T for *pseudo T Square (PST2)*.

3. Validation of the Final Solutions of Clusters

After one has narrowed down the number of solutions of the cluster analysis, in the above case to two, the issue arises how to assess the validity and reliability of the final choice. Without answering this question, one cannot be assured of having arrived at a meaningful and useful set of clusters (Punj and Stewart 1983).

From a scientific point of view, the final choice of clusters should be repeatable and valid based on some external criteria. Unfortunately, in most applied research projects, there are certain difficulties in strictly abiding by the scientific principle.

First, repeatability requires researchers to reach the same or similar solutions by using different methods on the same data or by analyzing different samples of the population in question with the same method. Both are hard to do in most real life clustering situations, because most studies, especially those in the social sciences, are exploratory in nature.

Second, one way to externally examine the validity of a cluster solution is to obtain some pre-knowledge of the number of clusters. This situation only exists when researchers are conducting a simulation study with an artificial data set. It is effectively impossible in real life applications, especially for exploratory research like this study.

Finally, the other way to validate a cluster solution with external criteria is to conduct statistical tests with external variables that are theoretically related to the clusters but that are not used in defining the clusters (Ketchen and Shook, 1996, p. 441-458). This is a very useful technique in many research fields, but it is hard to carry out in those fields where cluster analysis is rarely practiced, because researchers will have few case studies that they can rely on to establish a pool of meaningful variables. In this study, a strategy similar to this technique is explored and is discussed in some detail in the following sections. In particular, a combination of strategies and techniques is used to validate the cluster solutions. Among these are canonical R^2, variable loading power, and tests for neglected heterogeneity.

3.1. Canonical Discriminant Analysis

Canonical discriminant analysis is a dimension-reduction technique. It is similar to and related to principal component analysis. Given a data set with known subgroups of observations that are measured on several quantitative variables, canonical discriminant analysis can derive canonical variables (linear combinations of the quantitative variables) that summarize between-subgroup variation. One fact that needs to be understood is that discriminant analysis is different from cluster analysis. All varieties of discriminant analysis require prior knowledge of the clusters. In cluster analysis, constructing a classification is the final goal. There is no information available to researchers before the cluster analysis is performed.

The idea of discriminant analysis is to use a training data set, that is, a data set with known groups and associated quantitative variables, to derive so-called classification criteria that can be used to either study the groups in the data set or apply them to other samples of data from the same population. With the development of computing technology and applied statistical analysis, this technique is now extensively used in many fields (Klecka, 1980, p.12), often in conjunction with other statistical procedures. As is also the case in this study, both factor analysis and cluster analysis typically provide rather limited information about the final solution after the classification is done. Once the clusters are formed, canonical discriminant analysis, however, becomes a very powerful tool in revealing in-depth information about each cluster and the differences among the clusters.

The mathematical background of discriminant analysis is much like the one of principal component analysis. The procedure derives a linear combination of the quantitative variables that have the highest possible multiple correlation with the clusters. This highest multiple correlation is called the first canonical correlation. The second canonical correlation is defined in the same way as the first one, but is uncorrelated with the first one. As many canonical correlations as there are original variables can be obtained in this way.

Canonical discriminant analysis can be helpful for classification studies in two ways. First, in studying the ways in which groups differ, the researcher can draw on it to identify (i) how well the groups discriminate based on a certain set of characteristics and (ii) which characteristics are the most powerful discriminators. Second, canonical discriminant analysis can be helpful in deriving one, or more, mathematical equation from the known groups to set the criteria to accommodate new observations for future classification.

Before the decision on a final set of clusters is made for this particular study, a comparative study on the two most likely solution sets is performed. The comparison includes the following steps.

- Canonical discriminant analysis is performed on the two sets of clusters.[11] The particular method used performs univariate and multivariate one-way analyses of variance and computes squared distances between cluster means.

- Tests for heteroskedasticity and functional form on the log-linear Cobb-Douglas and on the more flexible functional form of the translog are performed in the regression analyses on all clusters of each final solution.

Following tables in this section provide statistics from the canonical discriminant analysis for the classification solutions containing seven clusters and ten clusters. In practice, there are only five clusters instead of seven retained from the solution of 7 cluster because two of them have observations less than 100, which is the minimum reasonable size for the application of canonical discriminant analysis. For the same reason, there are six clusters retained for the second solution set containing initially 10 clusters.

The above statistics for both solutions are compared with each other as follow. In general, the canonical disciminant analysis generates better values for the solution of 10 clusters.

Tables 4-03 and 4-04 show the *Pairwise Squared Distances* between clusters for both solutions of 7 clusters and 10 clusters, respectively. The statistics show that the distance between cluster means is smaller than Mahalanobis distance (SAS/STAT, 1990, p.388), where Mahalanobis distance is the distance between an observation and the centroid for each cluster in the multidimensional space defined by the variables (*PRIN1-PRIN3*).

[11] The *CANDISC* procedure provided by SAS (SAS/STAT, 1990, p.387-404) is employed for this analysis.

Table 4-03 Canonical Discriminant Analysis for Solution of 7 Clusters
Pairwise Squared Distances Between Groups

	Squared Distance to CLUSTER				
From CLUSTER	2	3	4	6	7
2	0 (0)				
3	5.26 (336.17)	0 (0)			
4	14.99 (583.54)	7.49 (836.70)	0 (0)		
6	14.99 (706.95)	7.72 (470.21)	8.98 (339.20)	0 (0)	
7	5.94 (357.35)	7.75 (654.34)	4.91 (224.59)	4.10 (235.53)	0 (0)

Notes: F statistics are in parentheses, they are all significant at the one percent level for the hypothesis that the distance between cluster means is smaller than Mahalanobis distance.

Table 4-04 Canonical Discriminant Analysis for Solution of 10 Clusters
Pairwise Squared Distances Between Groups

	Squared Distance to CLUSTER					
From CLUSTER	2	3	4	6	9	10
2	0 (0)					
3	17.71 (505.63)	0 (0)				
4	20.73 (806.63)	30.26 (750.94)	0			
6	18.99 (895.35)	15.02 (419.39)	9.88 (373.05)	0		
9	4.85 (265.27)	5.14 (156.29)	17.81 (755.36)	8.15 (426.85)	0	
10	6.41 (349.48)	18.06 (548.10)	6.03 (255.29)	5.08 (265.61)	5.17 (318.74)	0 (0)

Note: F statistics are in parentheses, they are all significant at the one percent level for the hypothesis that the distance between cluster means is smaller than Mahalanobis distance.

From Table 4-05 one can see that for the first solution (7 clusters), the correlation coefficients between the clusters and the individual variables that are retained from the previous factor analysis *(PRIN1-PRIN3)* are 0.61, 0.27 and 0.54. The average correlation coefficient is less than 0.5 for the set of quantitative variables.

