PN-ABZ-946

RESEARCH REPORT

104

ROLE OF FARM-LEVEL DIVERSIFICATION IN THE ADOPTION OF MODERN TECHNOLOGY IN BRAZIL

Marc Nerlove
Stephen Vosti
Wesley Basel

IFPRI Research Reports
Publications Review Committee and Procedures

Lawrence Haddad, Chairman
Suresh Babu
Christopher Delgado
Hans Lofgren
Ruth Meinzen-Dick
Philip Pardey
Agnes Quisumbing
Barbara Rose (ex officio)

All manuscripts submitted for publication as IFPRI Research Reports undergo extensive review. Prior to submission to the Publications Review Committee, each manuscript is circulated informally among the author's colleagues, presented in a formal seminar, and reviewed by two IFPRI reviewers. Following submission of the manuscript to the Committee, three additional reviewers—at least two external to IFPRI and one from the Committee—are selected to review the manuscript. Reviewers are chosen for their expertise in the manuscript's subject matter and methodology and, when applicable, their familiarity with the country setting. The Committee provides the author its reaction to the reviewers' comments. After revising as necessary, the author resubmits the manuscript to the Committee with a written response to the reviewers' and Committee's comments. The Committee then makes its recommendation on publication of the manuscript to the Director General of IFPRI. With the Director General's approval, the manuscript becomes part of the IFPRI Research Report series.

IFPRI Board of Trustees

David E. Bell
Chairman, U.S.A.

Gordon Conway
United Kingdom

Baba Dioum
Senegal

Wenche Barth Eide
Norway

Godfrey Gunatilleke
Sri Lanka

Ibrahim Saad Ahmed Hagrass
Egypt

Uwe Holtz
Federal Republic of Germany

Susan Horton
Canada

James Ingram
Australia

Nora Lustig
Mexico

Solita Monsod
Philippines

Benno Ndulu
Tanzania

I. G. Patel
India

Martin Piñeiro
Argentina

Hiroya Sano
Japan

Per Pinstrup-Andersen
Director General
Ex Officio, Denmark

B

PN-ABZ-946

ROLE OF FARM-LEVEL DIVERSIFICATION IN THE ADOPTION OF MODERN TECHNOLOGY IN BRAZIL

Marc Nerlove
Stephen Vosti
Wesley Basel

Research Report 104
International Food Policy Research Institute
Washington, D.C.

C

Copyright 1996 International Food Policy Research Institute.

All rights reserved. Sections of this report may be reproduced without the express permission of but with acknowledgment to the International Food Policy Research Institute.

Library of Congress Cataloging-in-Publication Data

Nerlove, Marc, 1933–
Role of farm-level diversification in the adoption of modern technology in Brazil / Marc Nerlove, Stephen Vosti, and Wesley Basel.
p. cm. — (Research report ; 104)
"December 1996."
Includes bibliographical references (p.).
ISBN 0-89629-107-3
1. Agricultural innovations—Brazil—Minas Gerais. 2. Agriculture—Technology transfer—Brazil—Minas Gerais. 3. Agricultural productivity—Brazil—Minas Gerais. 4. Farms, Small—Brazil—Minas Gerais. I. Vosti, Stephen A. II. Basel, Wesley. Title. IV. Series: Research report (International Food Policy Research Institute) ; 104.
S494.5.I5N465 1996
338.1'6'098151—dc21 96-49242
CIP

CONTENTS

List of Tables	iv
List of Illustrations	v
Foreword	vii
Acknowledgments	viii
1. Summary	1
2. Introduction	3
3. Data	6
4. Farm-level Diversification and Product Mix	8
5. Cluster Analysis	16
6. Univariate Analyses	25
7. Longitudinal Analysis	38
8. Conclusions and Policy Implications	42
Technical Appendix	44
Bibliography	51

TABLES

1. Product diversification matrix, 1982 8
2. Categories of agricultural production in the PRODEMATA survey 9
3. Corn price reports of sales versus on-farm consumption, 1979–84 11
4. Mean and standard deviation of prices across all farms reporting a price, 1979–84 12
5. Average product prices relative to the corn price, major products, 1979–84 14
6. Index of modern input use, by category, 1984 15
7. Cluster results, including off-farm labor, 1979 19
8. Cluster results, including off-farm labor, 1980 19
9. Cluster results, including off-farm labor, 1981 20
10. Cluster results, including off-farm labor, 1982 20
11. Cluster results, including off-farm labor, 1983 21
12. Cluster results, including off-farm labor, 1984 21
13. Descriptive statistics within clusters, 1984 22
14. Components of the index of modern inputs, mean values by cluster, 1984 23
15. Geographic distribution of clusters, 1984 24
16. Univariate relations 27
17. Bivariate relations 35
18. Descriptive statistics within transitional groups, 1984 40
19. Modern input use within and across transitional groups 41
20. Changes in total output and use of modern inputs by transitional groups, 1979–84 41
21. Agglomerative methods of hierarchical cluster analysis 46
22. Assumptions and objective functions of allocative methods 49

ILLUSTRATIONS

1. Zona da Mata, Minas Gerais, Brazil 7
2. Plot 1: Total output versus farm area 29
3. Plot 2: Use of modern inputs versus total output 31
4. Plot 3: Use of modern inputs versus percent of total output in top two crops 32
5. Plot 4: Percent of total output in top two products versus total output 34

FOREWORD

In their efforts to meet the challenges of growth, poverty alleviation, and natural resource conservation goals, and in reaction to the rapidly shrinking agricultural frontier in most areas of the tropics, countries throughout the developing world are taking a fresh look at the potential contribution that less-favored agricultural lands can make. These lands are not now or ever likely to be the "bread baskets" of the developing world. Rather, they suffer from inadequate or highly erratic rainfall, poor soils, or both, and are often located long distances from markets for agricultural inputs and products. Because they offer poor returns to investments in agricultural research and infrastructure, they have generally been ignored by policymakers and researchers alike.

Although these less-favored lands do not offer textbook conditions for large returns to agricultural investments, like those realized in more-favored lands over the past three decades, they represent an untapped source of growth—one that policymakers cannot afford to overlook at a time when returns from more-favored lands are declining. Investments in these areas may also be cost-effective means to alleviate poverty and assure sustainable management of natural resources.

But will development patterns in the less-favored lands be similar to those in the more-favored lands? Should we expect to witness the rapid and geographically comprehensive changes in production patterns and technology use that accompanied Green Revolution investments in more-favored lands? Will these farmers also come to rely on just a few crops that benefit from seed, fertilizer, and irrigation technology packages? Probably not, although little is known about the nature of changes in product mix and technology use associated with development in less-favored lands and the policies needed to promote them.

This research report by Marc Nerlove, Stephen Vosti, and Wesley Basel takes up these issues in the context of one agriculturally less-favored area in Brazil, the Zona da Mata in the state of Minas Gerais. The analysis describes farm-level changes associated with an integrated rural development project, with particular emphasis on how farming systems (based on product mixes) changed during the course of that project, and how these changes in product mix affected the use of modern agricultural technologies. The authors develop useful methodologies for the analyses of these issues.

The results suggest that improvements in agricultural productivity and dramatic changes in product mix are both possible, despite the ecological and economic conditions that tend to slow down agricultural development in these areas. This should come as welcome news to researchers and policymakers throughout the developing world, especially those now searching for more geographically comprehensive development strategies to meet food security objectives sustainably in the next century.

Per Pinstrup-Andersen
Director General

ACKNOWLEDGMENTS

The underlying thesis of this report evolved in conversations with Eliseu Alves of Empresa Brasileira de Pesquisa Agropecuaria (EMBRAPA) regarding the nature of the process of technology adoption in Brazilian agriculture. We owe much to his insights.

Our work is based in part on earlier work of Isabel Busom on farm-level diversification and modern input use in the Zona da Mata, Minas Gerais, Brazil.

Our colleagues at the Federal University of Viçosa have played a central role in all of our analyses of the PRODEMATA data. Not only have they made available an important data set, but they have also provided timely and helpful assistance in interpreting the data and improving our understanding of the socioeconomic context in which agricultural production occurs in the Zona da Mata of Minas Gerais. We are especially grateful to Sonia Coelho de Alvarenga, Evonir Batista de Oliveira, Carlos Antonio Moreira Leite, Francisco Armando da Costa, João Eustaquio de Lima, José Campos, Erly Cardoso Teixeira, and others who provided valuable comments on earlier drafts of this paper.

Our colleagues at IFPRI, especially Maarten Immink and Mark Rosegrant, have given us much useful and sound advice, as have two anonymous referees of the Publications Review Committee.

We wish to thank Lucy Centeno-Teodoro and Lourdes Hinayon for their tireless and efficient secretarial support throughout this research project and Julie Witcover for expert editing assistance on earlier drafts.

Generous financial support for this project was in part provided by EMBRAPA.

The usual caveat applies regarding responsibility for errors that remain.

1

SUMMARY

Until agricultural technology is developed for exploiting the surface of the sea and areas presently too arid or too cold for cultivation, most developing countries seeking to increase agricultural production have no alternative but to increase the productivity of agriculture on land already under cultivation. In relatively high-potential areas, here defined as those with soil and climatic conditions conducive to the production of basic foodgrains, how to apply Green Revolution technologies and the policies used to promote them and sustain their efficient use are generally known. While environmental externalities associated with the intensification of agriculture using such technologies are not well understood and the implications for policy of these externalities are not always clear, methods for securing essential production increases in high-potential areas currently not using modern agricultural technologies are known.

How to raise productivity on low-potential or marginal lands is less certain. Where Green Revolution technologies have been applied to marginal lands with diversified output, expected goals in yield increases, technology adoption rates, the degree of on-farm or regional specialization, and environmental sustainability have often not been reached. To improve success rates, rural development strategists and policymakers must have a better understanding of the process of adopting technology in these areas and the constraints to adoption. This study contributes to improved understanding by focusing on the links between technology adoption in agriculture and the (at least partially) exogenously determined farm-level product mix and output diversification decisions.

The setting for this study is the Zona da Mata area of the Brazilian state of Minas Gerais—a marginal area for agricultural production, principally because of erratic rainfall, limited irrigation potential, generally poor soils, and a rugged terrain that does not lend itself to mechanization or economies of scale in most aspects of agricultural production. Longitudinal data based on a stratified random sample of farm households linked to the government's Programa de Desenvolvimento Integrado da Zona da Mata (PRODEMATA) regional development program are the empirical underpinnings of this study.

First, the product mix is quantified using a product weighting scheme to determine the degree of farm-level output diversification. Second, farms are put into categories according to their degree of output diversification using cluster analysis, a statistical method for forming groups of observations. Off-farm labor is included in the list of ways farm households diversify their incomes. The five relatively stable categories that emerged from this analysis are farms that produce coffee, corn, dairy, and rice products; off-farm labor is the fifth category. These categories then serve as a basis for examining the factors that influence adoption of technological change, both within and across farm categories.

Third, regression analyses with single and multiple variables are used to explore the links among variables related to the scale of operation, the farm area, the degree of output diversification, and, as a measure of technology adoption, expenditures on modern inputs.

Finally, the characteristics of farms with a stable output mix (that is, those farms that did not alter their product mix during the sample period) are compared with those that made substantial changes over time in the composition of products produced. Among the characteristics compared is the use of modern agricultural technologies.

The following findings emerged from the analysis:

- The average degree of output diversity is high and varies substantially across farms, even within farm categories.
- Empirical results fail to show significant statistical links within and across farm categories between output diversification and expenditures on modern inputs, with the notable exception of the coffee category.
- Expenditures on modern inputs were proportional to the scale of operation for all farm categories. That is, expenditures on modern inputs were not linked to the measure of farm-level output diversification: less diversified farms spent approximately as much on modern technologies as their more diversified counterparts.
- Some farm categories have a much more stable product mix than others. Although the more stable and less stable farms differed substantially in their average use of modern inputs, the close and proportional link between the scale of operation and modern input purchases persists across all farms, regardless of the stability of their respective product mixes in the long run.

Several important policy implications are associated with these research results. If technology adoption is not linked to product mix or output diversification, then the returns to adoption of policies aimed at increasing output (such as extension programs) may not be large. The absence of links between product mix and technology adoption suggests that supply constraints that face all producers (more or less equally) in an area have a powerful impact on technological change. These supply constraints may include access to credit and climatic- or soil-based constraints to technological change. Finally, and perhaps most important in light of the low potential of the study area, many farms *did* quite radically alter their product mix during the sample period, 1979–84, suggesting that the potential for farmers to respond to appropriate changes in policy or technology or both is great.

2

INTRODUCTION

Over the past 30 years, the use of modern technology on high-potential agricultural lands has generated large increases in the yields of staple crops (Plucknett 1993). These yield increases have generally been accompanied by increased farm-level and regional specialization, particularly in the relatively few products for which new technologies were available (Hazell and Ramasamy 1991; Lipton and Longhurst 1989). The degree of success on high-potential land (measured in terms of efficiency and equity) has been so great (David and Otsuka 1994) that researchers and policymakers alike have been quick to attempt to replicate Green Revolution experiments in marginal areas, promoting their efforts as schemes for the alleviation of poverty as well as strategies for increasing food production (de Haen 1991).

Unfortunately, the results of applying Green Revolution technologies to marginal areas have often been disappointing. Even where technologies are appropriate, the rates at which adoption has occurred have been slow, yields have been significantly lower and less uniform (both across farms and over time) than in the high-potential areas, farm-level product specialization has been incomplete, and the environmental costs (especially in soil degradation and loss of forest cover) have been high (World Resources Institute 1994; World Bank 1992). Clearly, new strategies (and possibly new technologies) are required if marginal areas are to participate in technology-led, yield-based agricultural growth that does not degrade the environment.[1] Since the absence of farm-level product specialization and slow growth in productivity seem to go hand in hand, a logical first step in generating new development strategies is to explore the links between crop choice and technology adoption decisions in marginal areas composed of small farms with a range of output mixes.

Much of the literature on the adoption of modern agricultural technology and the use of modern inputs takes for granted that supply constraints, credit constraints, farmers' perceptions of the benefits of new technology, and their limited knowledge of modern inputs and their effects jointly or separately determine the adoption and spread of modern agricultural techniques (Reardon and Vosti 1992; Lipton and Longhurst 1989; Feder, Just, and Zilberman 1985). Further, the literature usually assumes (implicitly or explicitly) that the causation runs in one direction from the adoption of new technology to subsequent changes in product mix.

Agroclimatic constraints, the quality of the soil, topography, climate, availability of water, and the like are clearly fundamental in the determination of what crops and livestock products can be, and therefore may be, produced in a given region and on a given farm. But product- and input-specific price policies and infrastructure,

[1]This study follows Vosti and Witcover's (1996) conjecture that only private costs matter in farm-level resource allocation decisions, including those involving technical choice; therefore externalities are excluded from this analysis. On-farm resource degradation is included in the form of reduced yields on degraded or degrading lands.

especially as infrastructure relates to the marketing of specific modern inputs and farm products, greatly affect the extent to which favorable agroclimatic conditions may be translated into specific production decisions (Ayres et al. 1991).

By its very nature, biologically oriented agricultural research generates product-specific technologies. These technologies may interact with product mix and agro-climatic constraints in the adoption of modern technologies. Changes in product mix may clearly result from the decision to adopt new technologies, especially new varieties of crops or livestock as illustrated by the spread of soybean cultivation and of Zebu cattle into areas of Brazil considered too hot and dry to support them.[2] Moreover, a highly diverse agriculture may become increasingly concentrated in a few crop or livestock products especially favored by the existence of more efficient modern technologies, depending on whether infrastructure and agroclimatic conditions permit and price policies provide appropriate incentives.

But the causation may clearly work in reverse. Products already important in a region or on a specific farm (perhaps as the result of agroclimatic constraints) may weigh heavily in farmers' decisions to adopt or not adopt modern technologies and the extent to which these decisions are translated into purchases of inputs produced off-farm.

Thus, on the one hand, adoption decisions may lead to changes in product mix, particularly to concentration on a few products, while, on the other, particular mixes of products may facilitate or slow the adoption of new technologies. Ecological constraints to changes in product mix and the often product-specific nature of technical change will greatly influence these links between cause and effect.

As suggested above, the use of inputs purchased off-farm, such as insecticides, chemical fertilizers, or feed supplements, is one and often the best indicator of the extent to which modern technologies and practices have been adopted. But clearly, because infrastructure affects the availability of inputs and whether various crops and livestock products that make use of purchased inputs are produced, it determines the degree to which decisions to adopt modern practices are translated into the use of purchased modern inputs.

The purpose of this report is to explore the role of farm-level output diversification in the adoption and use of modern agricultural technologies in an agroecologically and economically marginal area. In doing so, questions such as the following are addressed: Does the production of certain types (or combinations) of agricultural products promote or inhibit the use of modern technologies? Does concentration of farm-level output in particular products reduce or increase technology use? Finally, are farmers who radically alter their product mix over time more or less likely to use modern technology intensively? These and a series of related hypotheses are tested empirically using data drawn from the agriculturally diversified but poor Zona da Mata region of Minas Gerais, Brazil.

The method of analysis used is part way between a true structural analysis and simple data analysis. Confluent relationships among the use of modern inputs, various measures of the size of producing units, and the degree of product diversifi-

[2]Personal correspondence with Dr. Eliseu Alves, former president of Empresa Brasileira de Pesquisa Agropecuaria (EMBRAPA).

cation are examined. The relationships are called "confluent" in the older econometric literature because they relate several endogenous variables (Haavelmo 1944).[3] Farms are grouped into product-mix types in order to focus attention on the relation between farm-level output mix and the adoption and use of modern agricultural technology (in the sense of use of purchased inputs). This emphasis on product mix is motivated primarily by the high degree of farm-level output diversification (both within and across farm types) in the area covered by the sample, and the presumption among agricultural economists working in the Zona da Mata that the series of large-scale policy intervention schemes (for example, PRODEMATA)[4] designed for this marginal region have not altered the patterns of general stagnation and lack of responsiveness in agriculture.