Table 4-05 Canonical Discriminant Analysis for Solution of 7 Clusters
Univariate Test Statistics

Variable	Total STD	Pooled STD	Between STD	R^2	R^2/ (1- R^2)	F	Pr > F
PRIN1	0.94	0.59	0.82	0.61	1.55	687.69	0.0001
PRIN2	0.84	0.72	0.49	0.27	0.37	165.61	0.0001
PRIN3	0.93	0.63	0.76	0.54	1.18	523.07	0.0001

Average R-Squared: Unweighted = 0.4732759 Weighted by Variance = 0.4877234

Notes: Total STD stands for standard deviations for total sample; *Pooled STD* for pooled standard deviations of within-cluster; *Between STD* for standard deviations of between-cluster; F test for the hypothesis that the distance between cluster means equals the distance between observations to their cluster means within clusters.

In Table 4-06, for the clustering solution of 10 clusters, the R^2 values between the clusters and the individual variables are 0.65, 0.38 and 0.59. The average R^2 is more than 0.5 for the same set of variables.

Table 4-06 Canonical Discriminant Analysis for Solution of 10 Clusters
Univariate Test Statistics

Variable	Total STD	Pooled STD	Between STD	R- Squared	RSQ/ (1-RSQ)	F	Pr > F
PRIN1	0.87	0.52	0.77	0.65	1.85	596.55	0.00
PRIN2	0.78	0.61	0.52	0.38	0.61	196.54	0.00
PRIN3	0.92	0.59	0.78	0.59	1.44	464.59	0.00

Average R-Squared: Unweighted = 0.5388271 Weighted by Variance = 0.5525458

Notes: Total STD stands for standard deviations for total sample; Pooled STD for pooled standard deviations of within-cluster; Between STD for standard deviations of between-cluster; F test for the hypothesis that the distance between cluster means equals the distance between observations to their cluster means within clusters.

In Table 4-07, for the solution of 7 clusters, the canonical R^2 $CanR^2$ values for the association between the clusters and the canonical variables are 0.66, 0.51 and 0.21, and the eigenvalues are 1.94, 1.03 and 0.26, respectively.

The eigenvalue is equal to $CanR^2/(1\text{-}CanR^2)$, where $CanR^2$ is the corresponding squared canonical correlation, and can be interpreted as the ratio of between-cluster variation to pooled within-cluster variation for the corresponding canonical variable. (SAS, 1990, p. 399) The larger the eigenvalue is, the larger is also the between-cluster variation and the smaller is the within-cluster variation. Therefore, a larger eigenvalue is associated with greater discrimination.

Table 4-07 Canonical Discriminant Analysis for Solution of 7 Clusters
Multivariate Test Statistics Eigenvalues = $\text{Can}R^2/(1-\text{Can}R^2)$

	Canonical Correlation	$\text{Can}R^2$	Eigenvalue
1st Canonical Correlation	0.810	0.660	1.943
2nd Canonical Correlation	0.712	0.507	1.029
3rd Canonical Correlation	0.453	0.205	0.259

Notes: $\text{Can}R^2$ is Squared Canonical Correlation.

In Table 4-08, for the solution of 10 clusters, the squared canonical correlations are 0.70, 0.60 and 0.27, the corresponding eigenvalues are 2.37, 1.49 and 0.37, respectively. All these statistics have greater values than those for the solution of 7 clusters.

Table 4-08 Canonical Discriminant Analysis for Solution of 10 Clusters
Multivariate Test Statistics Eigenvalues = CanR^2/(1- CanR^2)

	Canonical Correlation	CanR^2	Eigenvalue
1st Canonical Correlation	0.838	0.703	2.367
2nd Canonical Correlation	0.774	0.599	1.493
3rd Canonical Correlation	0.520	0.271	0.371

Notes: CanR^2 is Squared Canonical Correlation.

Table 4-09 presents the values of *Wilks' Lambda*, which is one of the multivariate statistics that can be used to assess the results of the discriminant analysis. It tests the hypothesis that the cluster means are equal in the data set. From the results given in Table 4-09, one can tell that both solutions reject the hypothesis of equal cluster means.

As a multivariate measure of cluster differences over the discriminating variables, *Wilks' Lambda* reveals the degree of differentiation among the clusters. Its value ranges from a maximum of 1.0 to a minimum of 0. Because lambda is an 'inverse' measure, values of lambda which are near zero denote high discrimination (the group centroids are greatly separated and very distinct relative to the amount of dispersion within the groups). As lambda increases toward its maximum value of 1.0, it is reporting progressively less discrimination. When lambda equals 1.0, the group centroids are identical; there are no group differences (Klecka, 1980, p. 38-39).

Compared to the value of 0.13 for the solution of 7 clusters, the lambda value of 0.08 for the solution of 10 clusters implies a higher degree of discrimination.

One should note that the comparison of the two sets of clusters by canonical discriminant analysis is not based on any formal hypothesis test of two sample means and variances. It is a practical attempt to analyze and compare different solutions of the clustering procedure. The conclusions provide some guidance for the researcher to identify the "better" from a set of alternative solutions.

Table 4-09 Multivariate Statistics(Wilks' Lambda) and F Approximations

Cluster Solutions	Value	F	P-value
10 Clusters	0.08	422.7901	0.0001
7 Clusters	0.13	447.4004	0.0001

From the above discussion of the results of the discriminant analysis, the statistics that are provided on the clusters reveal important information on the extent to which the clustering procedure has performed on two different solutions. However, one cannot draw a final conclusion yet without checking if the classification analysis served its main purpose, that is, whether it eliminates the problem of neglected heterogeneity in the subgroup data.

3.2. Tests for Neglected Parameter Heterogeneity

Another way to perform external validation on the final solution of the clustering process is the test for neglected parameter heterogeneity.

When neglected parameter heterogeneity exists in cross-section data, the appropriate classification that groups the data into homogeneous clusters will eliminate the differentiation among the observations within the clusters. To find and eliminate neglected heterogeneity is one of the purposes of this study. Hence, to the extent that one cannot find evidence for neglected parameter heterogeneity in the clusters, the suggested research methodology is supported and the validity of the particular cluster solution is corroborated.

As discussed in Chapter 2, the test of neglected heterogeneity is implemented in a dynamic process. Tests for heteroskedasticity and wrong functional form are employed to distinguish the problem of neglected heterogeneity from the problem of a wrong functional form. These tests have been performed on the unclassified data set for both the simple log-linear Cobb-Douglas production function and the more flexible translog production function. The results from these tests (Table 2-05) have indicated a problem of neglected heterogeneity with the unclassified data. The same tests are also performed on both models for the subgroup data sets obtained from the cluster analysis. The general-to-specific specification methodology used for translog model with the unclassified data is also applied to the translog model with the classified subgroup data sets.

Tables 4-10 and 4-11 show the results of tests for neglected heterogeneity for solutions of 7 and 10 clusters, respectively.

From the comparison of Tables 4-10 and 4-11, one finds that the test results for the Cobb-Douglas model for both solutions are all statistically significant at the one percent level, except for the *Reset* test for the fifth cluster in the first solution (7 clusters). Whether this is the problem of neglected heterogeneity or wrong functional form is still in question. By moving attention to the test results for the translog models, one hopes to get a clearer picture whether wrong functional form or neglected heterogeneity is the issue. In Table 4-10, one can find that in the first solution (7 clusters) there are only two clusters (cluster 1 and cluster 4) for which both tests for heteroskedasticity and wrong functional form are insignificant at the one percent level of statistical significance. These two clusters have 567 observations, which approximates about 32 percent of all observations in the solution. By contrast, there are four clusters for which both tests are insignificant in the second set of 10 clusters. These four clusters account for 1059 observations, which amounts to 65 percent of all observations in the solution.