It is important to note that off-farm labor provided by farm household members is explicitly included in the list of agricultural outputs. As physical infrastructure improves and demand for labor increases in both the agriculture and nonagriculture sectors, wage labor is likely to be an increasingly important component of household income, both as an alternative income source and as a potential source of cash needed for market purchases of production inputs as well as consumption goods (Reardon, Matlon, and Delgado 1988).

Chapter 3 contains a brief description of the study area and the data used. Chapter 4 highlights the diversified nature of agricultural production in the Zona da Mata and presents indexes of crop and livestock output and total output (including off-farm labor), which is subsequently used as a basis for the initial clustering exercise adopted. Chapter 5 provides an overview of the clustering procedure adopted and the resulting farm types suggested by the data. Chapter 6 contains a series of univariate and bivariate analyses with special attention paid to the interrelationship between the scale and concentration of output and the use of modern inputs in agriculture. Chapter 7 presents some preliminary intertemporal analyses focusing on the characteristics of farms that do not significantly vary their output mix over time and contrasts them with farms where substantial changes in output mix have been made over the six years covered by the sample (1979–84), with special reference to the expenditures on modern farm inputs. Chapter 8 concludes with directions for further research and a discussion of the policy implications of the results with respect to possible policies that might quicken the pace of agricultural modernization in output-diversified, marginal areas like the Zona da Mata. Finally, a technical appendix explains the technique of cluster analysis.

[3]In a system of simultaneous operations explaining several endogenous variables, a statistical association among two or more endogenous variables is called "confluent." The term "confluent relationship" is often contrasted with autonomous or structural relationships.

[4]Programa de Desenvolvimento Integrado da Zona da Mata (PRODEMATA) was a government-sponsored integrated rural development project for the Zona da Mata (semi-arid zone) of the state of Minas Gerais. It was cofinanced by the World Bank and took place during the period 1978–85.

3

DATA

Data for this study were drawn from surveys collected over the period 1979 through 1984 to monitor the progress of PRODEMATA in the impoverished Zona da Mata (see Figure 1). This hilly region contains soils of varying quality with declining fertility; erosion is more of a problem at higher elevations. The climate is tropical, with a rainy season spanning October to April and an annual average temperature of 22°C (peaking at an average of 32° in December). Close to 90 percent of the sample farms contain fewer than 100 hectares, with a substantial proportion of these with fewer than 10 hectares. Sharecropping arrangements are common. In addition to an initial baseline study undertaken in 1975, detailed agricultural production and socioeconomic information was collected using annual retrospective survey questionnaires in July of each year.

Production data on a large set of agricultural products were solicited. Data included inputs and outputs, details on the types and degrees of market linkages, as well as detailed information about the process of transferring knowledge through contact with agricultural extension agents, farmers' organizations, and the like. Special care was taken to note the extent and type of intercropping, although plot-level information was not available. Detailed data were also collected on livestock production and on income from off-farm sources. Land use and land reclamation practices also received special attention, especially farm-level investments in floodplains and reforestation. Finally, the annual retrospectives yielded information to construct yearly socioeconomic profiles of the families and the *municipios* (similar to a county or district) included in the study.

The initial sample for the PRODEMATA study included approximately 800 farmers. The sample was stratified by farm size, skewed toward the smaller, poorer farms targeted by the PRODEMATA project; it included a representative number of sharecroppers. Primarily due to the high percentage of older farmers in the initial sample population of landowners and the decision by the research group monitoring the project to drop observations in the event of land distribution following the death of an owner or sale of the property, the sample size fell to 384 by 1984. Therefore, the final panel used in this analysis contains the 384 farmers who were continuously present throughout the monitoring process.[5]

It should be noted that the sample selected in this region tends to be older than the norm, with the age of household heads averaging 57 years. While this may limit the ability to generalize the research results beyond the Latin American context, trends in rural to urban migration worldwide (and particularly in Latin America) are likely to make the Zona da Mata sample more rather than less typical.

[5]An analysis by Bradley (1990) suggests that no significant selectivity biases resulted from this decrease in sample size.

Figure 1—Zona da Mata, Minas Gerais, Brazil

4

FARM-LEVEL DIVERSIFICATION AND PRODUCT MIX

In order to examine the impact of output mix on technological adoption, it is first necessary to define a system of farming types based on product mix for the region. To do so, an index of total output is formed by combining crop and livestock output and off-farm labor, aggregating different products (including off-farm labor) according to a system of constant relative price weights. Within this framework, 13 different crop and livestock products are distinguished in addition to off-farm labor; then the percentage of each product in total output is calculated. The farms are then grouped by farm types based on these percentages, using statistical cluster analysis. (For technical details about cluster analysis, see the Technical Appendix.) On the basis of clustering results from several specifications, farms are divided into five different farm types, a classification that is found to be most stable over time. Classification into a greater number of clusters results in groupings of farms that change over time (the more clusters distinguished, the greater the instability), and fewer clusters did not adequately characterize production patterns.

Table 1 illustrates the highly diversified nature of agricultural production in the Zona da Mata. Most of the farmers produce three or more crops and two or more livestock products simultaneously. Certain site-specific exceptions to diversification do exist. For example, farms near refineries predominantly produce sugarcane, but they are clearly exceptions. The diversified nature of farming persisted throughout the 1979–84 period covered by the sample.

The PRODEMATA survey data on outputs used in the analysis are summarized in Table 2. Information on 17 categories of livestock production and 16 categories of crop production (including prices) are included. For livestock, the quantity produced is defined as the sum of quantities sold and consumed. For crops, total output is measured in annual gross production. Given the high degree of intertemporal price variation and

Table 1—Product diversification matrix, 1982

Number of Animal Products	Number of Crops						
	0	1	2	3	4	5	6 or more
0	2	2	2	4	3	1	1
1	0	0	1	2	3	0	0
2	2	4	12	16	12	4	1
3	1	9	15	38	51	12	2
4	1	2	13	25	26	16	3
5	1	9	4	12	22	11	7
6 or more	1	1	8	14	39	18	6

Source: Derived from the PRODEMATA data set, University of Viçosa, 1986.

Table 2—Categories of agricultural production in the PRODEMATA survey

Category	Translation	Unit
Livestock	Livestock	
Leite	Milk	Liters/day
Laticinios	Dairy products	Kilograms/month
Reprodutores	Bulls	Head per year
Vacas paridas	Cows with calves	Head per year
Vacas falhadas	Cows without calves	Head per year
Novilhas enxertadas	Bred heifers	Head per year
Garrotas não enxertadas	Unbred heifers	Head per year
Machos desmamodas	Steers/bull calves	Head per year
Bois-de-carro	Oxen	Head per year
Equinos e muares	Horses and mules	Head per year
Suinos	Hogs	Head per year
Caprinos	Goats	Head per year
Ovinos	Sheep	Head per year
Couro	Hides	Hides per year
Esterco	Manure	Tons per year
Aves	Chickens/fowl	Head per year
Ovos	Eggs	Dozens per year
Crops	Crops	
Arroz em casca	Rice	50 kilograms per year
Milho	Corn	60 kilograms per year
Feijão consorciado	Intercrop beans	60 kilograms per year
Feijão solteiro	Single-crop beans	60 kilograms per year
Fumo em corda	Tobacco	15 kilograms per year
Tomate	Tomato	20 kilograms per year
Batata inglesa	Potato	60 kilograms per year
Cebola	Onion	60 kilograms per year
Outras culturas annuais	Other annuals	Varies
Horticultural (outras)	Other horticultural crops	Varies
Café em côco	Coffee	30 kilograms per year
Cana-de-açucar	Sugarcane	Tons per year
Mandioca	Manioc	Tons per year
Citrus	Citrus	Hundreds per year
Banana	Banana	Hundreds per year
Fruticultura (outras)[a]	Other fruits[a]	Varies

Source: Derived from the PRODEMATA data set, University of Viçosa, 1986.
[a]The other horticultural group contained five separate categories each year. The other fruit group contained one category in 1979, and two categories in each subsequent year. In 1979, the other annual category was not reported but was replaced with a peanut category.

insufficient information on inventories, changes in stocks are not considered. For off-farm labor, 6 categories are reported, dividing labor into males, females, and children in agricultural and nonagricultural jobs. All categories are in annual production units, except for the dairy categories, which are multiplied to obtain annual units.

Once the product categories to be included in the analysis are determined, the next step is to choose an appropriate weight for each category (to determine the proportion of gross annual production value contributed by each group). Since the primary purpose of this output index is to differentiate farms by their product mix, it is desirable to eliminate as much variation caused by seasonal swings and inflationary trends as possible from these product weights. For this reason, the prices reported in each category are averaged over farmers and over each agricultural year. To adjust

for changes in relative prices over time, however, annual prices for each output category are normalized (or deflated) by that year's average price of corn. The corn price is chosen as the deflator because it is the most common product each year, and the market for corn is well integrated across the survey area.

Table 3 reports the mean price of corn over the sample period for farms reporting corn prices, whether or not the output was sold. Note that inflation was marked during the 1979–84 period, but that reported corn prices across farms did not deviate widely from the mean. Before computing the mean prices across farms, individual prices are examined for obvious outliers. Outliers are defined informally as a price observation isolated from the next closest observation. A more formal definition has not been adopted since the considerable variation in prices makes any definition based on standardized values impractical.

Table 4 presents the mean price and variation in prices over the period 1979–84, as well as the number of farms reporting prices in each category. Note that in every crop year most of the farmers in the sample produced corn, thereby lending credibility to its use as a deflator. In addition, the large number of farmers reporting nonzero production highlights the diverse nature of agriculture in the Zona da Mata.

The average relative prices for 35 output categories with corn as the deflator and the intertemporal average relative prices used as the value weights in calculating the shares of the value of total output attributable to each product are shown in Table 5. The formula for an individual product weight (w) is:

$$w = 1/6 \sum_{j=1}^{6} \left\{ [(1/n_j) \sum_{i=1}^{nj} p_{ij}] / \bar{p}_j^c \right\}, \tag{1}$$

where i indexes farms, j indexes years (1979–84), p_{ij} is the price of product i in year j, and $\bar{p}_j^c$ is the mean corn price for year j. These weights are multiplied by the quantity produced annually in each product category to obtain annual gross production value in each category. Summing across all categories gives total output value, and dividing each category's value by this total gives the percent of total annual gross output value within each category:

$$\sigma_{ij}^k = 100 \; w^k q_{ij}^k / \sum_{k=1}^{R} w^k q_{ij}^k, \tag{2}$$

where k indexes products and R is their total number, so σ_{ij}^k is the percentage of total production and q_{ij}^k is the physical quantity produced of the kth product by the ith farmer in the jth year.

The number of variables is reduced by aggregating several similar categories into larger product groups to simplify the analysis. Since the farm-level output is a key variable in defining diversity characteristics in the analysis that follows, it is important that this aggregation have a minimal effect on diversity. Initial aggregations combine different products and by-products of what are perceived to be single production activities. Milk, dairy products, bulls, cows, and manure are combined into a "dairy" variable; steers, heifer types, and hides into a "beef" variable; fowl and

Table 3—Corn price reports of sales versus on-farm consumption, 1979–84

Year	Region	Price Report from Farms with Sales		Price Report from Farms without Sales	
		Number	Mean	Number	Mean
			(Cr)		(Cr)
1979	1	27	229.6	79	241.8
			(49.7)		(52.0)
	2	49	226.7	191	233.0
			(28.5)		(38.8)
	3	61	203.8	115	209.7
			(25.7)		(21.7)
	All	137	217.1	385	227.8
			(34.6)		(39.8)
1980	1	27	375.2	66	387.9
			(73.6)		(56.3)
	2	48	409.0	190	411.6
			(69.0)		(53.9)
	3	41	402.0	123	403.3
			(50.3)		(51.5)
	All	116	398.6	379	404.8
			(65.0)		(54.1)
1981	1	41	731.4	35	724.9
			(120.6)		(129.7)
	2	53	757.2	141	823.4
			(135.8)		(537.2)
	3	46	706.4	86	728.1
			(120.9)		(420.5)
	All	140	733.0	262	779.0
			(127.5)		(465.7)
1982	1	32	1,062.5	43	1,065.1
			(156.1)		(174.4)
	2	46	1,128.3	145	1,107.6
			(200.7)		(148.2)
	3	80	1,034.2	49	994.9
			(149.0)		(70.9)
	All	158	1,067.3	237	1,076.6
			(170.8)		(147.8)
1983	1	28	2,796.4	47	2,929.8
			(656.9)		(605.0)
	2	20	3,175.0	146	3,379.5
			(437.5)		(679.6)
	3	37	4,076.2	108	3,875.9
			(669.5)		(471.8)
	All	85	3,442.6	301	3,487.4
			(840.9)		(683.7)
1984	1	22	10,759.1	48	11,500.0
			(1,609.7)		(1725.9)
	2	26	10,192.3	121	10,322.3
			(1,414.8)		(1,501.1)
	3	43	10,581.4	91	10,549.5
			(981.6)		(1,171.5)
	All	92	10,507.6	250	10,619.2
			(1,281.2)		(1,499.4)

Source: Derived from the PRODEMATA data set, University of Viçosa, 1986.
Note: Numbers in parentheses are standard deviations.

Table 4—Mean and standard deviation of prices across all farms reporting a price, 1979–84

Product Category	N	1979 Mean	Standard Deviation	N	1980 Mean	Standard Deviation	N	1981 Mean	Standard Deviation
Livestock									
Milk	278	5.0	2.7	277	10.4	4.8	244	25.8	6.3
Dairy products	40	40.4	11.4	49	80.1	15.6	53	152.5	39.5
Bulls	26	9,084.6	5,657.4	11	22,181.8	14,351.4	15	31,333.3	11,616.9
Cows with calves in the previous year	58	7,294.8	2,334.6	46	17,108.7	10,617.0	20	29,016.7	16,858.7
Cows without calves	83	6,877.1	5,982.2	62	14,933.9	7,367.8	54	19,257.4	8,987.6
Bred heifers	15	5,900.0	1,429.3	20	18,500.0	15,243.6	10	18,333.3	7,433.6
Unbred heifers	41	41,091.2	1,716.2	42	9,414.3	5,265.6	25	11,444.0	5,739.4
Steers	105	3,626.4	1,875.9	77	9,329.9	8,190.0	61	12,565.6	7,881.7
Oxen	23	9,163.0	6,850.5	10	20,600.0	8,896.9	12	28,979.2	8,788.2
Horses and mules	12	3,100.0	1,916.9	16	4,939.6	2,280.0	12	17,333.3	6,692.4
Hogs	352	531.6	417.1	397	916.2	188.4	342	1,337.0	225.9
Goats	4	550.0	369.7	7	1,280.0	1,213.3	6	1,435.0	604.1
Hides	...	...	...	1	50.0	...	2	500.0	424.3
Manure	15	274.0	168.6	57	398.7	150.1	171	850.0	524.2
Fowl	393	52.2	53.3	475	81.4	24.0	394	165.6	50.6
Eggs	...	...	...	424	35.1	6.7	394	63.1	16.4
Crops									
Rice	366	369.4	99.2	357	619.9	165.8	312	936.9	237.4
Corn	522	225.0	38.8	495	403.3	56.8	402	762.9	383.8
Intercrop beans	327	681.4	120.3	316	2,045.3	523.5	307	4,203.0	868.0
Single-crop beans	182	677.1	123.2	174	2,310.7	425.6	156	4,482.4	716.2
Tobacco	20	603.8	116.8	12	1,608.3	242.9	9	3,861.1	1,256.9
Tomatoes	8	127.5	34.9	6	273.3	129.9	3	466.7	208.2
Potatoes	4	227.5	93.2	3	900.0	346.4	3	1,400.0	529.2
Onions	0	0	0	0	0	0	0	0	0
Coffee	257	624.3	65.5	232	1,034.5	101.2	220	1,327.0	300.5
Sugarcane	74	246.8	81.9	83	540.8	291.4	80	934.5	544.0
Manioc	3	333.3	1,527.5	10	5,920.0	2,114.4	3	11,663.3	2,889.6
Citrus	12	50.0	48.4	8	81.3	37.2	14	91.4	26.3
Bananas	6	22.8	21.8	9	125.6	15.1	15	143.2	157.0
Wage labor									
Agriculture									
Male	135	62.7	29.6	126	123.8	110.4	108	193.5	63.6
Female	19	39.2	12.6	18	58.6	23.25	14	199.3	161.1
Child	9	35.8	15.7	6	79.2	54.8	8	73.8	32.5
Nonagriculture									
Male	51	191.4	184.1	82	444.8	710.0	47	811.0	981.1
Female	24	99.2	180.7	34	218.5	209.4	36	331.7	299.9
Child	0	0	0	0	0	0	0	0	0