Table 4-10 Tests for Neglected Heterogeneity for Clusters in Solution of 7 Clusters (Five clusters kept wirh observations over 100)

Cluster	1	2	3	4	5
Observations	294	553	194	273	468
Cobb-Douglas					
Breusch-Pagan	0	0	0	0	0
Reset	0	0	0	0	0.01
Translog Before Respecification					
Breusch-Pagan	0.16	0	0	0.01	0
Reset	0.18	0	0	0.01	0
Translog After Respecification					
Breusch-Pagan	0.43	0	0	0.01	0
Reset	0.17	0	0	0.01	0

Notes: Reset represents the Ramsey (1969) test for functional form, with second powers of the residuals being used. Probability values (p-values) are reported for all the statistical adequacy tests.

Table 4-11 Tests for Neglected Heterogeneity for clusters in Solution of 10 Clusters (Six clusters kept with observations over 100)

Cluster	1	2	3	4	5	6
Observations	294	121	194	273	371	369
Cobb-Douglas						
Breusch-Pagan	0	0	0	0	0	0
Reset	0	0	0	0	0	0
Translog Before respecification						
Breusch-Pagan	0.16	0.17	0	0.01	0.44	0
Reset	0.18	0.01	0	0	0.10	0.25
Translog After respecification						
Breusch-Pagan	0.43	0.24	0	0.02	0.31	0
Reset	0.17	0.01	0	0.06	0.10	0.29

Notes: Reset represents the Ramsey (1969) test for functional form, with second powers of the residuals being used. Probability values (p-values) are reported for all the statistical adequacy tests.

The above analysis on the test results indicates that with the correct functional form specified, the problem of neglected heterogeneity is eliminated from the majority of subgroup data sets. Based on this result, the comparison of the two alternative cluster solutions confirms that the solution based on 10 clusters for the classification process is the most appropriate choice.

CHAPTER 5

RESULTS ANALYSIS AND APPLICATIONS

A Note of Orientation

As discussed in Chapter 1, this study focuses on (a) identifying the possible problem of neglected heterogeneity in the empirical estimation process with cross-section data and (b) developing a practical solution to neglected heterogeneity with classification analysis. The idea of the classification analysis is to identify viable groups of economic actors with homogeneous behavioral response patterns. The statistical testing methodology (Zietz, 2000b) to identify the problem of neglected heterogeneity is described in Chapter 2 and is implemented on both the unclassified data set and the subgroup data sets which are obtained from the classification analysis. The test results indicate the existence of neglected heterogeneity for the unclassified data and significant improvement for the classified subgroup data.

The question that arises after the statistical analysis is whether the classification into subgroups leads to materially different economic results. Material results in the context of this study are interpreted as the estimates of total factor productivity, factor elasticities and economies of scale, which are discussed in this chapter. If the material results are different for the unclassified data and the classified subgroup data, one needs to ask whether the classification of the data into subgroups provides one with a better conceptual foundation and understanding of the results than an analysis on unclassified data. Since this study is not designed to provide specific recommendations for economic policy but instead focuses on the economic application of the material results, the discussion will emphasize the differences between the results from the unclassified data set and the ones from the classified subgroup data. An attempt is made to explain what these differences imply in more general terms for the relationship between the results that can be obtained from unclassified data relative to those from data sets that are delineated by the heterogeneous characteristics of individual groups.

Before going into the discussion, the estimation methodology is restated and the material results and corresponding statistical tests are introduced.

Based on the results of the classification analysis, which is discussed in Chapter 4, the final solution of 10 clusters is used for the remainder of the study. Six subgroups with 1622 observations are retained from the clustering solution. The other observations are left out because they are scattered in the multidimensional space defined by the classification variables.[12]

The economic model to be estimated is the translog production function model. The log-linear Cobb-Douglas production function introduced in previous chapters is estimated in order to identify the problem of neglected heterogeneity in a dynamic testing process. Since this analysis has been completed, the log-linear Cobb-Douglas model is not discussed any further in this study. The general-to-specific specification procedure used for the translog model with unclassified data is also applied to all the translog models with the group-specific data sets to test down these models to a more parsimonious form.

General statistics for the unclassified data and the subgroups are presented and discussed. It is hoped that these general statistics will provide insight into the economic behavior and characteristics of the economic agents in each group. Furthermore, these statistics along with the analysis of the material estimation results will help to reveal the relationship between economic behavior and individual heterogeneity.

All the material results described are obtained from estimates of respecified translog models for both the unclassified data set and the classified group-specific data sets.

Material Results in Discussion

Total Factor Productivity (TFP), which is introduced by Solow (1957), is widely used in economic empirical research. *TFP* measures the efficiency and effectiveness with which both labor and capital resources are used to produce output. In other words, it allows for better use of the available labor and capital resources. In this study, *TFPs* for the unclassified data and all subgroups are measured by the constant terms of the underlying regressions.

Factor Output Elasticity measures the elasticity of output with respect to the input resource in question. It is calculated as the derivative of the production function with respect to the independent variables that one is measuring. For the translog function that is used in this study, the Factor Output Elasticity for capital (K) is calculated with the following formula.

$$E_k = B_k + 2B_{kk}\,\underline{LogK} + B_{kh}\,\underline{LogH} + B_{kn}\,\underline{LogN} + B_{km}\,\underline{LogM}$$

where E_k stands for *Factor Output* Elasticity of K; the derivatives with underbars for means of the derivatives; and the Bs for parameters of the corresponding derivatives.

Economies of scale are being realized through operational efficiencies when the output increases by proportionately more than all the inputs. *Scale Elasticity* measures

[12] Not enough observations are available to form meaningful clusters for these observations.

this operational efficiency with respect to all the input resources. The scale elasticities for each model in this study are determined by summing the factor output elasticities (Beeson, 1987).

All these material results are defined and calculated as in Eff (1995). The difference to Eff's study is that (a) the results are derived for both the unclassified data set and for each of the subgroups and (b) tests of equality are performed for the different data sets.

5.1. Estimation Results

Table 5-01 presents the regression results for both the unclassified data and the subgroups from estimating the translog models, simplified with the general-to-specific method.

Table 5-01 shows that, after simplifying the equations, all the estimated models for the unclassified and subgroup data have different forms. The tests for heteroskedasticity and wrong functional form indicate that the problem of neglected heterogeneity that is identified for the unclassified data has been eliminated from subgroups 1, 2, 4 and 5.

Table 5-01 Translog Production Function Estimates for Unclassified and Subgroup Data

Variables	Unclassified	Subgroups					
		1	2	3	4	5	6
Constant	5.205	2.532	5.049	3.505	4.776	4.733	3.683
	(16.56)	(4.44)	(30.04)	(17.93)	(6.55)	(9.77)	(7.11)
LK	0.332	0.637	0.381	0.117	0.242		0.543
	(51.67)	(8.02)	(15.10)	(1.26)	(8.00)		(10.21)
LN	0.575	0.211	0.589	0.189	0.561	0.494	0.420
	(9.65)	(2.24)	(13.07)	(7.93)	(4.18)	(5.17)	(4.50)
LH	0.392	0.120	0.144	0.967	0.268	0.486	0.053
	(8.73)	(4.64)	(6.04)	(7.95)	(2.23)	(3.98)	(1.14)
LM	-0.114	0.147			0.103	0.178	0.148
	(-2.86)	(2.00)			(0.80)	(1.80)	(2.10)
LKN		0.023		-0.014		-0.019	
		(4.03)		(-1.58)		(-5.26)	
LKH					-0.045	-0.048	0.048
					(-3.05)	(-4.19)	(2.31)
LKM	-0.104	-0.118	-0.161	-0.072	-0.078	-0.143	-0.154
	(-45.87)	(-14.39)	(-11.99)	(-8.82)	(-5.23)	(-5.42)	(-6.23)
LNH	0.019			0.069	-0.019	0.031	
	(4.85)			(6.08)	(-1.48)	(2.95)	
LNM	-0.044	-0.028	-0.034	-0.031	-0.016	-0.029	-0.020
	(-12.09)	(-4.37)	(-11.44)	(-8.22)	(-1.32)	(-3.37)	(-3.12)
LHM	-0.035	-0.011		-0.072	-0.087		-0.026
	(-9.90)	(-2.48)		(-7.83)	(-5.66)		(-1.80)
LKK	0.054	-0.048	0.082	0.045	0.060	0.108	0.050
	(49.17)	(17.12)	(11.44)	(10.81)	(7.39)	(7.54)	(2.68)
LNN	0.021	0.009	0.024		0.026	0.017	0.017
	(7.22)	(2.22)	(9.36)		(4.01)	(3.45)	(4.19)
LHH			-0.014		0.075		-0.015
			(-3.12)		(6.41)		(-1.69)
LMM	0.087	0.078	0.101	0.079	0.086	0.082	0.098
	(44.80)	(15.95)	(15.89)	(16.19)	(8.86)	(6.61)	(9.70)
R	0.99	0.99	0.99	0.99	0.99	0.99	0.99
P-values:							
BP	0	0.43	0.24	0	0.02	0.31	0
Reset	0	0.17	0.01	0	0.06	0.10	0.29
Observations	1897	282	111	183	259	359	356

Notes: The prefix *L* stands for the natural logarithm. The dependent variable is *LQ* (Value of Shipment). *K* is capital input by subtracting total payment from value added, *H* is hours of production labor, *N* is non-production employment and *M* is material input. The naming convention for the independent variables follows this example for *K*: *LK* is logarithm of *K*, *LKN* is joint product of *LK* and *LN*, *LKK* is the second power of *LK*. T-values are reported in parenthesis below the estimated coefficients. *BP* stands for the Breusch-Pagan (1979) test for heteroskedasticity. *Reset* represents the Ramsey (1969) test for functional form, with second powers of the residuals being used. Probability values (p-values) are reported for the statistical adequacy tests.

5.2. General Statistics

General statistics for the unclassified and subgroup data are listed in Table 5-02 and 5-03. It is hoped that these statistics provide better insight into what parameter heterogeneity among these groups means in terms of economic behavior and individual characteristics. The discussion on these statistics is presented in following sections along with the analysis of the material estimation results.

Table 5-02 Input Expenditures in Manufacturing (Mill. 1982 $)[13]

	Unclass	Group1	Group2	Group3	Group4	Group5	Group6
Total Value of Shipment	1,796,602	402,602	204,360	29,099	44,464	671,825	101,574
Capital Expenditures	415,607	89,839	48,249	3,170	10,307	177,962	21,770
Nonproduction Labor Salaries	167,137	27,604	27,453	1,091	2,781	73,424	5,575
Production Labor Wages	191,255	32,252	28,139	2,696	7,086	79,670	11,253
Cost of Materials	1,017,326	254,739	98,261	21,854	23,861	338,785	62,589
Capital as Percent of VS	23.13	22.30	23.60	10.90	23.20	26.50	21.40
Salaries as Percent of VS	9.30	6.90	13.40	3.70	6.30	10.90	5.50
Wages as Percent of VS	10.64	8.00	13.80	9.30	15.90	11.90	11.10
Materials as Percent of VS	56.63	63.30	48.10	75.10	53.70	50.40	61.60

Note: VS stands for Total Value of Shipment.

[13] **The original input figure is used for capital expenditure. It is different from K that is redefined by Eff (1995) in his study.**

Table 5-03 Average Measures of Individual Characteristics

	Unclassified	Group1	Group2	Group3	Group4	Group5	Group6
Educational Level	12.10	12.40	12.40	11.80	11.30	12.40	12.10
Hourly Wage	7.55	8.49	9.60	6.59	5.90	8.22	6.65
Labor Share as Percentage	45.08	35.54	51.26	56.72	49.50	43.58	43.11
Median Firm Size	55.70	54.06	87.11	37.38	65.28	59.43	44.48
Per Capita K-Expenditure	3.43	5.19	2.63	2.26	1.30	2.63	2.45
ZABR	4	3	3	6	6	3	5

Note: Educational Level is measured in years; *Hourly Wage* in dollar; *Firm Size* in number of establishments; *Per Capita new capital expenditure* in thousands of dollars.

5.3. Total Factor Productivity

Total Factor Productivity (TFP) is determined by a host of causes that interact with one another in subtle ways. Key causal factors include:

- Changes in the Quality of Labor – improvement in the variables that affect the productive capacity of workers. Educational level and up-to-date working skills are the key determinants.
- Capital Deepening – A rise in the amount of capital per worker (Solow, 1957).
- Technical Progress – Advances in knowledge includes research and development, technology catch-up, innovations, etc.

Table 5-04 lists the *TFPs* from all regressions. *TFP* for the unclassified model is higher than the *TFPs* for all the subgroups and the statistical tests indicate clear differences between the unclassified model and all the subgroups. *TFPs* also vary among the subgroups. Similar tests of equality are performed for *TFPs* among the subgroups and in the majority of cases the results show that they are different from each other.[14]

[14] The actual results are not presented here because of the overwhelming information from the massive number of calculations that are performed.

Table 5-04 Total Factor Productivity for Unclassified and Subgroup Models

	Unclassified	Group1	Group2	Group3	Group4	Group5	Group6
TFP	5.205	2.532	5.049	3.505	4.776	4.733	3.683
t-value		-76.9	-8.9	-105.1	-9.4	-17.7	-53.6

Note: t-value obtained from the test of equality of each TFP from each individual subgroup data comparing with TFP from unclassified data.[15]

If one studied only the results from the unclassified data, the total factor productivity figure would suggest that all the economic agents (US counties) are performing well and the same across the board. If one compares the *TFPs* for the subgroups to that of the unclassified data, one arrives at a different conclusion.