Table 4—Continued

Product Category	N	1982 Mean	Standard Deviation	N	1983 Mean	Standard Deviation	N	1984 Mean	Standard Deviation
Livestock									
Milk	229	39.7	7.1	226	106.5	16.8	209	286.8	57.1
Dairy products	47	274.5	59.4	50	678.2	143.5	50	1,906.0	390.9
Bulls	16	31,875.0	10,249.4	22	82,772.7	34,206.6	18	412,777.8	169,099.9
Cows with calves in the previous year	17	34,411.8	13,805.8	38	82,184.2	33,602.5	31	276,000.0	123,470.6
Cows without calves	55	29,645.3	16,249.7	68	71,171.6	44,016.7	52	239,173.1	103,674.6
Bred heifers	11	25,272.7	7,617.0	5	68,000.0	19,235.4	10	324,400.0	164,689.1
Unbred heifers	34	19,441.2	15,639.8	35	41,675.5	19,235.4	34	15,600.0	92,258.2
Steers	66	18,019.7	13,237.9	75	35,506.7	21,787.3	61	144,245.9	104,238.3
Oxen	12	40,666.7	16,897.2	19	103,842.1	56,067.7	15	374,900.0	152,369.9
Horses and mules	12	24,833.3	14,989.9	19	71,789.5	44,187.2	11	149,909.1	85,616.0
Hogs	337	3,430.9	550.3	313	6,719.5	706.6	309	28,918.1	3,172.8
Goats	4	2,500.0	707.1	9	8,055.6	5,198.8	12	31,458.3	16,461.3
Hides	0	0	0	5	1,040.0	1,152.4	0	0	0
Manure	131	1,981.7	1,065.5	158	5,416.9	4,154.3	2	13,275.3	8,144.8
Fowl	402	343.9	87.9	383	786.1	203.9	348	2,459.9	718.8
Eggs	398	145.3	28.1	384	302.4	45.3	346	1,248.1	239.6
Crops									
Rice	301	2,601.0	523.1	305	5,492.8	794.1	285	15,612.3	3,400.9
Corn	395	1,072.9	157.2	386	3,477.5	720.4	352	10,590.1	1,444.6
Intercrop beans	255	4,872.5	820.1	272	15,152.6	3,108.1	283	44,941.7	7,954.5
Single-crop beans	130	4,902.3	953.0	141	16,152.5	3,224.3	108	43,583.3	6,190.3
Tobacco	19	9,968.4	5,348.1	15	9,666.7	6,463.3	9	40,500.0	9,307.3
Tomatoes	4	1,200.0	355.9	4	1,750.0	500.0	9	5,500.0	2,397.9
Potatoes	6	2,483.3	895.4	3	7,666.7	2,309.4	3	23,000.0	2,645.8
Onions	0	0	0	1	18,000.0	...	1	200.0	...
Coffee	214	3,119.7	252.4	224	6,242.6	573.4	216	26,663.0	3,400.9
Sugarcane	74	1,890.1	604.5	84	3,661.6	1,169.1	89	9,540.5	4,246.8
Manioc	11	15,236.4	8,814.5	9	53,333.3	28,722.8	13	17,615.4	61,228.5
Citrus	6	358.3	241.7	5	460.0	240.8	7	2,071.4	1,367.1
Bananas	18	251.4	327.7	20	279.4	236.8	22	1,300.7	1,083.3
Wage labor									
Agriculture									
Male	111	446.4	398.5	104	966.1	659.2	93	2,393.5	1,398.1
Female	17	259.4	111.3	14	525.0	196.9	13	1,707.7	1,133.2
Child	3	116.7	28.9	4	342.5	160.9	4	1,675.0	689.8
Nonagriculture									
Male	64	1,165.7	1072.9	53	2,383.5	1,969.5	45	8,069.0	9,007.4
Female	43	962.6	1038.5	40	1,327.0	1,360.5	48	3,560.9	4,336.6
Child	1	40.0	...	1	400.0	...	2	1,300.0	424.3

Source: Derived from the PRODEMATA data set, University of Viçosa, 1986.
Note: A total of 384 farms were included in the sample.

Table 5—Average product prices relative to the corn price, major products, 1979–84

Product Category	1979	1980	1981	1982	1983	1984	Mean
Livestock							
Milk	0.022	0.026	0.034	0.037	0.031	0.027	0.029
Dairy products	0.179	0.199	0.200	0.256	0.195	0.180	0.201
Bulls	40.370	54.990	41.070	29.710	23.800	38.980	38.160
Cows with calves	32.420	42.420	38.030	32.070	23.630	26.060	32.440
Cows without calves	30.560	37.030	25.240	27.630	20.470	22.580	27.250
Bred heifers	26.220	45.870	24.030	23.560	19.550	30.630	28.310
Unbred heifers	18.230	23.340	15.000	18.120	11.980	14.730	16.900
Steers	16.120	23.130	16.470	16.800	10.210	13.620	16.060
Oxen	40.720	51.070	37.980	37.900	29.860	35.400	38.820
Horses/mules	13.780	12.250	22.720	23.150	20.640	14.160	17.780
Pigs	2.363	2.272	1.752	3.198	1.932	2.731	2.375
Goats	2.444	3.173	1.881	2.330	2.316	2.971	2.519
Hides	...	0.372	0.655	...	0.299	...	0.442
Manure	1.218	0.988	1.114	1.847	1.558	1.254	1.330
Fowl	0.232	0.202	0.217	0.321	0.226	0.232	0.238
Eggs	...	0.087	0.083	0.135	0.087	0.118	0.102
Crops							
Rice	1.642	1.537	1.228	2.424	1.580	1.474	1.647
Corn	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Intercrop beans	3.028	5.071	5.509	4.542	4.357	4.244	4.458
Single crop beans	3.009	5.729	5.875	4.569	4.645	4.116	4.657
Tobacco	2.683	3.987	5.061	9.291	2.780	3.824	4.604
Tomatoes	0.567	0.678	0.612	1.118	0.503	0.519	0.666
Potatoes	1.011	2.231	1.835	2.315	2.205	2.172	1.961
Onions	...	...	...	...	5.176	0.019	2.597
Coffee	2.775	2.565	1.739	2.908	1.795	2.518	2.383
Sugarcane	1.097	1.341	1.225	1.762	1.053	0.901	1.230
Manioc	14.810	14.680	15.290	14.200	15.340	16.770	15.180
Citrus	0.222	0.201	0.120	0.334	0.132	0.196	0.201
Bananas	0.102	0.311	0.188	0.234	0.080	0.123	0.173
Wage labor							
Agriculture							
Male	0.279	0.307	0.254	0.416	0.278	0.226	0.293
Female	0.174	0.145	0.261	0.241	0.151	0.161	0.189
Child	0.159	0.196	0.097	0.109	0.098	0.158	0.136
Nonagriculture							
Male	0.851	1.103	1.063	1.087	0.685	0.762	0.925
Female	0.441	0.542	0.435	0.897	0.382	0.336	0.505
Child	...	...	...	0.037	0.115	0.123	0.092

Source: Derived from the PRODEMATA data set, University of Viçosa, 1986.

eggs into a "fowl" variable; and single-crop and intercropped beans into a "beans" variable. Also, for off-farm labor, all six categories are aggregated. A second type of aggregation is made up of products that are similar or generally do not represent significant percentages of total production. Horses and mules, oxen, and goats are combined into an "other animal" variable; tomatoes, potatoes, onions, other annuals, other horticultural crops, citrus, bananas, and other fruits are combined into a "perishables" variable. Prior examination of these combinations indicates that relatively few farms produced more than one type within the same group, and therefore these aggregations do not greatly affect the measures of diversity adopted. The resulting

output data for each farm include the total output index for each year and the percentage of the total index in the five livestock and eight crop production groups: dairy, beef, pork, fowl, and other animals; rice, corn, beans, tobacco, coffee, sugar, manioc, and perishables; and off-farm labor.

A similar value-weighted index of modern inputs used is also constructed in exactly the same manner as the output index. Price outliers are deleted before the average prices in each year are calculated. These prices are then deflated by the mean corn price in that year. Finally, the resulting averages are again averaged over the six-year period to obtain the weight for each category of modern input. The weights and quantities of inputs used are available for chemical fertilizer, mechanical tillage, modern feed supplements, and most veterinary expenses. For pesticides, hybrid seeds, and some veterinary expenses only expenditure figures are available; separate prices and quantities are not provided. These expenditures are deflated directly by the average corn price in each year and aggregated into the total to obtain the overall index. The results for 1984, presented in Table 6, as an example, indicate that expenditures on all types of modern inputs varied greatly at the farm level.

Table 6—Index of modern input use, by category, 1984

Input	Mean	Standard Deviation	Minimum	Maximum
Hybrid seeds	3.76	6.49	0	51.94
Chemical fertilizer	47.44	99.14	0	818.99
Mechanical tillage	8.93	27.81	0	251.96
Pesticides	2.36	7.38	0	21.97
Total crop	62.48	125.17	0	1,070.07
Modern feeds	20.22	212.14	0	4,082.57
Veterinary services	2.58	3.06		22.95
Total livestock[a]	25.97	214.84	0	4,119.11
Total crop and livestock	88.45	247.58	0	4,154.49

Source: Derived from the PRODEMATA data set, University of Viçosa, 1986.
[a]The total livestock index includes a small amount of modern inputs for forage crops.

5

CLUSTER ANALYSIS

In the statistical cluster analysis, farms are assigned to groups within which similar shares of farm output are devoted to various products. Since the idea is to identify farms by their product mixture, other variables such as gross product value or farm area are not included in the cluster analysis. However, statistics that describe farmer and farm characteristics are provided for the 1984 clusters. Similarity is based solely on the percentage contribution to total output of each of the 14 product categories: beef, dairy, pork, fowl, other livestock, corn, beans, rice, coffee, sugar, tobacco, manioc, perishables, and off-farm labor. Only farms that took part in the survey in all six years, 1979–84, are used in the cluster analysis. The analysis is repeated for each year, rather than only once based on a six-year average, to allow for annual variations in the farmers' choices of product mix. As a consistency check, the results of cluster analysis on the panel subset are compared with a clustering using all farms surveyed in each year. Only a few farms are grouped differently when using the entire data set.

A survey of clustering methods and general difficulties associated with their use is included as the Technical Appendix. To summarize, there are essentially three decisions to be made when choosing a clustering method. The first is to determine how similarities between observations are to be measured. The second is to choose a specific clustering algorithm, and the third is to determine the precise number of clusters in the data set. As previously stated, differences in the percentage of total production of 14 categories of production is the measure of interfarm similarity. The similarity between observations is defined mathematically as the standard Euclidean distance between the observations viewed as vectors in the R_{14} product-percentage space.

In clustering algorithms, both hierarchical (including single, average, and Ward linkages) and nonhierarchical techniques are available. A nonhierarchical method is used here because it generally results in clusters with lower pooled within-cluster sums of squared deviations than any hierarchical method, and because it does not impose a strict and artificial a priori hierarchy on the data. The algorithm used is similar to the *K*-means, or nearest centroid method of sorting algorithms described in the Appendix. Specifically, it is an iterative technique that requires as initial input the number of clusters and an initial centroid, or mean vector, for each cluster. In the first iteration, the observations are considered one at a time and assigned to the cluster with the nearest centroid. The centroids remain unchanged until all observations are assigned. Before the second iteration, new centroids are determined from the mean vector of all observations allocated to specific clusters during the first iteration. The observations are then reassigned using these new centroids. Iterations continue in this manner until the cluster centroids converge, that is, until the centroids change less than a prespecified tolerance limit between iterations. It has been shown that convergence will always occur and will result in an approximate local minimum of the

pooled within-cluster sum of squared deviation called trace (W) or trW,[6] known as a one-move local optimum.[7] By the standard decomposition of variance, the total deviations matrix is the sum of W and the between-cluster deviations matrix,

$$B = \sum_{s=1}^{k} n_s \bar{\sigma}_s \times \bar{\sigma}_s, \tag{3}$$

where $\bar{\sigma}_s$ is the 14 × 1 vector of average product percentages for clusters. So, minimizing the pooled within-cluster sum of squared deviations, $tr(W)$, is the same as maximizing the between-cluster sum of squared deviations, $tr(B)$. This objective function results in the maximum-likelihood allocation under the assumption that the observations within clusters are distributed spherical multivariate normal, that is, $f_s = (\mu_s, \sigma^2 I)$ is the distribution of clusters (Hawkins, Muller, and ten Krooden 1982, 326). This assumption states that there is no correlation between variables within a cluster, that variances are equal for each variable, and that the covariance matrix is identical for all clusters. This is obviously not realistic, but satisfactory procedures using more general assumptions have yet to be developed.

Given the $tr(W)$ objective, the centroid sorting method will result in an approximate local optimum, regardless of initial centroids. However, as with many other iterative optimizing techniques, a local minimum frequently may not coincide with the global minimum. So, for any prespecified number of clusters, some method of determining a global optimum is necessary. Since practical dynamic programming algorithms converging to the global optimum are unavailable, a search procedure over alternative initial centroids is used. For the production data used in this study, preliminary results indicate that the groupings obtained from alternative initial centroids generally differ only in the reallocation of a few borderline observations, with little change in the overall cluster character.[8] Therefore, it was deemed unnecessary to conduct a large formal grid search over possible initial centroids. Rather, the centroid sorting algorithm is repeated several times, using two different types of initial centroids. Given K clusters, the first type of initial centroid is obtained by first performing a Ward hierarchical clustering of the data set and then using the mean vectors of the Kth-level clusters as the K initial centroids. The second type of initial centroid is a random drawing of K farms from the data set. To check for local minima, the clustering is repeated nine times for the specified number of clusters, once with the first type of initial centroids and eight times with random drawings as initial centroids. The final clustering is then chosen as the one giving the minimum $tr(W)$ out of these nine.

In determining the number of clusters in the data set, none of the standard statistics are satisfactory (see the Appendix). This is particularly true for cases in

[6] See the Technical Appendix for the definition of trace (W) or trW.

[7] It is possible that the centroid sorting method will not result in an exact minimum $tr(W)$. However, the difference between the exact minimum $tr(W)$ and the one resulting from centroid sorting will be relatively small. See the Technical Appendix for a more precise discussion of this result.

[8] Throughout this discussion, cluster character refers to the number of observations, the mean vector, and the covariance matrix of the cluster.

which a relatively small number of clusters capture the majority of variation in the underlying data. This study falls into this category since relatively few clusters of production activities effectively characterize most of the sample farms. The usual "number of clusters" criteria are especially inadequate for detecting a relatively small number, say, less than eight. For this reason, it is necessary to rely more on intuitive and informal procedures. The primary informal criterion is comparison of clusters over time. To be useful in a panel study, the clusters obtained each year should be identifiable in other years.

The results for experiments with the number of clusters equaling four, five, six, seven, and eight, using 1979 production data, are given in Table 7. When five clusters are specified, the major farm types are corn, coffee, dairy, rice, and off-farm labor. The number of farms in each cluster are in parentheses and the percentages below each cluster heading indicate the mean percent of total output attributed to each product category. For example, the corn produced on the 109 farms that made up the fifth-level corn cluster in 1979 averaged 31 percent of the value of total output, whereas pork contributed an average of 15 percent and beans 13 percent. It must be emphasized that the names given these clusters are for mnemonic purposes only. The "meaning" of a cluster is expressed in terms of the centroid percentages, which represent average percentage contributions to total output of particular products (or groups of products) to total farm output.

As the number of clusters is increased, larger and more heterogeneous clusters begin to subdivide. For example, as analysis proceeds from five to six clusters, the dairy cluster splits into diversified and specialized dairy subclusters. As the number of clusters increases further to seven the corn cluster splits as well. Finally, with eight clusters, a small group of farmers specializing in commercial crops (sugar or tobacco) splits from the diversified corn subcluster.

These results indicate a rough hierarchy in the sense that the dairy group at the fifth level is represented by two groups at the eighth level, and the corn group at the fifth level is represented by three groups at the eighth. Since the procedure is nonhierarchical, there are numerous farms reallocated between the fourth and eighth levels. For instance, 2 farms from the fifth-level coffee cluster, 10 farms from the corn cluster, 3 farms from the rice cluster, and 3 from the off-farm labor cluster are reallocated to the diversified dairy cluster at the eighth level. Also reallocated between the fifth and eighth levels are 3 farms from the off-farm labor cluster to corn subclusters and 2 farms from the dairy cluster to the diversified corn cluster. In total 23 farms are reallocated nonhierarchically, that is, to clusters with average vectors having a different major product.

Tables 8 to 12 give the results of the cluster analysis using the PRODEMATA panel data for each year over the 1980–84 period. The results for 1979 are similar, but there are significant differences among years. For example, in all years except 1979, only one dairy cluster is present at the eighth level. At the fourth and fifth levels, however, the general character of the clusters is relatively stable. In every year, at the fifth level, there is a cluster with dairy as the main product, one with coffee, one with rice, one with corn, and one with off-farm labor. At all levels above the fifth, this stability over time does not occur. For this reason, fifth-level clustering is chosen for the panel studies of the links between farm-level output, diversity, and agricultural modernization.

Table 7—Cluster results, including off-farm labor, 1979

4 Clusters	5 Clusters	6 Clusters	7 Clusters	8 Clusters
Corn (131) 28% corn 15% rice 13% pork 12% beans	Corn (109) 31% corn 15% pork 13% beans	Corn (99) 33% corn 15% pork 13% beans	Corn, spc (45) 49% corn 19% beans	Corn, spc (54) 46% corn 18% beans
...	...	...	Corn, div (58) 22% pork 19% corn	Corn, div (34) 32% pork 16% corn 14% fowl
...	Rice (27) 58% rice 12% dairy 10% corn	Rice (24) 60% rice 10% corn	Rice (24) 60% rice 10% corn	Rice (24) 60% rice 10% corn
Ofl (55) 52% ofl 15% corn	Ofl (56) 51% ofl 15% corn	Ofl (53) 52% ofl 15% corn	Ofl (50) 54% ofl 15% corn	Ofl (50) 54% ofl 15% corn
Coffee (87) 63% coffee	Coffee (89) 62% coffee	Coffee (87) 63% coffee	Coffee (88) 62% coffee	Coffee (87) 63% coffee
Dairy (109) 54% dairy	Dairy (101) 56% dairy 10% corn	Dairy, spc (42) 73% dairy	Dairy, spc (42) 73% dairy	Dairy, spc (42) 73% dairy
...	...	Dairy, div (77) 40% dairy 12% corn 10% rice 10% coffee	Dairy, div (75) 40% dairy 13% corn 10% rice	Dairy, div (75) 40% dairy 13% corn 10% rice 10% coffee
...	...	...	...	Sugar/ tobacco (16) 29% tobacco 23% sugar 19% corn

Notes: Spc is specialized; div is diversified; and ofl is off-farm labor.
Format: Cluster label (number of farms)
main components of mean vector

Table 8—Cluster results, including off-farm labor, 1980

4 Clusters	5 Clusters	6 Clusters	7 Clusters	8 Clusters
Corn (137) 26% corn 17% rice 14% pork 10% fowl	Corn (113) 28% corn 16% pork 12% fowl 11% beans	Corn (105) 29% corn 17% pork 12% fowl 10% beans	Corn (101) 31% corn 17% pork 12% beans	Corn, spc (47) 45% corn 13% beans 10% pork
...	...	...	...	Corn, div (55) 23% pork 18% corn 10% fowl
...	Rice (29) 54% rice 13% corn	Rice (29) 54% rice 13% corn	Rice (29) 54% rice 13% corn	Rice (28) 55% rice 13% corn
Ofl (63) 52% ofl 14% corn	Ofl (61) 52% ofl 14% corn	Ofl (62) 52% ofl 14% corn	Ofl (61) 52% ofl 14% corn	Ofl (58) 53% ofl 13% corn
Coffee (89) 56% coffee 10% dairy 10% corn	Coffee (87) 57% coffee 10% dairy	Coffee, spc (50) 69% coffee	Coffee, spc (50) 69% coffee	Coffee, spc (50) 69% coffee
...	...	Coffee, div (59) 34% coffee 21% dairy 12% corn	Coffee, div (56) 34% coffee 22% dairy 12% corn	Coffee, div (60) 33% coffee 21% dairy 12% corn
Dairy (94) 52% dairy 10% corn 10% rice	Dairy (93) 53% dairy 10% corn	Dairy (78) 56% dairy 10% rice 10% corn	Dairy (79) 56% dairy 10% corn 10% rice	Dairy (78) 56% dairy 10% corn 10% rice
...	...	...	Fowl (7) 67% fowl	Fowl (7) 67% fowl

Notes: Spc is specialized; div is diversified; and ofl is off-farm labor.
Format: Cluster label (number of farms)
main components of mean vector

Table 9—Cluster results, including off-farm labor, 1981

4 Clusters	5 Clusters	6 Clusters	7 Clusters	8 Clusters
Corn (131) 25% corn 15% rice 13% beans 10% pork 10% fowl	Corn (101) 29% corn 15% beans 12% pork 11% fowl	Corn (102) 28% corn 15% beans 11% pork 11% fowl	Corn (87) 31% corn 16% beans 12% pork	Corn, spc (58) 36% corn 21% beans 13% pork
...	...	...	...	Corn, div (57) 20% dairy 16% corn
...	Rice (40) 40% rice 11% corn 10% dairy	Rice (26) 51% rice 13% corn	Rice (33) 46% rice 12% corn 11% dairy	Rice (25) 52% rice 13% corn
Ofl (52) 54% ofl 12% corn	Ofl (53) 54% ofl 12% corn	Ofl (50) 55% ofl 12% corn	Ofl (51) 54% ofl 13% corn	Ofl (44) 58% ofl 12% corn
Coffee (115) 65% coffee	Coffee (114) 66% coffee	Coffee, spc (79) 74% coffee	Coffee, spc (62) 78% coffee	Coffee, spc (76) 75% coffee
...	...	Coffee, div (52) 41% coffee 14% corn 13% dairy	Coffee, div (65) 47% coffee 13% corn 11% dairy	Coffee, div (53) 43% coffee 14% corn 12% dairy
Dairy (86) 54% dairy	Dairy (76) 57% dairy	Dairy (75) 57% dairy	Dairy (79) 56% dairy	Dairy (65) 60% dairy
...	...	...	Fowl (7) 67% fowl	Fowl (6) 71% fowl

Notes: Spc is specialized; div is diversified; and ofl is off-farm labor.
Format: Cluster label (number of farms)
main components of mean vector

Table 10—Cluster results, including off-farm labor, 1982

4 Clusters	5 Clusters	6 Clusters	7 Clusters	8 Clusters
Corn (150) 25% corn 15% rice 12% pork 10% fowl	Corn (131) 26% corn 12% pork 11% beans 11% fowl	Corn (107) 28% corn 12% pork 11% fowl 11% beans	Corn (86) 29% corn 14% pork 12% fowl 11% beans	Corn (65) 33% corn 14% pork 13% beans
...	...	...	Corn/Ofl (56) 32% ofl 18% corn 10% pork	Corn/Ofl (52) 33% ofl 18% corn 11% pork
...	Rice (25) 51% rice 12% corn	Rice (22) 53% rice 11% corn 10% dairy	Rice (21) 54% rice 11% corn 11% dairy	Rice (31) 45% rice 13% corn
Ofl (57) 56% ofl 10% corn	Ofl (54) 57% ofl 10% corn	Ofl (52) 58% ofl 10% corn	Ofl, spc (28) 73% ofl	Ofl, spc (31) 73% ofl
Coffee (96) 61% coffee	Coffee (96) 61% coffee	Coffee, spc (63) 71% coffee	Coffee, spc (61) 71% coffee	Coffee, spc (63) 71% coffee
...	...	Coffee, div (69) 34% coffee 16% corn 11% dairy 11% pork	Coffee, div (62) 36% coffee 15% corn 13% dairy 10% pork	Coffee, div (63) 35% coffee 16% corn 12% dairy 11% pork
Dairy (81) 53% dairy	Dairy (78) 53% dairy	Dairy (71) 56% dairy	Dairy (70) 56% dairy	Dairy (70) 56% dairy
...	...	...	...	Fowl (12) 47% fowl 18% corn 12% pork

Notes: Spc is specialized; div is diversified; and ofl is off-farm labor.
Format: Cluster label (number of farms)
main components of mean vector

Table 11—Cluster results, including off-farm labor, 1983

4 Clusters	5 Clusters	6 Clusters	7 Clusters	8 Clusters
Corn (131) 22% corn 15% rice 11% pork 10% fowl	Corn (99) 26% corn 13% pork 12% fowl 10% beans	Corn (101) 26% corn 13% pork 12% fowl 10% beans	Corn (83) 28% corn 13% fowl 13% pork 10% beans	Corn (66) 30% corn 14% fowl 13% pork 11% beans
...	...	...	...	Corn/Ofl (40) 37% ofl 15% corn 10% pork
...	Rice (32) 39% rice 17% perishables	Rice (22) 52% rice 12% ofl	Rice (22) 52% rice 12% ofl	Rice (22) 52% rice 12% corn
Ofl (51) 60% ofl	Ofl (52) 59% ofl 10% corn	Ofl (51) 59% ofl 10% corn	Ofl (50) 60% ofl 10% corn	Ofl, spc (27) 73% ofl
Coffee (123) 67% coffee	Coffee (122) 68% coffee	Coffee (122) 68% coffee	Coffee, spc (84) 76% coffee	Coffee, spc (84) 76% coffee
...	...	...	Coffee, div (62) 42% coffee 12% corn 11% dairy	Coffee, div (62) 42% coffee 12% corn 11% dairy
Dairy (79) 54% dairy	Dairy (79) 54% dairy	Dairy (79) 54% dairy	Dairy (74) 55% dairy	Dairy (74) 55% dairy
...	...	Perishables (9) 63% perishables	Perishables (9) 63% perishables	Perishables (9) 63% perishables

Notes: Spc is specialized; div is diversified; and Ofl is off-farm labor.
Format: Cluster label (number of farms)
main components of mean vector

Table 12—Cluster results, including off-farm labor, 1984

4 Clusters	5 Clusters	6 Clusters	7 Clusters	8 Clusters
Corn (131) 25% corn 13% rice 11% beans 11% pork	Corn (115) 27% corn 12% beans 11% ppork 10% dairy	Corn (112) 27% corn 12% beans 12% pork 10% fowl	Corn (92) 30% corn 13% beans 12% pork 12% pork	Corn (58) 34% corn 14% beans 12% fowl
...	...	...	...	Corn/dairy (48) 31% dairy 17% corn 10% pork
...	Rice (23) 50% rice 11% dairy	Rice (23) 50% rice 11% dairy	Rice (23) 50% rice 11% dairy	Rice (25) 49% rice 14% dairy
Ofl (51) 59% ofl	Ofl (48) 61% ofl	Ofl (48) 61% ofl	Ofl (47) 61% ofl	Ofl (45) 62% ofl
Coffee (124) 67% coffee	Coffee (124) 67% coffee	Coffee (121) 68% coffee	Coffee, spc (86) 76% coffee	Coffee, spc (95) 74% coffee
...	...	...	Coffee, div (57) 42% coffee 12% corn 11% dairy	Coffee, div (54) 37% coffee 16% corn 10% dairy
Dairy (77) 56% dairy 10% rice	Dairy (73) 57% dairy	Dairy (70) 58% dairy	Dairy (71) 57% dairy	Dairy, spc (51) 64% dairy
...	...	Perishables (9) 59% perishables	Perishables (7) 68% perishables	Perishables (7) 68% perishables

Notes: Spc is specialized; div is diversified; and ofl is off-farm labor.
Format: Cluster label (number of farms)
main components of mean vector

Table 13—Descriptive statistics within clusters, 1984

Farm Characteristic	Coffee N = 122	Corn N = 114	Dairy N = 71	Off-farm Labor N = 48	Rice N = 22
Mean age of household head (years)[a]	56.0	57.8	56.8	56.4	60.8
	(11.2)	(11.6)	(11.8)	(10.2)	(9.2)
Education of household head (years)	2.1	1.9	2.7	1.4	1.5
	(1.8)	(2.2)	(2.0)	(1.7)	(1.5)
Family size (number of persons)	6.1	5.2	4.9	6.0	4.8
	(2.9)	(2.5)	(2.3)	(2.5)	(2.7)
Dependency ratio (0–15+755/16–54)	0.21	0.19	0.17	0.17	0.15
	(0.21)	(0.22)	(0.19)	(0.20)	(0.18)
Percent electrified	48	28	59	25	23
	(50)	(45)	(49)	(43)	(42)
Farm size (hectares)	35.9	25.6	59.8	16.5	22.4
	(36.5)	(33.9)	(46.2)	(24.3)	(22.6)
Value of total output (Cr)[b]	1,209	478	980	449	487
	(1,160)	(625)	(739)	(296)	(462)
Value of on-farm production (Cr)[b]	1,191	459	951	176	454
	(1,159)	(623)	(726)	(157)	(449)
Value of off-farm labor (Cr)[b]	18.2	19.1	29.7	273.0	33.3
	(54.8)	(33.5)	(55.1)	(215.1)	(52.2)
Yield (output/hectare)[b]	51.9	27.7	20.2	29.4	32.9
	(45.3)	(27.1)	(14.5)	(28.4)	(29.2)
Percent sharecroppers	12	11	0	29	23
	(33)	(32)	(0)	(45)	(42)
Distance to market (kilometers)	26.8	19.5	21.8	22.5	18.4
	(12.9)	(13.4)	(13.1)	(15.2)	(10.3)
Percent who owned transport vehicles	75	61	85	38	64
	(43)	(49)	(36)	(48)	(48)
Percent co-op members	15	12	55	2	23
	(35)	(33)	(50)	(14)	(42)
Percent of output in top two products	81.6	59.9	75.6	81.3	76.4
	(10.1)	(13.4)	(12.3)	(11.8)	(13.5)
Total modern inputs used (Cr)[b]	147.7	38.6	64.6	26.8	27.7
	(190.5)	(62.5)	(101.0)	(58.9)	(40.5)
Total modern input use/ total output (crops and livestock)b	0.117	0.090	0.069	0.154	0.077
	(0.103)	(0.072)	(0.070)	(0.305)	(0.121)

Note: The numbers in parentheses are standard deviations.

[a]The likelihood ratio test for equality of variance among the five clusters rejected equality at the 5 percent significance level for all variables except age, education, family size, dependency ratio, percent electrified, distance to market, and percent owning transport. The F-test for equality of means, adjusted for heteroskedasticity where appropriate, rejected equality of means at the 5 percent level for all variables except age and dependency ratio.

[b]All value figures for outputs and inputs are expressed in terms of the corn-based indexes derived in Chapter 4.

The descriptive statistics provided in Table 13 for a series of farmer- and farm-specific characteristics for five clusters in 1984 highlight the similarities and differences both across and within clusters.[9] The demographic composition of households is quite similar. Household heads are generally in their mid- to late 50s and have approximately two years of formal education. Family size averages slightly over five persons, and dependency ratios are low and similar, which is not surprising given the

[9] In this and all succeeding analyses and tables, seven farms have been deleted from the sample of 384 farms. Two farms were deleted because they reported zero output in at least one year. Five farms reported extremely high livestock input and output in several years. Inclusion of these farms in one of the five clusters values the results; therefore, they were deleted.

age structure of household heads. The geographic distribution of households (measured in terms of distances to markets) is also uniform across clusters.

When the agricultural characteristics of farms are analyzed, the similarities disappear. Average farm size varies from 16.5 hectares for farms specializing in off-farm labor to 59.8 hectares for dairy farmers. The value of total output (expressed in terms of a corn-based index) for coffee producers is 1,209, nearly three times that of farmers in the corn and off-farm labor clusters. Predictably, farms specializing in off-farm labor derive nearly 10 times more income from the sale of household labor than do other farms.

While distance to markets is similar for most farms in the sample, *access* to markets does vary substantially.[10] The percentage of farmers who own transport vehicles, which are necessary to participate effectively in input and output markets, ranges from a low of 38 percent in the off-farm labor cluster to a high of 85 percent in the dairy cluster. The same pattern emerges for co-op membership, with 55 percent of dairy farmers reporting membership versus 2 percent for farms specializing in off-farm labor.

Spending on modern agricultural inputs, measured in terms of absolute expenditures and relative to the value of total output, varies across (and in some cases within) clusters (Table 14). Coffee producers led the way in absolute levels with an index value of 148 corn units[11] in 1984, while at the other end of the spectrum farmers in the off-farm labor cluster reported using only 27 units.

Table 14—Components of the index of modern inputs, mean values, by cluster, 1984

Component	Coffee N = 122	Corn N = 114	Dairy N = 71	Off-farm Labor N = 48	Rice N = 22
Hybrid seeds	5.27	4.34	2.72	0.99	1.41
	(7.83)	(7.60)	(2.92)	(1.62)	(3.48)
Chemical fertilizer	113.16	20.89	16.40	8.97	5.75
	(147.43)	(43.31)	(25.42)	(16.36)	(11.46)
Mechanical tillage	16.46	4.78	3.07	6.98	9.36
	(37.88)	(13.85)	(5.62)	(36.10)	(26.95)
Pesticides	5.46	1.06	1.07	0.51	0.36
	(11.79)	(3.84)	(1.90)	(1.80)	(0.74)
Total crop	140.34	31.07	23.26	17.46	16.87
	(186.89)	(56.72)	(27.56)	(50.50)	(37.09)
Modern feeds	4.11	3.75	28.19	6.57	7.07
	(15.51)	(9.22)	(87.04)	(31.56)	(14.68)
Veterinary services	2.13	2.58	4.28	1.09	2.04
	(2.47)	(3.63)	(2.20)	(1.84)	(3.66)
Total livestock[a]	7.38	7.55	41.37	9.36	10.86
	(17.58)	(12.46)	(94.18)	(32.10)	(14.94)
Total (crops and livestock)	147.72	38.62	64.62	26.83	27.74
	(190.51)	(62.46)	(100.99)	(58.88)	(40.47)

Source: Derived from the PRODEMATA data set, University of Viçosa, 1986.
Note: The numbers in parentheses are standard deviations.
[a]The total livestock index includes a small amount of modern inputs for forage crops.

[10] For a more detailed discussion of definitions of market access, see Wanmali 1992.
[11] Corn units will be defined and discussed in the next chapter.

Regarding expenditures on modern inputs relative to the index of the value of total crop and livestock production, farms in the off-farm labor cluster are the most intensive users. This result confirms the important link between the cash flow generated by wage labor and decisions on use of technology at the farm level (Reardon and Vosti 1992). Factors influencing farm-level use of modern inputs are the focus of the analysis in subsequent chapters.

As expected, the composition of expenditures on modern inputs varies substantially across clusters. Table 14 presents a breakdown of expenditures by type for the entire sample, as well as for the five major clusters. While farms in the coffee cluster use consistently more of all types of modern inputs for crop production, their use of chemical fertilizer is disproportionately high vis-à-vis farms from other clusters. In the use of modern inputs in livestock production, dairy farms predictably use more modern feeds and veterinary services than the other farms.

As a result of differences in land quality, access to infrastructure and markets, and the past and current region-specific emphases on particular agricultural products, the geographic distribution of clusters across the 12 *municipios* represented in the PRODEMATA study is not uniform. Table 15 presents the locations of farms by cluster for 1984. Farms specializing in coffee production are concentrated in Carangola, Manhuaca, and Raul Soares. The distribution of corn producers is somewhat more uniform, but Alto Rio Doce contains the greatest number. Milk production is concentrated in Carangola, Muriaé, and Raul Soares, and off-farm labor comes predominantly from farms located in Santos Dumont and Uba. The few rice producing farms in the sample are concentrated in the *municipio* of Muriaé.[12]

Table 15—Geographic distribution of clusters, 1984

Municipio	Coffee	Corn	Dairy	Off-farm Labor	Rice	Total
Alto Rio Doce	0	32	1	4	1	38
Juiz De Fora	0	4	7	4	0	15
São João Nepomuceno	0	0	3	2	2	7
Santos Dumont	0	1	4	8	0	13
Carangola	23	5	10	3	0	41
Leopoldina	0	5	7	2	6	20
Manhuaca	57	6	0	1	0	64
Muriaé	10	15	15	6	9	55
Ervalia	10	11	1	1	0	23
Ponte Nova	3	5	3	5	0	16
Raul Soares	18	15	16	0	0	49
Uba	1	15	4	12	4	36
Total	122	114	71	48	22	377

Source: Derived from the PRODEMATA data set, University of Viçosa, 1986.