In Tables 5-01 and 5-02, one can find that different characteristics of individual groups play important roles for the behavior of each homogeneous group. For example, groups 1, 2 and 5 have a value of 3 for variable *ZABR*. This small value reveals that the counties of these three groups are closer to an MSA area and have larger populations than those in the other groups. Because urban areas usually possess both greater agglomeration economies and higher levels of human capital, *TFPs* for these counties should be higher than the one for others. This is borne out for groups 2 and 5, which have higher educational levels, higher hourly wages and a larger median firm size. As a consequence, there are higher *TFP* values.

Group 1 shares many features with groups 2 and 5. It has the highest educational level and also the highest per capita capital expenditures, but the contribution of total capital expenditures to the total value of shipments is lower than those for groups 2 and 5. The biggest difference for group 1 is that it does not take advantage of the higher level of human capital. The contribution to total shipped value from production labor wages is

[15] The formula is found in DeGroot (1975, p.433) and is used by Eff (1995).

only 8 percent, which is the lowest among all groups. On the other hand, *TFP* for group 4 is much higher than for group 1. Group 4, with lower educational levels and hourly wages than group 1, contributes up to 16 percent from production labor to total output. Also, group 4, with the lowest per capita capital expenditure among all the groups, generates a higher contribution from capital input to the final product than group 1. This paradox makes an interesting point that would provide one with motivation to look further into other causal factors such as capital structure, technology improvement and management style in the production process.

Based on the above description, one would have a clearer understanding of the relationship between economic behavior and individual characteristics of the economic agents by analyzing the heterogeneity among them, instead of believing that all the economic agents are performing the same.

5.4. Factor Elasticities

The factor output elasticities derived from the estimated models for the unclassified data set and the 6 subgroups are listed in Table 5-05. The output elasticity of production labor (*H*) is higher for groups 2, 3, 4 and 6, of which three are groups of counties located in rural areas. By contrast, the output factor elasticity of non-production labor (*N*) is higher in groups 2 and 5, which are groups of counties that are located near MSA areas. These findings match those of Eff's (1995) comparison of MSA counties with those of rural counties. Both studies reveal the same pattern of factor output elasticities: in general, the elasticities are higher in the regions that most intensively use a particular factor.

Table 5-05 Factor Output Elasticities

	Unclassified	Group1	Group2	Group3	Group4	Group5	Group6
E_H	0.08	0.05	0.08	0.08	0.10	0.05	0.08
		(-76.7)	(1.3)	(4.5)	(45.4)	(-117.5)	(-6.0)
E_K	0.25	0.26	0.29	0.14	0.27	0.30	0.24
		(44.5)	(70.1)	(-145.0)	(95.2)	178.8)	(-5.4)
E_M	0.60	0.64	0.53	0.73	0.57	0.56	0.63
		(154.1)	(-109.9)	(163.1)	(-81.2)	(-166.5)	(115.2)
E_N	0.08	0.05	0.10	0.05	0.06	0.09	0.06
		(-146.4)	(37.7)	(-52.6)	(-69.9)	(36.7)	(-149.5)

Notes: *E* stands for factor output elasticity and the subscripts for each input factor; numbers in parentheses are t-values for the null hypothesis that the subgroup statistic equals the one for the unclassified model. The p-values for all tests are 0.00.

Closely examining the output elasticities of production labor (*H*) and capital input (*K*) and comparing group 4 with group 1, one will find that group 4 shows a higher output elasticity of capital input and a much higher output elasticity of production labor than group 1. Table 5-03 reveals that group 4 has much lower per capita expenditure and possesses lower human capital than group 1. In line with the analysis in the above section on *TFP*, these findings indicate that group 4 manages to put available resources to better use and, therefore, can produce more effectively with less labor and capital than group 1.

5.5. Scale Elasticity

Table 5-06 lists the scale elasticities calculated for all the models estimated with unclassified and subgroup data.

Table 5-06 Economies of Scale Elasticities

	Unclassified	Group1	Group2	Group3	Group4	Group5	Group6
SC	.009	.001	.004	.003	.002	.001	.005
	(9.35)	(0.79)	(1.77)	(0.52)	(0.66)	(0.68)	(3.92)
p-value	0.00	0.43	0.08	0.60	0.51	0.50	0.00

Notes: *SC* stands for Scale Elasticity. Numbers in parentheses are t-values for the null hypothesis of constant returns to scale, for which the scale elasticity equals 1; p-values for the test are also reported.

The scale elasticity for the model estimated with unclassified data does not indicate constant returns to scale. By contrast, unitary scale elasticities prevail in majority of the subgroups.

In conclusion, if one compares the results for the unclassified data analysis to those for the subgroups, it is difficult to conclude that the results for the unclassified data can be representative for all the economic agents in this particular study. By just relying on the unclassified data that neglect heterogeneous characteristics among the economic agents, many important aspects of economic behavior are lost, the resulting estimates tend to be biased and some even contradict the results that can be obtained from the classified data. Estimation and analysis of the classified data not only gives one better insight into the relationship between the economic behavior and individual heterogeneity but also provides motivation for further research on the characteristics that affect the underlying economic processes.