[12] Initial within-*municipio* sampling schemes and subsequent nonuniform attrition rates led to large differences across *municipios* in absolute sample size, which may have biased the geographic distribution of farm clusters in 1984.

6

UNIVARIATE ANALYSES

This chapter examines some simple relationships between scale of operation (as measured by area of farm or an index of total output), farm-level diversity of product mix, and measures of agricultural modernization (based on modern input use).

The relationships are estimated separately for each cluster. It is apparent from the previous discussion that such relationships are "confluent," that is, they represent joint relationships between two endogenous variables. The variable used to stratify the sample—the cluster to which the observations belong—is not strictly exogenous either. But the cluster to which an observation belongs captures many of the unobserved exogenous effects of land, soil type, climate, and particular attributes of management or infrastructure that can explain both scale and diversification within cluster groups and modernization. Within such a context, the question of the joint, net relation between scale and diversity and use of modern inputs can be addressed.

Results are presented here only for 1984. Analyses have been carried out separately for five groups:

	Output Mix	Number
Coffee	67 percent coffee, 8 percent corn	122
Corn	27 percent corn, 12 percent beans, 11 percent pork	114
Dairy	57 percent dairy, 9 percent rice	71
Rice	50 percent rice, 11 percent dairy	22
Off-farm labor	61 percent off-farm labor, 9 percent corn	48

The number of farms included in the plots and in the regression analyses is sometimes less than the number in the original groups because of missing data.

The term "corn units" used throughout this analysis means that all variables expressed in cruzeiros have been divided by the mean corn price in the year in question (here, 1984) for the entire sample of corn-producing farmers. (Corn is a good choice for a deflator because most farmers produce it in every subregion of the Zona da Mata. Clearly, for a single cross-section there is no need to deflate generally, but recall that analyses have been carried out for each of the six years, 1979–84.) In 1984, the average price per 50-kilogram sack was Cr10,590. Natural logarithms of all the indexes, except for the index of diversity, which is in percentage terms, are taken; for reference, $4 \equiv \log 55$, $5 \equiv \log 150$, $6 \equiv \log 400$, all in corn units.

The construction of the indexes was discussed in Chapter 4. The index of modern input use is a value-weighted quantity index of inputs defined as modern. For crops, modern inputs are hybrid seeds, chemical fertilizers, insecticides, herbicides, and

mechanical traction. Traditional inputs are labor and standard seeds, organic fertilizer, and animal traction. For livestock, modern inputs are protein concentrates, mineral supplements, balanced rations, vaccines, and other veterinarian expenses. Traditional inputs to livestock production are labor and standard feeds such as ground corn. The indexes used in this chapter are as follows:

- Total area under cultivation in hectares
- Total output
- Total crop output
- Total livestock output
- Off-farm labor income
- Modern inputs use
- Modern inputs used in crop production
- Modern inputs used in livestock production
- Diversity: percentage of total output accounted for by the top two products
- Diversity of crop production: percentage of total crop output accounted for by the top crop produced
- Diversity of livestock production: percentage of total livestock output accounted for by the top livestock product

Simple regressions between pairs of these variables or their logarithms make it possible to examine the relation of scale and product diversity to the use of modern inputs in crop and livestock production in the five product clusters chosen for analysis. Plots of four of these relationships are presented: (1) log of total output versus log area, (2) log of spending on modern input use index versus log of total output, (3) log of spending on modern input index versus percent of total output accounted for by the two top products, and (4) percent of total output accounted for by the two top products versus the log of total output.

Since logarithms are taken of all but the percentage variables, this report deals for the most part with double log relationships:

$$\log y = a + b \log x, \tag{4}$$

which in absolute form corresponds to $y = Ax^b$, where $\log A = a$. Thus, a negative intercept corresponds to $A < 1$. The slope b is the elasticity or percentage increase in y given a 1 percent increase in x. So $b > 1$ implies that y is increasing more than proportionately with x, while $b < 1$ implies that y increases less than proportionately with x.

With these facts in mind, the regression results and plots for each of the five groups at the fifth cluster level in 1984 are examined. The results are summarized in Table 16, the last two columns of which give regression results for the pooled cluster data. The second to the last column constrains all coefficients to be the same between clusters, and the last column constrains all coefficients except the intercept to be the same. The F-test in the second to the last column then tests the equality of both the intercepts and the slope coefficients between clusters. The F-test in the last column tests the equality of only the slope coefficients between

Table 16—Univariate relations

Independent Variable	Group: Coffee (1)	Corn (2)	Dairy (3)	Off-farm Labor (4)	Rice (5)	Coefficient: All Restricted (6)	Slope Restricted (7)
	Dependent variable: log total output (crop and livestock)						
1. Intercept	4.92	3.98	3.83	4.43	4.33	4.16	...
(Standard error)	(0.19)	(0.20)	(0.37)	(0.22)	(0.49)	(0.12)	...
Log area	0.58	0.62	0.72	0.25	0.52	0.65	0.55
(Standard error)	(0.06)	(0.07)	(0.09)	(0.10)	(0.18)	(0.04)	(0.04)
R^2	0.45	0.40	0.45	0.12	0.29	0.44	0.57
	N=122	N=114	N=71	N=47	N=22	F=26.51**	F=3.51**
	Dependent variable: log modern input use (total)						
2. Intercept	–3.10	–2.47	–2.18	–2.67	–3.66	–3.15	...
(Standard error)	(0.69)	(0.53)	(0.72)	(1.94)	(1.75)	(0.34)	...
Log total output (total)	1.09	0.95	0.87	0.88	1.06	1.05	0.98
(Standard error)	(0.10)	(0.09)	(0.11)	(0.33)	(0.30)	(0.05)	(0.06)
R^2	0.50	0.50	0.49	0.16	0.41	0.52	0.56
	N=118	N=109	N=70	N=40	N=20	F=8.62**	F=0.55
	Dependent variable: log modern input use (crop)						
3. Intercept	–2.81	–1.91	–0.87	–1.37	–4.19	–2.36	...
(Standard error)	(0.69)	(0.45)	(0.71)	(0.79)	(2.47)	(0.27)	...
Log total output (crop)	1.06	0.88	0.68	0.79	1.14	0.97	0.98
(Standard error)	(0.10)	(0.09)	(0.13)	(0.17)	(0.47)	(0.05)	(0.06)
R^2	0.47	0.50	0.31	0.38	0.31	0.56	0.54
	N=118	N=108	N=63	N=38	N=15	F=4.75**	F=1.16
	Dependent variable: log modern input use (livestock)						
4. Intercept	–2.23	–2.02	–2.98	1.06	0.57	–1.27	...
(Standard error)	(0.52)	(0.77)	(0.82)	(1.00)	(0.91)	(0.30)	...
Log total output (livestock)	0.80	0.77	0.92	0.21	0.36	0.64	0.70
(Standard error)	(0.11)	(0.16)	(0.13)	(0.25)	(0.19)	(0.06)	(0.07)
R^2	0.42	0.22	0.42	0.03	0.20	0.31	0.35
	N=82	N=81	N=70	N=21	N=16	F=3.94**	F=2.92**
	Dependent variable: log modern input use (total)						
5. Intercept	1.95	3.93	4.31	4.07	5.47	2.58	...
(Standard error)	(1.00)	(0.60)	(0.81)	(1.43)	(2.13)	(0.39)	...
Percent max 2	0.029	–0.017	–0.010	–0.019	–0.040	0.012	–0.007
(Standard error)	(0.012)	(0.01)	(0.011)	(0.018)	(0.028)	(0.005)	(0.006)
R^2	0.05	0.03	0.01	0.03	0.10	0.01	0.20
	N=118	N=109	N=70	N=40	N=20	F=21.83**	F=3.09**
	Dependent variable: log modern input use (crop)						
6. Intercept	2.63	3.10	3.08	1.32	5.59	2.09	...
(Standard error)	(0.83)	(0.43)	(0.44)	(0.73)	(2.03)	(0.29)	...
Percent max (crop)	0.020	–0.008	–0.006	0.013	–0.054	0.015	–0.002
(Standard error)	(0.010)	(0.008)	(0.007)	(0.010)	(0.027)	(0.004)	(0.004)
R^2	0.03	0.01	0.01	0.05	0.23	0.04	0.27
	N=118	N=108	N=63	N=38	N=15	F=27.51**	F=3.53**

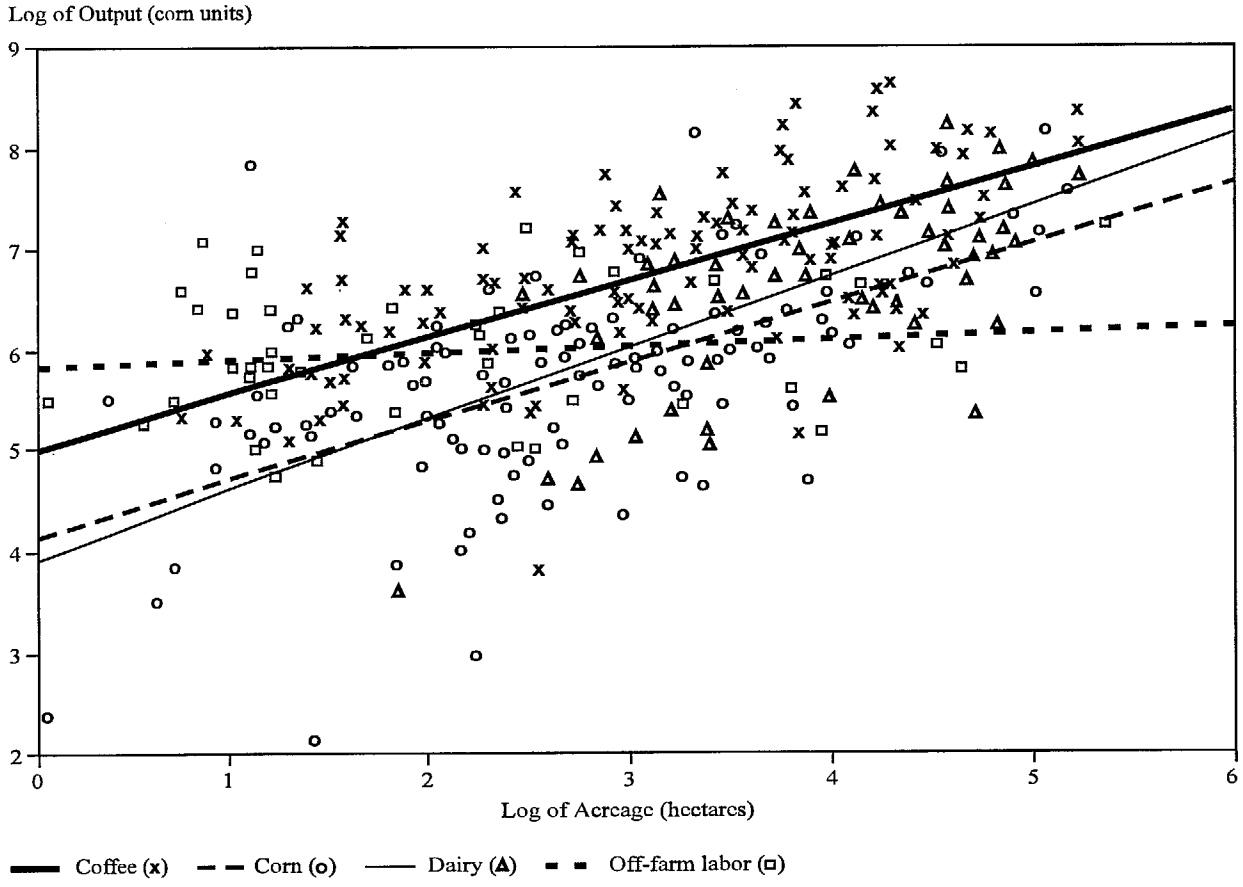
(Continued)

Table 16—Continued

Independent Variable	Group: Coffee (1)	Corn (2)	Dairy (3)	Off-farm Labor (4)	Rice (5)	Coefficient: All Restricted (6)	Slope Restricted (7)
	Dependent variable: log total output (crop and livestock)						
7. Intercept	1.91	2.62	1.27	1.40	3.03	1.36	...
(Standard error)	(0.50)	(0.73)	(1.11)	(1.46)	(0.93)	(0.34)	...
Percent max 2 (livestock)	–0.004	–0.017	0.018	0.006	–0.012	0.009	–0.004
(Standard error)	(0.008)	(0.012)	(0.014)	(0.019)	(0.013)	(0.005)	(0.005)
R^2	0.004	0.03	0.03	0.01	0.06	0.12	0.12
	N=82	N=81	N=70	N=21	N=16	F=7.45**	F=1.11
	Dependent variable: percent max 2 (total)						
8. Intercept	64.84	70.85	66.55	65.20	73.37	51.92	...
(Standard error)	(6.60)	(6.78)	(10.97)	(16.57)	(18.33)	(4.58)	...
Log total output (total)	2.50	–1.95	1.38	2.72	–0.53	3.49	0.45
(Standard error)	(0.98)	(1.18)	(1.65)	(2.79)	(3.15)	(0.73)	(0.67)
R^2	0.05	0.02	0.01	0.02	0.001	0.06	0.38
	N=122	N=114	N=71	N=48	N=22	F=47.21**	F=2.28
	Dependent variable: percent max (crop)						
9. Intercept	56.96	66.92	63.32	78.77	66.20	48.08	...
(Standard error)	(7.40)	(7.36)	(13.97)	(12.71)	(24.04)	(4.49)	...
Log total output (crop)	3.43	–2.38	–0.42	–2.34	1.77	3.41	–0.22
(Standard error)	(1.12)	(1.40)	(2.64)	(2.78)	(4.43)	(0.79)	(0.85)
R^2	0.07	0.03	0.00	0.02	0.01	0.05	0.28
	N=112	N=114	N=67	N=46	N=22	F=28.74**	F=2.21
	Dependent variable: percent max (livestock)						
10. Intercept	80.83	96.78	61.20	108.99	92.97	78.12	...
(Standard error)	(6.68)	(5.11)	(8.82)	(7.02)	(11.82)	(3.19)	...
Log total output (livestock)	–3.63	–7.91	3.02	–8.81	–4.44	–1.90	–5.00
(Standard error)	(1.47)	(1.15)	(1.41)	(1.94)	(2.77)	(0.67)	(0.70)
R^2	0.05	0.30	0.06	0.34	0.12	0.02	0.26
	N=117	N=113	N=71	N=42	N=21	F=28.26**	F=6.60**
	Dependent variable: log modern input use (total)						
11. Intercept	3.32	2.31	0.90	0.55	0.78	2.74	...
(Standard error)	(0.67)	(0.53)	(1.12)	(1.66)	(0.94)	(0.36)	...
Log off-farm labor	0.16	0.11	0.67	0.38	0.57	0.08	0.27
(Standard error)	(0.18)	(0.15)	(0.26)	(0.31)	(0.28)	(0.08)	(0.10)
R^2	0.03	0.01	0.24	0.04	0.32	0.01	0.23
	N=29	N=44	N=23	N=40	N=11	F=10.53**	F=1.05
	Dependent variable: modern input use/output (crop and livestock)						
12. Intercept	0.200	0.077	–0.026	0.373	0.104	0.070	...
(Standard error)	(0.058)	(0.032)	(0.100)	(0.417)	(0.148)	(0.052)	...
Log off-farm labor	–0.0226	–0.0005	0.0238	–0.0359	0.0032	0.0111	–0.0099
(Standard error)	(0.0153)	(0.0089)	(0.0231)	(0.0772)	(0.0433)	(0.0122)	(0.0256)
R^2	0.074	0.000	0.048	0.006	0.001	0.006	0.054
	N=29	N=44	N=23	N=40	N=11	F=1.83	F=0.30

**Significant at the 5 percent level.

Figure 2—Plot 1: Total output versus farm area



Source: Derived from the PRODEMATA data set, University of Viçosa, 1986.
Notes: N = 355 farms in 1984: 122 coffee farms, 114 corn, 71 dairy, and 48 off-farm laborers.
$R^2 = 0.57$

clusters (leaving the intercepts unconstrained). Asterisks following the F-statistic indicate rejection of the hypothesis that the coefficients are equal across clusters at the 5 percent significance level.

Plot 1: Log of Total Output versus Log of Farm Area

In Plot 1 (Figure 2), all slope values are less than one and significantly different from one another (for example, $F = 3.51$ in section 1, column 7 of Table 16); and the intercept terms also differ significantly (for example, $F = 26.51$ in section 1, column 6 of Table 16).[13] The main intercluster difference is the off-farm labor cluster, whose slope coefficient is markedly lower than the others. The other four groups, whose major outputs are on-farm products, are similar, with slope coefficients ranging from 0.52 for rice farms to 0.72 for dairy farms. Thus there is evidence of sharply diminishing returns in on-farm output to size of farm (measured by area).

For the purposes of this report, area is not a good measure of the scale of the operation. Presumably, the returns diminish because farms with a larger area under cultivation have poorer quality land. This is borne out by the fact that the regression for farms in the dairy cluster, which have a larger mean size, has the largest slope coefficient; these results imply more land in pasture but a smaller decline in output (and presumably, quality) as size increases. It is remarkable, however, that despite the great differences between clusters with respect to product mix, after allowing for differences in intercept (which primarily reflect differences in means), the different types of farms with major on-farm production are characterized by a similar degree of diminishing returns to land area.

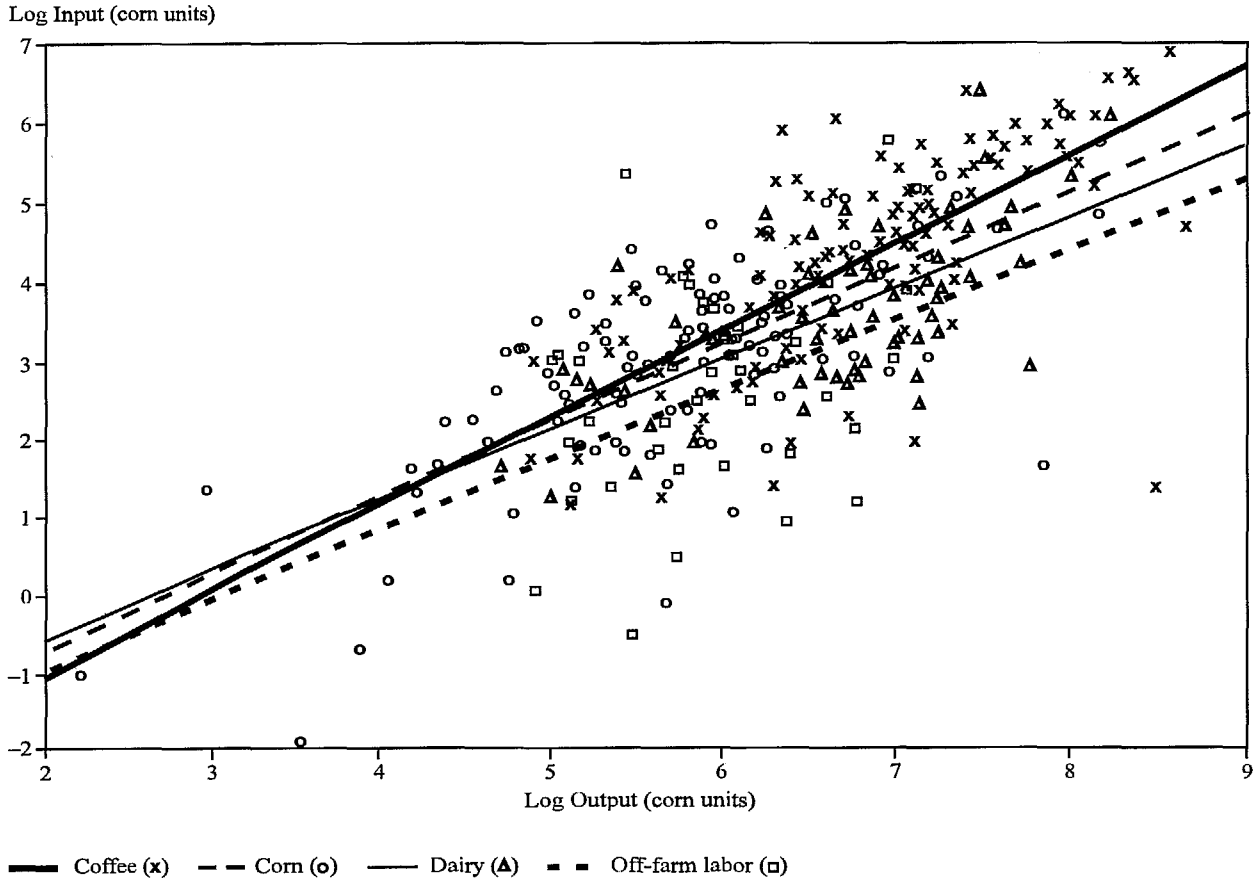
Plot 2: Log of Modern Input Use versus Log of Total Output

The next set of regressions deals with the relation between total use of modern inputs for crops and livestock and the scale of farming operations measured by the product index. Plot 2 (Figure 3) depicts the relations between the logs of these variables for each cluster. The slopes, which are much closer to 1 than those in Plot 1, differ insignificantly across clusters.

To test the consistency of this relationship across crops and livestock products, an additional equation is estimated for each. Splitting modern input use and output between crops and livestock reduces the common slope to 0.70 for livestock but leaves crops unchanged (as reported in sections 3 and 4 of Table 16). With these disaggregated variables, however, more significant differences show up between clusters. In the livestock relation, the off-farm labor cluster has a much smaller slope coefficient (actually insignificantly different from zero), indicating the traditional nature of livestock production on these smaller farms. In the crop relation, the slope

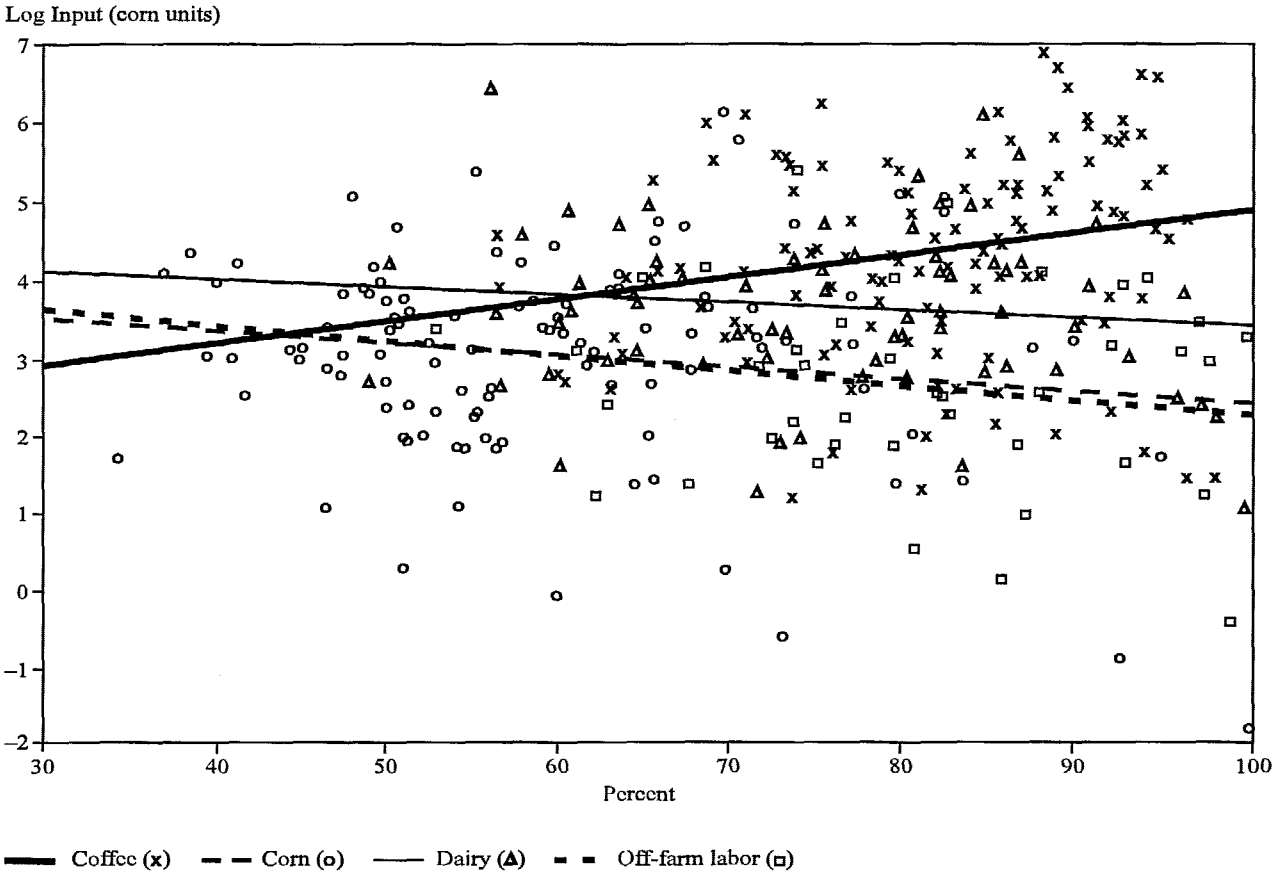
[13]When equality of slopes is rejected, the equality of intercepts test in column 6 is not valid; that is, the test in column 6 is valid only if one assumes slopes are equal. Once equality of slopes is rejected, the equality of intercepts is generally uninteresting.

Figure 3—Plot 2: Use of modern inputs versus total output



Source: Derived from the PRODEMATA data set, University of Viçosa, 1986.
Notes: N = 337 farms in 1984: 118 coffee farms, 109 corn, 70 dairy, and 40 off-farm laborers.
$R^2 = 0.56$

Figure 4—Plot 3: Use of modern inputs versus percent of total output in top two crops



Source: Derived from the PRODEMATA data set, University of Viçosa, 1986.
Notes: N = 337 farms in 1984: 118 coffee farms, 109 corn, 70 dairy, and 40 off-farm laborers.
$R^2 = 0.20$

coefficient for the rice cluster is considerably different from the others, but has a very high standard error, in part due to the small number of farms in this cluster.

Spending on modern inputs was expected to increase more than proportionately with scale of operation, but these findings show that they do not in the Zona da Mata. Moreover, this finding is robust with respect to type of farming.

Plot 3: Log of Modern Input Use versus Percent of Total Output Accounted For by the Top Two Products

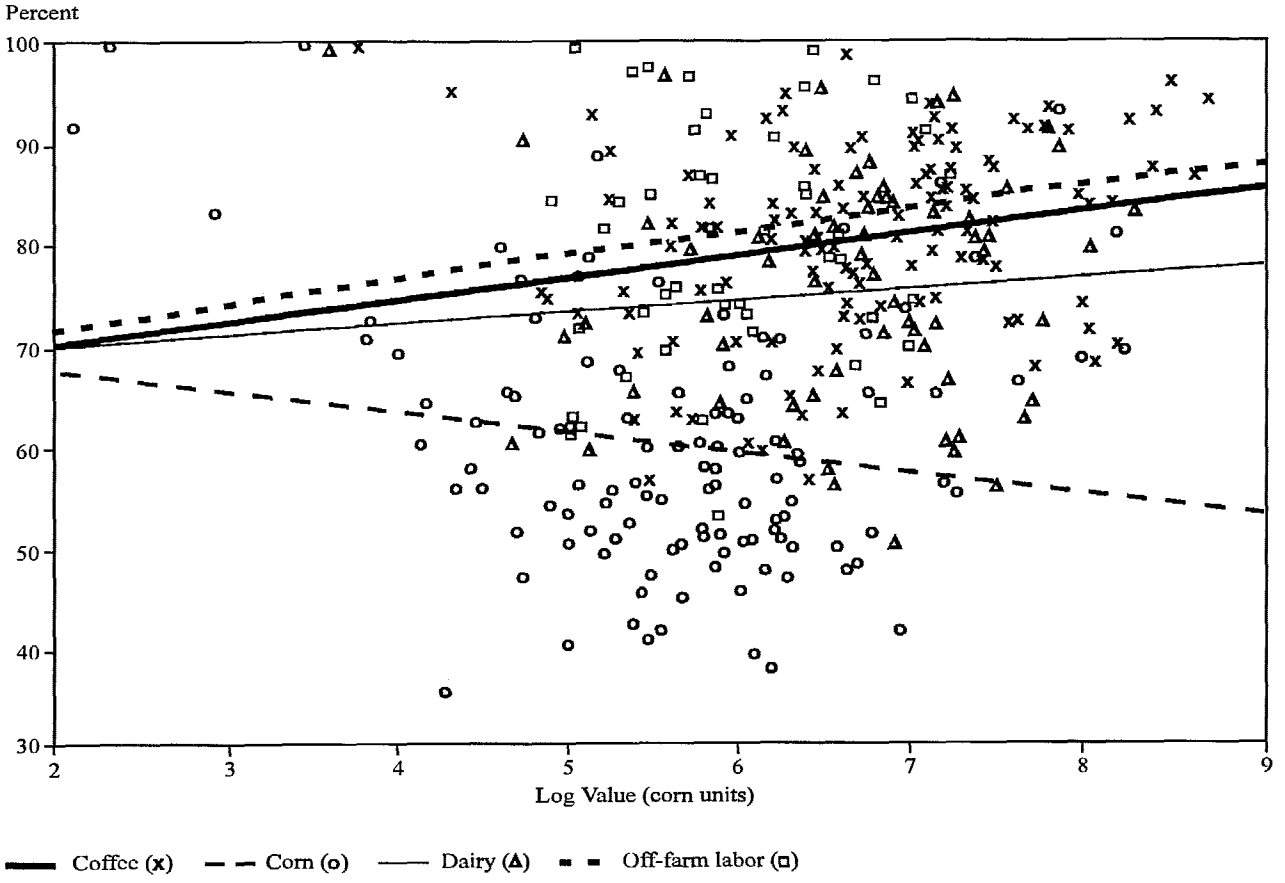
The percent of total output accounted for by the top two products is a measure of the degree of farm-level output specialization (that is, it is inversely related to how diversified the farm is). Plot 3 (Figure 4) indicates that there is a small, but significant, relation between farm-level diversity and the use of modern inputs (see section 5, column 6 of Table 16) for some types of farms. The F-test suggests that there are significant differences in slope coefficients between clusters, with coffee farms responding positively (in terms of modern input use) to increased specialization, and corn farmers displaying just the opposite tendency.

This negative relation, –0.007 (± 0.012 for a 5 percent confidence interval), characterizing the slope of the regression line relating specialization (this time measured in terms of the top crop or livestock product) to use of modern inputs, generally reacts strongly to distinctions between crop and livestock inputs (sections 6 and 7, Table 16). Such a negative relation indicates a tendency for farms that spend more on modern inputs to be less specialized. The total and crop regression, however, shows significant differences among clusters. Coffee is the cause here; increases in product concentration are significantly associated with increased use of modern inputs.

Plot 4: Percent of Total Output Accounted For by the Top Two Products versus Log of Total Output

Plot 4 (Figure 5) and sections 8, 9, and 10 of Table 16 show large differences among farms of different scale (measured in total output) with respect to diversity in production within clusters and among farms across different clusters. However, because of the low overall correlation between these two variables, the differences across clusters are not statistically significant. For coffee farms, specialization increases significantly with scale, with a 1 percent increase in gross output leading to a 2.5 percentage point increase in the fraction accounted for by the top two products (section 8, column 1 in Table 16). Corn farms, on the other hand, indicate a negative relation between product specialization and total output, while the other three groups (dairy, off-farm labor, and rice) show no significant link between the two. A similar pattern emerges from the crop regression, when crop and livestock output are disaggregated. The livestock regression, however, shows a significant negative relation between specialization in livestock production and total livestock output for all groups except the dairy cluster (which shows a significant, positive relation between the two variables).

Figure 5—Plot 4: Percent of total output in top two products versus total output



Source: Derived from the PRODEMATA data set, University of Viçosa, 1986.
Notes: N = 355 farms in 1984: 122 coffee farms, 114 corn, 71 dairy, and 48 off-farm laborers.
$R^2 = 0.38$

Plot 5: Modern Input Use versus Off-Farm Labor

No plots are presented. The regression slopes are not significantly different from one another but are significantly positive and significantly less than one (see Table 16, section 11). However, as discussed below, when one controls for scale of operation, no significant links can be found between off-farm labor earnings and use of modern inputs (Table 16, section 12).

To explore these relationships further, three multiple regressions are examined:

- Log of modern input use versus percent of total output in the top two products and log of total output;
- Percent of total output in the top two products versus log of modern input use and log of total output; and
- log of modern input use versus log of total output and log of off-farm labor.

Table 17—Bivariate relations

	Group					Coefficient	
				Off-farm		All Restricted[a]	Slope Restricted[b]
Independent Variable	Coffee (1)	Corn (2)	Dairy (3)	Labor (4)	Rice (5)	(6)	(7)
	Dependent variable: log modern input use (total)						
1. Intercept	–2.84	–1.71	–0.99	–1.54	–0.39	–2.65	. . .
(Standard error)	(0.87)	(0.70)	(0.83)	(2.07)	(2.08)	(0.38)	
Percent max 2	–0.005	–0.012	–0.019	–0.030	–0.048	–0.011	–0.016
(Standard error)	(0.009)	(0.007)	(0.007)	(0.016)	(0.020)	(0.004)	(0.004)
Log total output (total)	1.11	0.94	0.91	1.02	1.12	1.11	1.01
(Standard error)	(0.11)	(0.09)	(0.10)	(0.32)	(0.27)	(0.06)	(0.06)
R^2	0.50	0.51	0.54	0.24	0.56	0.53	0.58
	N=118	N=109	N=70	N=40	N=20	F=9.68**	F=2.63**
	Dependent variable: percent max 2						
2. Intercept	53.10	59.00	51.72	45.60	49.14	37.12	. . .
(Standard error)	(7.38)	(7.92)	(11.50)	(18.54)	(18.08)	(5.23)	
Log modern input use	–0.46	–2.20	–4.71	–2.83	–5.19	–2.20	–2.27
(Standard error)	(0.91)	(1.31)	(1.82)	(1.52)	(2.18)	(0.73)	(0.63)
Log output (total)	4.45	–1.21	6.14	7.03	6.78	6.94	3.81
(Standard error)	(1.41)	(1.76)	(2.27)	(3.31)	(3.62)	(1.07)	(0.93)
R^2	0.12	0.03	0.11	0.13	0.26	0.12	0.41
	N=118	N=109	N=70	N=40	N=20	F=43.19**	F=4.43**
	Dependent variable: log modern input use (total)						
3. Intercept	–3.35	–1.05	–2.32	–1.93	–0.42	–1.87	. . .
(Standard error)	(1.06)	(0.95)	(1.31)	(1.59)	(1.79)	(0.48)	
Log total output (crop and livestock)	1.26	0.70	0.80	0.88	0.31	0.83	0.85
(Standard error)	(0.18)	(0.15)	(0.23)	(0.25)	(0.39)	(0.07)	(.10)
Log off-farm labor	–0.20	–0.05	0.18	0.03	0.41	0.05	–0.02
(Standard error)	(0.12)	(0.13)	(0.25)	(0.27)	(0.35)	(0.06)	(0.08)
R^2	0.65	0.35	0.52	0.28	0.37	0.47	0.51
	N=29	N=44	N=23	N=39	N=11	F=2.29	F=2.26

[a]The *F*-statistic in column (6) tests the equality of the intercept across groups.
[b]The *F*-statistic in column (7) tests the equality of all slope coefficients across groups.
**Significant at the 5 percent level.