CHAPTER 6

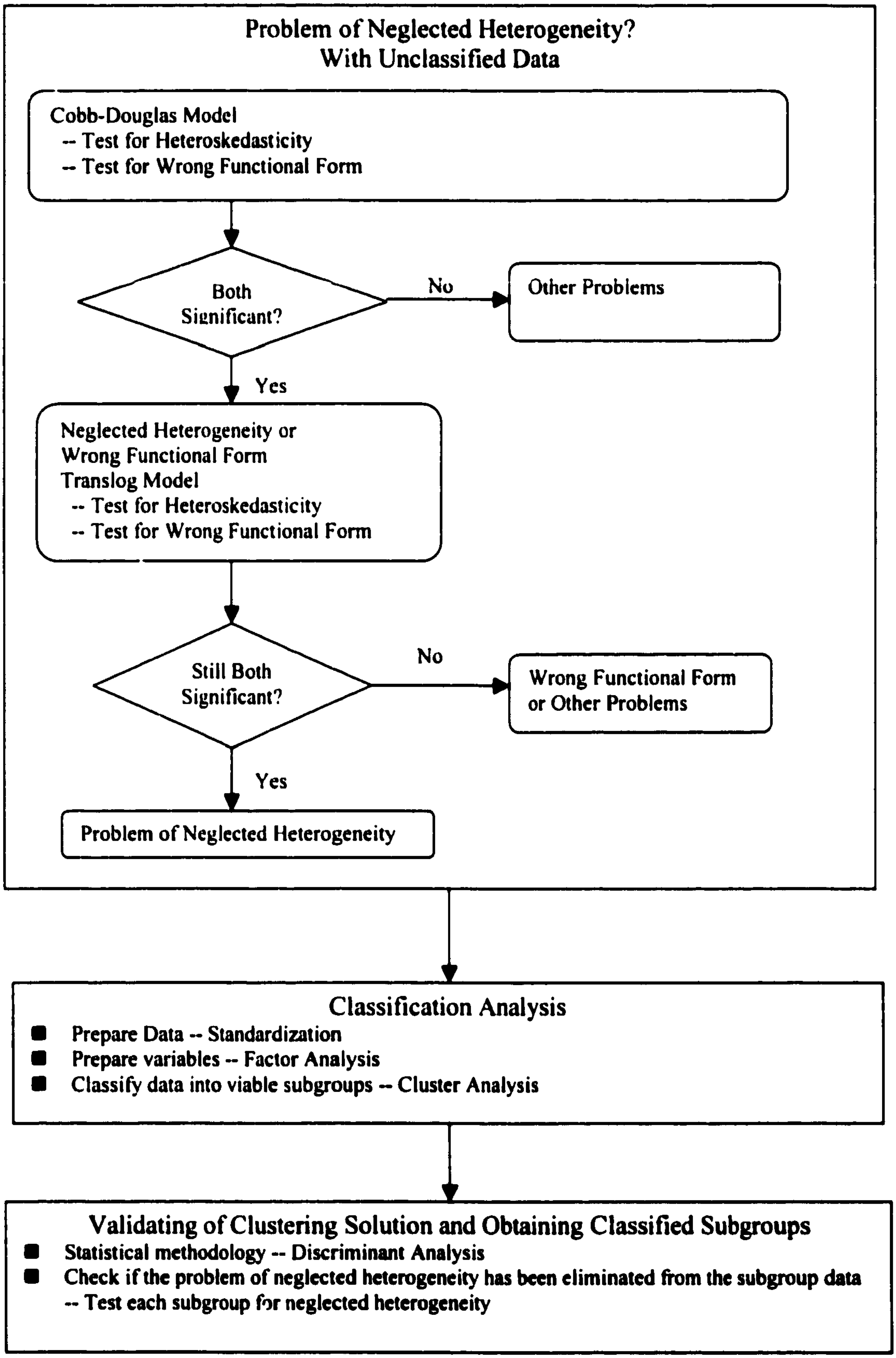
SUMMARY AND CONCLUSIONS

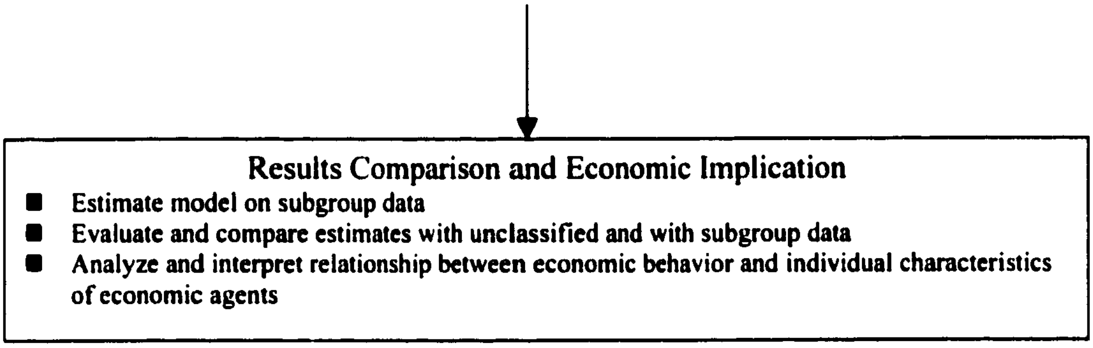
As discussed at the beginning of this study, neglecting individual heterogeneity and doing research instead on unclassified data only will most likely result in estimates that arc biascd and impossiblc to intcrprct in any meaningful way. The purpose of this study has been to demonstrate a classification methodology for cross-sectional data that incorporate behavioral heterogeneity. The suggested methodology allows the researcher to identify viable groups of economic actors with homogeneous behavioral response patterns.

6.1. Summary of Empirical Process

The flowchart on the next page illustrates what has been done in this study.

Figure 6-01 Flowchart of Summary


Problem of Neglected Heterogeneity?
With Unclassified Data
Cobb-Douglas Model
-- Test for Heteroskedasticity
-- Test for Wrong Functional Form
Both
Significant?
No
Other Problems
Yes
Neglected Heterogeneity or
Wrong Functional Form
Translog Model
-- Test for Heteroskedasticity
-- Test for Wrong Functional Form
Still Both
Significant?
No
Wrong Functional Form
or Other Problems
Yes
Problem of Neglected Heterogeneity
Classification Analysis
Prepare Data -- Standardization
Prepare variables -- Factor Analysis
Classify data into viable subgroups -- Cluster Analysis
Validating of Clustering Solution and Obtaining Classified Subgroups
Statistical methodology -- Discriminant Analysis
Check if the problem of neglected heterogeneity has been eliminated from the subgroup data
-- Test each subgroup for neglected heterogeneity


Results Comparison and Economic Implication
■ Estimate model on subgroup data
■ Evaluate and compare estimates with unclassified and with subgroup data
■ Analyze and interpret relationship between economic behavior and individual characteristics of economic agents

In econometric practice, the problem of neglected parameter heterogeneity arises when the constructed model is estimated with undifferentiated data and systematic differences in the behavioral response patterns of the individual economic agents in the data set are ignored. Parameter heterogeneity is interpreted as a case where the regression coefficients of the estimating model differ across individual observations or groups of observations.

Zietz (2000b) has suggested a simple testing methodology for neglected heterogeneity on the basis of a set of Monte-Carlo experiments. The suggested test strategy involves a dynamic application of tests for heteroskedasticity and wrong functional form.

The present study uses the 1982 manufacturing census data for US counties (Eff, 1995) to estimate county-specific production functions. The study starts with the testing method suggested by Zietz (2000b) to identify whether neglected heterogeneity exists in the unclassified data. This is shown in the first part of the flowchart.

The testing procedure is implemented first by estimating a simple log-linear Cobb-Douglas production function for the unclassified data and carrying out the tests for heteroskedasticity and wrong functional form. Statistical significance on both tests indicates the possibility of a problem of either neglected heterogeneity or wrong functional form. To distinguish between the two problems, the testing procedure continues to estimate the more flexible translog production function for the unclassified data set and executes the same tests for heteroskedasticity and wrong functional form. Since both tests remain highly significant statistically, there is reason to believe that neglected heterogeneity is a potential problem for the original unclassified data set.

There are many economic researchers who have devoted great effort to developing methodologies, such as exact aggregation, micro simulation and weight adjusted aggregation models, to incorporate the impact of individual heterogeneity on economic behavior. In his book on clustering and aggregation in economics, Fisher (1969) concludes that a need is felt for solutions to clustering and aggregation problems that arise in economics. Based on his research experience and observation, he points out that the solutions should be specific and of a form that could be applied to concrete, numerical problems, and that a reduction of the data to a smaller scale is most desirable for easier management and comprehension.

Faced with the potential of neglected heterogeneity in the unclassified data, this study has tried to develop an objective and effective classification methodology with a series of multivariate statistical procedures to discover possible homogeneous subgroups with similar individual characteristics that affect their economic behavior. The benefit of the methodology is a potentially much better understanding of economic behavior because individual behavioral heterogeneity is revealed.

The classification methodology developed in this study is discussed in a practical manner. In Chapter 3, the theoretical and practical issues are laid out and discussed.

These issues include data preparation, the necessity and method of reducing the clustering dimensional space by factor analysis, as well as the commonly used classification method called cluster analysis.