These results are presented in Table 17 in the same format as in Table 16 (sections 2 and 3). In the second equation in Table 17, both intercepts and slopes differ among farming clusters. Moreover, the slope coefficients are virtually identical to the corresponding coefficients in the univariate relations (sections 2 and 5, column 7 of Table 16). This indicates that although the relation between use of modern inputs and specialization is significantly negative, it is relatively small compared with the proportionate positive relation to scale of operation. Remember, however, that a 1 percentage point increase in specialization represents a much larger percentage increase, since the specialization index is always less than 1.

The regression treating diversity as a dependent variable and log use of modern inputs and log gross output as independent variables (section 3, Table 17) shows significant differences among clusters. Bear in mind that the relations among these variables are not structural but jointly confluent. Again there is a negative correlation between specialization and use of modern inputs. The effect of scale on specialization is significantly positive for all clusters except corn.

Returning to the relation between use of modern inputs and total output value (section 2, column 4, Table 16), one notes that the group whose primary output is off-farm labor has a small coefficient relative to the other clusters (although with a relatively high standard error). To determine the effect of off-farm labor as a secondary product, several further regressions are examined (sections 11 and 12, Table 16, and section 1, Table 17). In section 11, Table 16, which shows the results for the univariate regression of log use of modern inputs on the log of the index of off-farm labor, there is a small positive relation. Since, however, a significant positive proportionate relation has already been found between expenditures on modern inputs and scale of operation (section 2, Table 16), the composition of total output needs to be controlled.

Such a regression is shown in section 3, Table 17: log use of modern inputs on log output of crops and livestock and, separately, log of off-farm labor are regressed, allowing the effects of these separate components of total output to be distinguished. The coefficients on log output of crops and livestock are large, positive, and significant, while the coefficients on log of off-farm labor for the clusters are not significantly different from zero. The rice group has no coefficients that are significantly different from zero, not even the intercept. The high standard errors are caused in part by the small sample of rice farms.

Section 12, Table 16 uses an alternative measure of modern input use (adjusted for scale of operation by dividing by the index of output of crops and livestock). The regression of this variable on the log of output of off-farm labor shows no significant relation between off-farm labor and the degree of modernization.

To summarize, some differences were found among types of farms in the simple univariate and bivariate relationships studied in this chapter. There is clear indication that returns diminish as farms grow larger, which is related to the quality of land on larger farms. When, however, this factor is eliminated by measuring scale of operation in terms of gross output at constant prices, use of modern inputs (overall or specific to crops or livestock) increases roughly proportionately with scale for all types of farms. Output of off-farm labor as a secondary product has no significant relation to use of modern inputs. Adjusted for scale of operation, farms with off-farm labor as primary output use proportionately more modern inputs. The notable exception to uniformity across clusters is the relation between the use of modern inputs and the degree of specialization, which is positive for some types of farms and negative for others.

To disentangle the joint confluent relation between scale of operation and specialization, on the one hand, and use of modern inputs, on the other, requires a multiple regression analysis. The analysis presented in Table 16, section 2, suggests that the univariate relation (or lack thereof) holds up quite well when allowance is made for the relation between scale and modern input use: expenditure on modern inputs varies proportionately with scale irrespective of type of farming; diversity varies with scale depending on product mix and type of farming (Table 16, section 3); and there is a small but significantly negative correlation between specialization and use of modern inputs.

7

LONGITUDINAL ANALYSIS

To this point, attention has focused on defining farm types by their concentration on various agricultural products and exploring differences in expenditures on modern farm technology among the farm types at a particular moment in time. In this chapter, the analysis is extended to a longitudinal study of farm-level changes in product mix, to test for the influence of changes in product mix on modern input use.

Several factors are expected to influence farmers' decisions to change product mix radically over time (Vosti and Witcover 1996).[14] First, neither the price policies nor the agricultural extension available to farmers were uniform for all products. For example, a long-term initiative to increase coffee production was introduced in the Zona da Mata over part of the sample period, with preferential input and output prices and credit terms (vis-à-vis competing crop and livestock products) (University of Viçosa 1986, 62). These measures successfully induced 36 farms in the sample used here to become primarily coffee producers. This represents a 20 percent increase in the size of the coffee cluster from 1979 to 1984. Second, in spite of the highly variable agroecological microregions in the Zona da Mata, movements across the four clusters retained in the longitudinal analysis (coffee, corn, dairy, and off-farm labor) are quite feasible. While this mountainous region does not generally lend itself to agricultural production, farmers can (technologically) shift from one of the four major product clusters to another without encountering insurmountable agroecological obstacles. The same is probably not true for rice and perishable crops, which require flat irrigated lands (*varzeas*) for production.

Third, the timing and magnitude of investments and returns are expected to vary dramatically across products, and, in the absence of perfect capital markets, may serve to limit the flow of farmers from one production cluster to another, in response to economic and other incentives.

Based on these expected differences and the preliminary descriptive analyses, farms in the PRODEMATA data panel are divided into four groups based on the stability of their product mix and the direction of cluster movements over the six years covered in the data series. Farms that remained within the same cluster for all of the six years are labeled "stable." They represent 38 percent of the farms included in the panel. Farms located at the border of a cluster or farm type and those that switched clusters for only one year and then returned to their original cluster are labeled "marginal." These "marginally stable" farms make up 33 percent of the sample. Finally, farms that clearly jumped from one cluster to another during the six-year period are labeled "jumpers"; they make up 29 percent of the sample.

[14]It should be emphasized that shifts from one cluster to another over time do *not* refer to changes in relative prices in these measures of farm-level production, since stable, inflation-free corn units are used to aggregate farm output, although changes in output mix could be a response to relative price variations.

Each cluster is divided into "central" and "peripheral" areas for each year. Farms that remained in the central area of their respective cluster over the entire sample period are labeled "stable." Two-year presence in the central areas of at most two clusters is required to define the "to coffee" and "to other" transitional groups. For example, farms that spent two or more years in the central area of a noncoffee cluster at the beginning of the sample period and later moved to the central area of the coffee cluster for at least two years are labeled "to coffee." Farms that remained in peripheral areas but did not display erratic behavior are labeled "marginally stable." Due to erratic product mix changes, 27 farms are excluded from the longitudinal analysis.

Jumpers are further subdivided into two categories: those that switched to coffee production, labeled "to coffee," and those that switched to another cluster, labeled "to other."

The characteristics of the stable, marginally stable, to coffee, and to other groups are then examined to determine how they differed (if at all), especially with regard to their respective use of modern farm technology.

Table 18 reports the means and standard deviations for a series of farm-level and other characteristics in 1984 for all groups. With the exception of age of household head, family size, the dependency ratio, and the percent of farms electrified, the equality of all variable means can be rejected at the 5 percent level.

A fairly consistent hierarchy emerges from an examination of the descriptive statistics in Table 18. The farms classified as stable tend to be best off in access to land and infrastructure as well as in endowments of human capital. Stable farms are least likely to be involved in sharecropping arrangements and are well above other groups in total output. Average modern input use is highest among the stable farmers—both absolutely and in relation to total on-farm output.

Farms that jumped to coffee are somewhat smaller, are much more likely to be involved in sharecropping, and tend to be situated farther from major market towns than stable farms. However, farms that adjusted their output mix in favor of coffee production enjoyed higher yields than any other group. Total use of modern inputs on farms that jumped to coffee is second only to that of stable farms.

Marginally stable and farms that jumped to other products tend to be the worst off. Farm size drops substantially in these groups, as does output, access to infrastructure, and cooperative membership. Income derived from off-farm sources rises sharply. Total modern input use decreases markedly to a low of 27.2 input units for farms that jumped to other products.

A similar pattern emerges for the intensity with which modern inputs are used (vis-à-vis total farm output). The ratio of modern input use to crop and livestock output is highest for stable farms and lowest for farms that jumped to other products.

In summary, descriptive statistics indicate that stable and jumped-to-coffee farms tend to be better off, and their privileged status (especially in terms of total output) is linked to modern input use. However, this association must be tested.

Table 19 presents univariate regression results for each of the longitudinal groups. The model tests for differences within and across groups in the influence of the scale of operation (total output) on modern input use. Results support the mean differences across groups in modern input use suggested by the descriptive analyses, but they do not support differences across groups with respect to the effects of output generation on input use. Indeed, as indicated in Chapter 6, for cross-cluster analysis, modern input use increases approximately proportionately with increased agricul-

Table 18—Descriptive statistics within transitional groups, 1984

Group	Stable N = 133	Marginally Stable N = 116	Jump to Coffee N = 36	Jump to Other N = 65
Mean age of household head	57.5	55.9	55.9	57.6
	(10.6)	(11.8)	(11.8)	(11.6)
Education of household head	2.5	1.8	1.9	1.8
	(1.9)	(1.9)	(1.9)	(2.2)
Family size	5.4	5.7	5.6	5.1
	(2.7)	(2.6)	(3.1)	(2.4)
Dependency ratio	0.17	0.19	0.23	0.20
	(0.20)	(0.20)	(0.24)	(0.22)
Percent electrified	48	34	44	32
	(50)	(48)	(50)	(47)
Farm size	45.5	24.4	39.9	29.9
	(44.1)	(30.6)	(38.4)	(35.8)
Value of total output[a]	1,162	675	873	455
	(1,129)	(768)	(663)	(405)
Value of on-farm production[a]	1,130	611	844	383
	(1,132)	(779)	(659)	(396)
Value of off-farm labor[a]	32.4	64.3	28.4	71.9
	(76.9)	(126.5)	(11.8)	(142.7)
Yield (output/hectare)[a]	36.4	37.6	40.7	24.8
	(31.6)	(43.1)	(34.5)	(25.5)
Percent sharecroppers	5	13	22	17
	(22)	(34)	(42)	(37)
Distance to market	23.2	21.3	28.4	21.7
	(13.7)	(13.8)	(11.8)	(13.5)
Percent who owned transport vehicles	80	57	75	62
	(40)	(50)	(43)	(49)
Percent co-op members	32	10	28	18
	(46)	(30)	(45)	(39)
Percent output in top two products	77.4	73.0	76.0	65.0
	(13.3)	(15.9)	(10.9)	(16.0)
Total value of modern inputs used[a]	129.1	52.3	85.1	27.2
	(189.8)	(76.8)	(104.6)	(48.3)
Total modern inputs used/ total output (crop and livestock)	0.112	0.102	0.093	0.087
	(0.105)	(0.194)	(0.063)	(0.114)

Source: Derived from the PRODEMATA data set, University of Viçosa, 1986.

Notes: The numbers in parentheses are standard deviations. The likelihood ratio test for equality of variance among the four groups rejected equality at the 5 percent significance level for all variables except age, education, family size, dependency ratio, percent electrified, distance to market, and percent with transport. The *F*-test for equality of means, adjusted for heteroskedasticity where appropriate, rejected equality of means at the 5 percent level for all variables except age, family size, dependency ratio, and percent electrified.

[a]All values for outputs and inputs are expressed in terms of the corn-based index derived in Chapter 4.

tural production for stable and marginally stable farms and both types of jumpers. That is, a 1 percent increase in the index of total output will generate (on average) a 1 percent increase in the index of modern input use, regardless of the nature of the farm-level product mix and the stability of product mix over time.

Finally, in Table 20, which presents changes in total output and input use over time by transitional groups, there is a high degree of variability of output production and input use among all farms during the 1979–84 period. Supply constraints associated with problems such as credit availability and poor physical and marketing

Table 19—Modern input use within and across transitional groups

Independent Variable	Stable	Marginally Stable	To Coffee	To Other
	Dependent variable: log of modern input use			
Intercept	–2.55	–2.78	–3.04	–2.57
(standard error)	(0.57)	(0.68)	(1.06)	(0.86)
Log total output	0.99	0.98	1.06	0.90
(standard error)	(0.08)	(0.11)	(0.16)	(0.15)
R^2	0.51	0.44	0.57	0.39
	N=133	N=116	N=36	N=65

Notes: Equality of intercept: F=4.27** (asterisks imply significant differences at the 5 percent level). Equality of slopes (unrestricted intercepts): F=0.14

Table 20—Changes in total output and use of modern inputs by transitional groups, 1979–84

Group	Number of Farms	1979 Mean	Change from Previous Year				
			1980	1981	1982	1983	1984
			(percent)				
Stable	133						
Value of total output		912.1 (920.3)	–3.4	24.6	–7.9	18.7	–3.1
Value of modern inputs		94.1 (133.4)	29.8	–2.8	8.6	4.1	–3.8
Marginally stable	116						
Value of total output		540.0 (672.8)	–1.0	35.2	–14.2	15.8	–6.0
Value of modern inputs		59.6 (123.6)	6.5	18.5	–12.2	7.7	–26.5
Jumpers to coffee	36						
Value of total output		565.3 (575.9)	14.0	27.9	–11.2	26.7	–5.8
Value of modern inputs		74.4 (112.5)	14.5	11.3	20.1	–8.5	–18.4
Jumpers to others	65						
Value of total output		472.5 (591.1)	–5.0	–1.1	–7.2	25.7	–12.2
Value of modern inputs		50.5 (132.2)	17.6	–25.8	–39.0	–18.7	24.2
All groups	350						
Value of total output		671.5 (779.2)	–1.5	24.5	–10.0	19.4	–5.2
Value of modern inputs		72.5 (129.2)	20.3	0.8	–0.4	2.0	–10.3

Note: The numbers in parentheses are standard deviations.

infrastructure are probably responsible for some of these variations, which seem to affect virtually all farms regardless of characteristics. While no clear pattern emerges, growth rates of total output and input use among stable and jump-to-coffee farms tend to be higher and more stable than those for marginally stable farms and those that jumped to other products.

8

CONCLUSIONS AND POLICY IMPLICATIONS

Past literature on adoption of technology has emphasized the causal link between changes in agricultural technology and subsequent changes in product mix, emphasizing the product-specific nature of most technological innovations and the policies designed to speed their adoption and use. Implicit in this framework is the notion that farmers producing technologically privileged products and those willing to switch to these products will use modern inputs more intensively, both absolutely and relatively. In this report, a broader framework has been adopted, emphasizing though not explicitly modeling the complex joint relationships among product mix and diversification, adoption of modern agricultural technologies, and use of purchased (modern) inputs.

In the detailed analyses of the degree of output diversity presented in this report, product mix and output concentration are found to vary substantially across farms in this diversified agricultural region, but there is little evidence of complete specialization (as predicted by some theories) in any particular agricultural product.

Off-farm employment is an important "output" of many farms in this region. It represents farm-level interaction with local agriculture and nonagriculture sectors, not seasonal or long-term migration by household members.

When product concentration ratios are used to construct by cluster analysis a series of farm types and to test for differences in the factors influencing modern input use, univariate analyses fail to show significant links between measures of output diversification and expenditures on modern inputs within farm types. The notable exception is the coffee group, where the relation is significant and positive. Bivariate regression analyses confirm the significant and proportional relation between scale of operation (as measured by a corn-based index of total output) and modern input use, but again no relationship is shown between the concentration in product mix and use of modern agricultural inputs for any product mix, except coffee. In sum, empirical results suggest no significant cross-sectional impact of product mix on technology adoption—either as the concentration of product mix changes within farm types or as absolute product mix changes across farm types.

Distinguishing between farmers with unstable versus stable product mixes over time, descriptive statistics suggest a statistically robust hierarchy of transitional farm groups, with stable farmers being best off in aggregate output, followed by farms that jumped to coffee, marginally stable farms, and farms that jumped to production of other crops. Substantial mean differences in modern input use were confirmed across transitional groups, but the link between modern input use and scale of operation across transitional groups was quite similar and roughly one-to-one (on a double log scale).

The differences (or lack of differences) revealed by the univariate, multivariate, and longitudinal analyses are potentially important to policymakers and to the individuals and institutions involved in the generation and dissemination of new agricul-

tural technologies. First, if the univariate analyses are correct and the use of modern inputs varies more or less proportionately with the scale of output for all farms, regardless of product mix or output diversification, then focusing technology generation and agricultural extension efforts on particular types of farming systems or farm sizes because of expected differences in adoption rates is unlikely to yield greater returns than approaching farms more uniformly. That is to say, the long-term stability of the confluent relationships examined here suggests no significant gains to focusing on one particular set or scale of farming operation. Symmetrically, the results also suggest that the calculation of expected benefits of agricultural research projects need not be weighted by adoption probabilities based on product mix or output diversification in order to establish research priorities or design policies.

Second, that there are no clear longitudinal trends in input use suggests that supply constraints (and, particularly in this setting, credit constraints) may be a primary determinant of technology adoption, and these supply constraints affect all farms more or less equally.