In Chapter 4, the results from the classification analysis are analyzed with a series of statistical tests to validate and determine the final solution of clusters. Two clustering solutions, a solution with 7 clusters and one with 10 clusters, are retained for further analysis. The two clustering solutions are compared and evaluated with discriminant analysis and tests for neglected heterogeneity. Discriminant analysis is applied as a statistical tool to derive various statistics on these two clustering solutions. The test for neglected heterogeneity illustrated in the first step is also executed on all subgroup data set for the two solutions. Both discriminant analysis and the test for neglected heterogeneity validate the two solutions and indicate that the solution with 10 clusters is the most appropriate choice.

There are 6 subgroups with sufficient observations (more than 100) retained from the solution of 10 clusters for the economic analysis following the classification analysis. Translog production function models are specified based on the well-known general-to-specific methodology and estimated for each of the 6 subgroup data sets. Chapter 5 analyzes and discusses what the estimates imply in terms of general economic characteristics, such as differences among the groups in terms of educational level and per capita capital expenditure, and in terms of more specific economic results, such as total factor productivity and factor elasticities. The analysis of these results across groups raises serious doubts of the usefulness of the model that is estimated with unclassified data. It also reveals some interesting relationships between economic behavior and individual characteristics among the subgroups and sheds some light on further economic research that may be useful as follow-up studies.

Based on the summary given above, four aspects of this study stand out in terms of their contributions to dealing with the problems associated with neglected heterogeneity.

First, the testing methodology (Zietz, 2000b) for the problem of neglected heterogeneity is successfully implemented in practice. The test results confirm the existence of the problem of neglected heterogeneity for the original unclassified cross-section data used in this study.

Second, this study is able to develop an objective and effective classification methodology for a given specific data set to discover homogeneous subgroups with similar individual characteristics that are related to their economic behavior.

Third, as demonstrated by this study, one can establish a close and economically meaningful relationship between group-specific economic behavior and the individual characteristics of the economic agents in each group. Such relationships should be of significant economic value. For example, economic policies that try to target specific groups need exactly this type of group-specific information. It is also valuable for such issues as deriving forecasts for the aggregate of all observations. Specifically, the results provide not only useful weighting that can be used to aggregate the subgroups but also the different behavioral patterns that need to be aggregated.

Finally, the procedures developed in this study apply to many research areas in addition to economics. Specifically, they apply whenever one is facing the problem of neglected heterogeneity. The area of economic education would be a good example and is illustrated in the section.

6.2. Applications in Economic Education

It is important to note that the classification methods developed in this study with a production function example is applicable to a wide rage of applied questions. An example from the area of educational research will help illustrate this. The same or a very similar type of classification analysis can be applied to study thc learning behavior of different student groups based on data that measure class, course, teacher, and student characteristics, such as class size and faculty teaching load, student performance and individual student characteristics, such as age and grade point average. Such analysis would be helpful in providing better insight into the relationship between a student's classroom success and his/her individual characteristics. The results of such a study could be used to improve administrative decisions as well as classroom pedagogy. In what follows, a simple illustration of these points is provided.

In teaching economics and other courses, student evaluations of the course being taught are a very important feedback mechanism. The information returned from students reflects how the material is introduced, how the class is taught and perceived by students, and how effective the class is in terms of delivering knowledge. A benchmark study of the teaching of introductory economics presented by Saunders (1994), which uses the TUCE III [16] database, reports how introductory economics was taught in 53 different two-year and four-year colleges and universities in the U.S. during the 1989-1990 academic year. This benchmark study provides a wide range of information on the teaching of the courses and the characteristics of the classes and of the individual instructors and students. Because of the nature of individual differences in such a cross-section database, studies conducted from the survey are likely to be subject to the same problem of neglected heterogeneity as the research that has been undertaken as a demonstration project in this paper. A methodology similar to that employed by this study would be a possible solution for this kind of problem.

In his study, Saunders (1994) emphasizes many differential characteristics among classes, instructors, and students, such as class size, instructor's teaching experience, student's previous economic education, and time spent studying, and so on. The TUCE data have been used for many studies (e.g. Cochran and Zietz, 1996, Zietz and Cochran, 1998, Kennedy and Siegfried, 1997). Researchers have paid great attention to the detailed characteristics information and made efforts to try to incorporate as much information as possible to explain teaching and learning behavior from different angles. However, many of these studies find that it is difficult to organize all this information into one unclassified model. For example, one typically finds that student behavior is modeled either without incorporating heterogeneity in behavioral responses among classes, instructors, and students altogether or by assuming that the differences can all be captured by the addition of simple intercept dummy variables. For the TUCE cross-section data, such attempts always seem to lead to the problems of heteroskedasticity and possibly wrong functional form. As amply discussed in this study, these two statistical regression problems are most likely caused by the problem with neglected heterogeneity and, therefore, the estimation results on most of the studies on the TUCE data will be difficult to interpret in a meaningful way. On the other hand, research that adopts the cluster

[16] TUCE III stands for the third edition of the "Test of Understanding College Economics."

analysis method presented here, by taking advantage of the information at the individual student level for subgrouping the data set, would most likely provide better insight into the determinants of learning.

Suppose, for example, that one is conducting a study on the effectiveness of student learning based on the TUCE III data set introduced above. In designing such a study, one may want to use a student's performance on the TUCE test or the difference between his/her pre- and post-test results as the dependent variable. As for the independent variables, there are many variables to choose from, such as class size, number and length of class meetings, use of class time by instructors, teaching style of instructors, instruction method, instructors' after-class availability, etc. All this information and more are available in the database. A simple regression model can be easily established and estimated. But there are some fundamental questions that need to be answered. First, are all observations (individual students) homogeneous across the board in their behavior? Second, would the data used for the regression model by themselves be able to generate meaningful estimates, or is there other important information in the data set that has not been included in the regression? Clearly, there is much more information in the data set than is typically incorporated in a regression model. An example will help illustrate this and suggest how to answer the above questions.

Consider two groups of instructors and students who may be from two different types of schools, top ranked universities (group 1) and two-year colleges (group 2). It is reasonable to believe that the instructors in group 1 have on average a terminal degree and a better understanding of the subject matter they are teaching than their counterparts in group 2. Also, the students in group 1 are more likely to have better SAT scores, higher expectations, more time to spend studying than the students in group 2. One would rationally expect that if one estimates the effect of classroom teaching on students' TUCE scores, the statistics would result in significant bias and "strange" parameter estimates if one combined the observations from both groups in one data set and estimated a single equation on the assumption that the combined data set reflects homogeneous behavioral responses.

To solve the problem, one could take advantage of the procedures developed in this study. With classification analysis, one can divide the original data set into subgroups based on the available individual characteristics of both instructors and students. Estimation on such subgroups is more likely to give unbiased results and meaningful interpretations.

There are many other examples from the education field, such as educational program marketing (Boughan, 1991) and educational psychology (Wang, 1994) where neglected heterogeneity could be a potential problem. For example, the teaching of English to new immigrants may have much better results if it were done in region-specific groups based on differences in cultural and linguistic backgrounds (Kojic-Sabo and Lightbown, 1999). Many of these educational studies provide a potential opportunity for applying the classification methodology developed here to treat the problem of neglected heterogeneity.

BIBLIOGRAPHY

Anderberg, M.R. 1973, *Cluster Analysis for Applications*, New York: Academic Press, Inc.

Beeson, E.P. 1990, "Sources of the Decline of Manufacturing in Large Metropolitan Areas", *Journal of Urban Economics*, Vol. 28, 71-86.

Berlage, L. and Terweduwe, D. 1988, "The Classification of Countries by Cluster and by Factor Analysis", *World Development*, Vol. 16, 1527-1545.

Blundell, R., Pashardes, P. and Weber, G. 1993, "What Do We Learn About Consumer Demand Patterns from Micro Data?" *The American Economic Review* Vol. 83, 570-597.

Bock, H.H. 1985, "On Some Significance Tests in Cluster Analysis", *Journal of Classification*, Vol. 2, 77-108.

Boughan, K. 1991, "A Cluster Analysis of the 1985-1989 Non-Credit Student Body: Implementing GAO-Demographic Marketing at P.G.C.C., Part II", *Market Analysis MA91-5, Prince George's Community College, Office of Institutional Research and Analysis.*

Breusch, T.S. and Pagan, A.R. 1979, "A Simple Test for Heteroscedasticity and Random Coefficient Variation", *Econometrica*, Vol. 47, 1287-1294.

Brynjolfsson, E. and Hitt, L. 1995, "Information Technology As a Factor of Production: The Role of Differences Among Firms", *Economics of Innovation and New Technology*, Vol. 3, 183-200.

Calinski, T. and Harabasz, J. 1974, "A Dendrite Method for Cluster Analysis", *Communications in Statistics*, Vol. 3, 1-27.

Census of Population, 1980, Vol. 1, *U.S. Bureau of the Census.*

Census of Manufactures, 1982, Vol. 3, *U.S. Bureau of the Census.*

Chesher, A. 1984, "Testing for Neglected Heterogeneity", *Econometrica*, Vol. 52, 865-872.

Christensen, L.R. and Jorgenson, D.W. 1969, "The Measurement of U.S. Real Capital Input, 1929-1967", *Review of Income and Wealth*, Vol. 15, 293-320.

Cochran, H.H. and Zietz, J. 1996, "How to Effectively Manage Principles of College Economics: New Evidence from a Large National Database", Journal of the Tennessee Economics Association, Vol. 1, 31-38.

Dubes, R. and Jain, A.K., 1979, "Validity Studies in Clustering Methodologies", *Pattern Recognition*, Vol. 11, 235-254.

Duda, R.O. and Hart, P.E. 1973, *Pattern Classification and Scene Analysis*, New York: John Wiley & Sons, Inc.

Eff, E.A. 1995, "Urban-Rural Differences in Manufacturing Technology: U.S. Counties, 1982 and 1987", Working Paper, Department of Economics and Finance, Middle Tennessee State University, Murfreesboro, TN.

Everitt, B.S. 1974, *Cluster Analysis*, London: Heineman Educational Books Ltd.

Everitt, B.S. 1979, "Unsolved Problems in Cluster Analysis", *Biometrics*, Vol. 35, 169-181.

Fisher, D.W. 1969, *Clustering and Aggregation in Economics*, Baltimore: The Johns Hopkins Press.

Gilbert, C.L. 1986, "Professor Hendry's Econometric Methodology", *Oxford Bulletin of Economics and Statistics*, Vol. 48, 283-307.

Hartigan, J.A. 1985, "Statistical Theory in Clustering", *Journal of Classification*, Vol. 2, 63-76.

Heckman, J.J. and Sedlacek, G. 1985, "Heterogeneity, Aggregation, and Market Wage Functions: An Empirical Model of Self-Selection in the Labor Market." *Journal of Political Economy*. Vol. 93, 1077-1125.

Heikkila, J.E. 1996, "Are Municipalities Tieboutian Clubs?" *Regional Science and Urban Economics*, Vol. 26, 203-226.

Jain, K.A. and Dubes, C.R. 1988, *Algorithms for Clustering Data*, Englewood Cliffs, NJ: Prentice Hall.

Ketchen, D.J. and Shook, C.L. 1996, "The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique", *Application of Cluster Analysis*, John Wiley & Sons, Ltd., 441-458.

Kim, Jae-On and Mueller, W. Charles, 1978, *Factor Analysis*, A SAGE University Paper, Sara Miller McCune, Sage Publications, Inc.

Kirman, P.A. 1992, "Whom or What Does the Representative Individual Represent?" *Journal of Economic Perspectives*, Vol. 6, 117-136.

Klecka, R.W. 1980, *Discriminant Analysis*, A SAGE University Paper, Sara Miller McCune, Sage Publications, Inc.

Kojic-Sabo, I. and Lightbown, M.P. 1999, "Students' Approaches to Vocabulary Learning and Their Relationship to Success." *Modern Language Journal*, Vol. 83, 176-192.

Lewbel, A. 1989, "Exact Aggregation and A Representative Consumer", *The Quarterly Journal of Economics*, August, 621-633.

MacQueen, J.B. 1967, "Some Methods for Classification and Analysis of Multivariate Observations", *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 281-297.

Milligan, G.W. 1980, "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algoritnms", *Psychometrika*, Vol. 45, 325-343.

Milligan, G.W. and Cooper, M.C. 1985, "An Examination of Procedures for Determining the Number of Clusters in a Data Set", *Psychometrika*, Vol. 50, 159-179.

Punj, G. and Stewart, D.W. 1983, "Cluster Analysis in Marketing Research: Review and Suggestions for Application", *Journal of Marketing Research*, Vol. 20, 134-148.

Ramsey, J.B. 1969, "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis", *Journal of the Royal Statistical Society*, Vol. 31, 350-371.

Sarle, W.S. 1983, "Cubic Clustering Criterion", *SAS Technical Report A-108, SAS Institute, Inc.*

SAS/STAT® User's Guide, 1990, Version 6, SAS institute, Inc.

Solow, R. 1957, "Technical progress and the aggregate production function", *Review of Economics and Statistics*, Vol. 39, 312-320.

Stoker, M.T. 1993, "Empirical Approaches to the Problem of Aggregation Over Individuals", *Journal of Economic Literature*, Vol. 31, 1827-1874.

Ueltschy, C.L. 1997 "The Influence of Acculturation on Advertising Effectiveness to the Hispanic Market", *Journal of Applied Business Research,* Vol. 13, 87-101.

Wang, B. 1994, "An Empirical View on Performance Indicators in Higher Education of Taiwan", *Journal of Education and Psychology,* Vol. 17, 61-98.

Zietz, J. and Cochran, H.H. 1998, "How Much Does Teaching Methodology Matter for Learning Principles of Economics?" *Kentucky Journal of Economics and Business,* Vol. 17, 39-54.

Zietz, J. 2000a, "Heteroskedasticity and Neglected Parameter Heterogeneity", Working Paper, Department of Economics and Finance, Middle Tennessee State University, Murfreesboro, TN.

Zietz, J. 2000b, "Neglected Hetrogeneity and Spurious Non-Linearity", Working Paper, Department of Economics and Finance, Middle Tennessee State University, Murfreesboro, TN.