Finally, and perhaps most important in this agriculturally marginal area, a substantial number of farms radically altered output mix over the sample period, suggesting that ecological factors play a limited role in determining product mix, even in this ecologically highly heterogeneous area. This demonstrated willingness (and ability) to react to economic and other incentives suggests that consistent agricultural policies play an important part in bringing about desired change even among Brazil's poorest and most traditional farmers.

TECHNICAL APPENDIX

The purpose of cluster analysis is to group observations of a data set into clusters such that observations within a cluster are similar—that is, they resemble each other more than they resemble those in different clusters. The term encompasses a wide variety of techniques, each requiring different prior assumptions about the structure of the data set. In comparison with the assumptions imposed by other methods, however, the assumptions associated with cluster analysis require little previous knowledge of the data. Thus the technique is useful for preliminary characterization of a data set that the researcher suspects may consist of several differing groups. The clusters obtained can be useful in the specification of more explicit modeling techniques such as regression analysis.

Since cluster analysis does not require explicit specification of a model, many of the serious errors caused by model misspecification are avoided. However, since the common grouping algorithms and testing procedures do not generally have a statistical basis, most clustering procedures are highly subjective. In most types of cluster analysis, therefore, retaining this subjective character means trading off the statistical foundation that accompanies more explicit assumptions.

The preceding paragraph gives an intuitive definition of a cluster. Any specific clustering method, however, must have a more precise definition. Put another way, given a data set of N observations on p variables, what is meant by two observations being "similar?" If all the variables are continuous and of a quantitative type, the standard gauge of similarity is a familiar distance measure. When the variables are all of similar scale and range, a standard metric is usually appropriate, such as the Euclidean distance in p-dimensional vector space,

$$\left[\sum_{k=1}^{P}\left(x_{ik}-x_{jk}\right)2\right]^{1/2}, \tag{5}$$

where x_{ik} and x_{jk} represent the value for the kth variable associated with the ith and j^{th} observations, or the city block metric,

$$\sum_{k=1}^{P}\left|x_{ik}-x_{jk}\right|. \tag{6}$$

If the variables have relatively different scale or range, the measure may be standardized by first dividing by the overall standard deviation, or by using the Mahalanobis' distance, $(X_i - X_j)' \; S^{-1} \; (X_i - X_j)$, where X_i and X_j are the p-vectors representing the ith and j^{th} observations, $'$ denotes transpose of the vectors, and S^{-1} is the inverse of the data set's $p \times p$ sample covariance matrix.

Note that most distance measures will be biased toward finding clusters of a spherical shape. If it is known that the clusters have a specific elongated shape, it might be better to use a measure where some variables carry greater weight than others. Since in most preliminary studies such information is not available, however, the equal weight distance measures are most practical. If some of the variables are discrete or qualitative, many different measures of similarity, usually based on

matching coefficients, have been suggested, each appropriate to a particular type of variable and clustering algorithm. Anderberg (1973) presents a good overview of these measures.

Once a suitable similarity measure has been defined, a specific algorithm for clustering the observations must be chosen. There are two basic types of methods, hierarchical and nonhierarchical. In hierarchical clustering, successive groupings or divisions of the data set result in a clustering tree or dendrogram indicating the cluster in which each observation is placed in each level, from one (all the observations in the same cluster) to N (each observation in its own cluster). At each level, say $K + 1$, the clusters are subsets of those in the preceding level, K. That is, at any given level K, one cluster will subdivide into exactly two level $K + 1$ clusters, with each observation appearing in exactly one cluster per level. The actual generation of the dendrogram can proceed from either end of the "tree."

In nonhierarchical clustering, there is no such link between the different levels. Each "level" here represents the grouping of the data set into a different number of clusters. For each level, a grouping results that may have no relation to the grouping of the preceding level. Many nonhierarchical techniques do not use the distance measure directly. Rather, an objective function such as the pooled within-cluster sum of squared deviations is used to compare different clustering alternatives. Included in this group are several iterative maximum-likelihood reallocation procedures.

Hierarchical techniques are useful when there are strong prior indications of a hierarchical structure, such as the taxonomic classification in biology; they also generally require less computing power. Nonhierarchical techniques, on the other hand, generally impose more distributional assumptions on the data. A third type of procedure often regarded as cluster analysis is called "mixture analysis." In mixture procedures, rather than giving each observation a specific cluster allocation, a vector of probabilities is generated indicating the probability that the observation is in any given cluster. These procedures are generally of more statistical validity, but will not be discussed as they were not useful for this study (see Hawkins, Muller, and ten Krooden 1982).

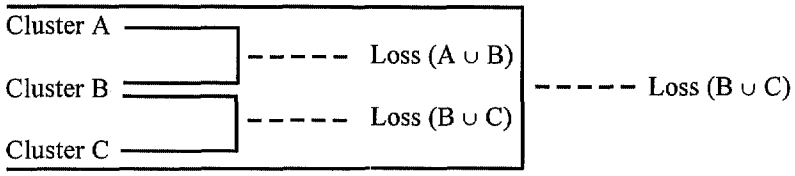
Since none of the techniques usually sets an internal criterion for the proper number of clusters in the data set, this decision is left to the researcher, regardless of the choice of techniques. Numerous criteria of varying statistical basis can be useful in making this decision. Most criteria require some subjective input.

Once the cluster results are obtained, estimates of the mean and covariance matrix of the cluster could be obtained, or further analysis such as regressions or analysis of variance could be performed. Except for mixture analysis, such estimates will in general be biased.

Hierarchical Techniques

Hierarchical techniques include agglomerative methods, which proceed by successive fusions of the N observations into progressively fewer clusters, and divisive methods, which begin with one cluster and successively divide the observations into progressively more clusters. All agglomerative techniques can be understood within a framework proposed by Ward (1963). In Ward's paradigm, the two clusters merged at each stage are chosen to maximize a specific objective or loss

function. Agglomerative methods differ only in the choice of objective function. The figure below outlines this procedure from the three-cluster stage to the two-cluster stage, where Loss (A ∪ B) refers to the value of the loss function associated with combining the clusters A and B into one cluster.



The combination resulting in the minimum loss then becomes the clustering for the next stage. Proceeding to each stage, K clusters, requires computing the objective function $\sum_{j=1}^{K} j$ times.

Table 21 lists several common agglomerative methods and their associated objective functions. Average linkage and Ward's method use objective functions closer to the intuitive concept of a cluster, but in real or simulated data situations, single and complete linkages often recover known clusters just as well (see for example, Milligan 1980).

Divisive methods of hierarchical clusters work in the opposite direction, proceeding from one cluster containing all observations by successive divisions until each observation is separate. The calculations necessary are much greater. For example, in making the first division of the data set, $2^{N-1} - 1$ different combinations are possible. Divisive techniques are accordingly less common and were not considered for this study. See Dillon and Goldstein (1984) for further details.

Table 21—Agglomerative methods of hierarchical cluster analysis

Method	Description
Single linkage	Minimum distance between two observations, one from one cluster and one from another
Complete linkage	Maximum distance between two observations, one from one cluster and one from another
Average linkage	Average distance between all pairs of observations, one from one cluster and one from another
Ward's method	Increase in the pooled within-cluster sum of squared deviations: $\sum_{i=1}^{k} E_k E_k$, is the sum of squared deviations for cluster k; $E_k = \sum_{i=j}^{p} \sum_{j=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2$, where i indexes variables, j indexes observations in cluster k, and $\bar{x}_i$ is the cluster mean for variable i

As mentioned above, the result of hierarchical cluster analysis is an entire tree, with allocations of the data set at every level. Nothing in the procedure itself indicates which level best represents the data set; that is, what the true number of clusters is. For agglomerative procedures, the value of the loss function at each successive stage is a monotonically increasing sequence, since if there existed a combination of clusters resulting in a lower loss than the minimum of the preceding stage, such a fusion should have occurred at that level. So for hierarchical techniques, inferring the number of clusters is equivalent to choosing the maximum value of the loss function that the researcher finds acceptable (Hawkins, Muller, and ten Krooden 1982). Posed in this fashion, the choice of clusters must, in general, be a subjective decision. However, as some guidance is frequently desirable, statistics used for nonhierarchical techniques are often consulted. These will be discussed below.

In conclusion, although hierarchical techniques only apply intuitively when a hierarchical structure is suspected, the agglomerative methods generally perform well. The primary advantages of hierarchical analysis include less computation and the absence of any distributional assumptions underlying the techniques. In this respect, these procedures may actually be more robust than maximum-likelihood methods.

Nonhierarchical Techniques

In nonhierarchical analysis, each level corresponding to a specific number of clusters in the data set is analyzed separately (with no relation to either preceding or successive levels). If the researcher is reasonably certain of the number of clusters present, only that level and perhaps a few around it need be clustered. Because each level requires separate analysis, nonhierarchical methods are usually computer-intensive, iterative algorithms, requiring some assumptions about the probability distribution of the data.

The methods discussed are generally called "partitioning" or "allocation" techniques, since for a given number of clusters the result is an allocation of each observation to a specific cluster. In theory, all these methods operate by optimizing a specified objective function across all the possible combinations of data points. Obviously, the number of comparisons necessary is usually prohibitive to an exhaustive search; for instance, partitioning 16 observations into 4 clusters results in 171,798,901 possible combinations (Hawkins, Muller, and ten Krooden 1982). Therefore, most allocation methods use either a simple sorting procedure or an iterative hill-climbing approach to optimize the objective function. As with all such algorithms, the solution approaches a local optimum, which may not coincide with the global optimum across all combinations. Thus, some process for checking global optimization is often suggested.

The simplest of the allocative methods is called K-means clustering (MacQueen 1967), or the nearest-centroid sorting method (Forgy 1965). Given a p-variate profile on the data points, the centroid of a cluster is defined as the mean vector of all p variables across the observations allocated to the cluster, $\bar{x}_k = \sum_{i=1}^{n} x_{ik}$, where, x_{ik} is the p-vector associated with the ith observation in the kth cluster.

To initiate the K-means clustering algorithm, K initial centroids must be chosen. Each observation is then considered in turn and allocated to the cluster whose centroid is closest to it. As with hierarchical procedures, different distance measures could be considered but the standard Euclidean distance is most common. To result in a local optimum, iteration of this simple process is necessary. For the second iteration, the new cluster centroids are computed by averaging the observations allocated in the first iteration and then performing a second pass through the data as before.

Forgy's centroid sorting method differs in that the cluster centroid is recomputed each time an observation is allocated rather than at the end of one iteration, as in K-means sorting. For this reason, the order in which the data points are considered will affect the allocation. It has been shown that these procedures will iterate to convergence and result in a local optimum of the pooled within-cluster sum of squared deviations as defined in Table 21 (Anderberg 1973, 165).

Note that the specific local optimum achieved is dependent on the initial centroids chosen and, for Forgy's method, the order of the observations in the data set. Methods of choosing the initial centroids include choosing the first K observations in the data set, choosing K observations from the data at random, and choosing K observations spaced relatively far apart, using more complicated algorithms. Another suggestion is to first perform a hierarchical analysis of the data, such as average linkage or Ward's method. Then take the clustering at the Kth level and compute these centroids to use as initial points. Usually, there are relatively few observations reallocated from the hierarchical solution; thus, this method can decrease the number of iterations required for convergence.

Regardless of how the initial centroids are chosen, however, a local optimum is the result. To check against the global optimum the algorithm should be repeated several times with different initial points. A common procedure is to define a grid over the possible initial points and search over this grid for the maximum local optimum. Such grids, however, become progressively less practical as the number of variables increases.

The objective function used in these allocative procedures can usually be identified with an assumption about the observations' probability distribution implicit in the choice of objective. Following the framework of Hawkins, Muller, and ten Krooden (1982), the data points can be viewed as a random sample from a mixture of K probability distributions, usually the normal distribution. The approach of allocative procedures is to view each observation as generated by two independent samplings. First, the specific distribution is randomly chosen from the K present in the population; and second, the specific observation, X, is drawn from this distribution. By this interpretation, each observation can be associated with one specific distribution or cluster. Table 22 shows several common distributional assumptions and the associated objective function whose optimization results in the maximum-likelihood allocation under that assumption. In each case, the assumed distribution is normal, $N()$, and the optimum objective is the minimum. W is the pooled within-cluster scatter or sample covariance matrix,

$$W_k = \sum_{j=1}^{n} (x_j - \bar{x}_k)(x_j - \bar{x}_k)', \tag{7}$$

Table 22—Assumptions and objective functions of allocative methods

Distribution Assumed	Resulting Objective Function
$N(\mu_k, \delta^2 I)$: homoskedastic and spherical,	Trace (W)
$N(\mu_k, \Sigma)$: homoskedastic,	Determinant $(W) = \mid W \mid$
$N(\mu_k, \Sigma_k)$: general heteroskedastic normal	$\prod_{k=1}^{k} \mid W_k/n_k \mid n_k$

where X_j is the p-vector representing the jth data point and $\bar{x}_k$ is the mean vector for cluster k, and $W = \sum_{k=1}^{k} W_k$. The trace(W) objective is then used in Ward's hierarchical method or the K-means sorting algorithm. So, using either of the iterative sorting methods discussed above involves the implicit assumption that all the clusters in the data set represent a random sample from a mixture of homoskedastic, spherical normal distributions. In other words, one assumes no correlation between variables and the same variance for all clusters. These are obviously questionable assumptions, but the more complicated objective functions are orders of magnitude more difficult to optimize. More sophisticated algorithms have been developed by Friedman and Rubin (1967), Scott and Symons (1971), and Jensen (1968).

Such allocative procedures are intuitively appealing, especially to a statistician, due to their maximum-likelihood interpretation. But as mentioned before, misspecification of the implied assumptions can theoretically decrease the efficiency of these methods. However, for the preliminary studies in which cluster analysis is most often used, such theoretical concerns are not critical.

Inference Testing and Estimation for Allocative Methods

One inference usually of interest is testing model fit; that is, testing the underlying assumptions used in choosing the objective function, particularly the homoskedasticity and spherical nature of the covariance matrix. The simplest procedure for this is to examine the sample covariance matrix for the solution obtained. This is obviously not statistically rigorous, but for allocative methods few of the classical procedures are applicable.

The second inference of interest is the determination of the true number of clusters in the data set. This can be viewed in two different ways. The first would be testing the null hypothesis of no clusters in the data set versus an alternative of K clusters, while the second would be a test between the hypothesis of K clusters versus $K + 1$ clusters. For the first test, the most obvious statistic, particularly given the homoskedastic assumption, is the classical F statistic for equality of several means,

$F = \frac{(N-K)\ tr(B)}{(K-1)\ tr(W)}$, where $tr(\ )$ refers to the trace, or sum of diagonal elements, of the

matrix; W is the pooled within-cluster sample covariance matrix defined previously, and B is the between-cluster covariance matrix,

$$B = \sum_{k=1}^{k} n_k(\bar{x}_k - \bar{x})\,(\bar{x}_k - \bar{x})', \tag{8}$$

$\bar{x}$ being the mean vector across all data points in the set. Other suggested statistics include $tr(W)$ and $|T|/|W|$, where $T = W + B$ is the total covariance matrix across all data points. However, since the clusters are generated by minimizing an objective function related directly rather than randomly to these statistics, these tests are not strictly valid. The appropriate statistics are max(F), min(trW), and max($|T|/|W|$), where max( ) means the maximum value of the statistic across all possible groupings.

The distribution of max(F) will not be the standard F distribution, and thus the actual distribution also needs to be approximated. Lee (1979) and Arnold (1979) both propose versions of the max($|T|/|W|$) statistic along with approximations of its distribution function. Arnold believes that the standard test as usually stated is not well posed. He prefers a two-step procedure that first tests the null hypothesis that the data are concentrated about one point, unimodal distribution, versus the alternative that the data are either uniformly distributed or grouped into clusters. If this null hypothesis is rejected, the second test is the null hypothesis of uniformly distributed data versus the remaining alternative, two or more clusters. His version of the statistic is log[max($|T|/|W|$)]. Both Lee's and Arnold's statistics were successful in distinguishing between clustered and unclustered data, using both simulated and real data sets.

These tests can be used to infer the presence of clusters, but they do not help distinguish between K clusters and $K + 1$ clusters. Attempts to extend statistics like those Lee or Arnold proposed to test K clusters against $K + 1$ clusters have met with little success. Most commonly used criteria are extensions of the classical statistics, such as F or $tr(W)$. For instance, a frequent suggestion is to graph the F statistic versus the number of clusters and pick out the maximum. This procedure is usually known as the Carlinski-Harbasz criterion. Milligan and Cooper (1985) report the performance of this criterion and several others in detecting the number of clusters in simulated data. The Carlinski-Harbasz criterion was one of the better performers under all conditions. In general, any of these ad hoc criteria perform better or worse as the data set comes closer to or moves farther from their underlying assumptions. One difficulty with most of these criteria is that dividing by $(K - 1)$ means that the criterion is usually strictly increasing in the lower range of cluster numbers and as such is not likely to distinguish between, say, three and four clusters.

Since the aim of clustering the data is to examine the differences and similarities of the obtained clusters, estimation of cluster parameters like the mean and variance are important. Unfortunately, the main result in this area is that for allocative and hierarchical methods the usual estimates will have significant bias. As stated earlier, mixture analysis techniques theoretically do not result in biased estimates. So if unbiased estimates are important, it might be necessary to do mixture analysis along with the allocative method to obtain both a specific allocation and good estimates.

BIBLIOGRAPHY

Aldenderfer, M. S., and R. K. Bashfield. 1984. *Cluster analysis*. Beverly Hills, Calif.: Sage Publications.

Anderberg, M. R. 1973. *Cluster analysis for applications*. New York: Academic Press.

Arnold, S. J. 1979. A test for clusters. *Journal of Marketing Research* 16 (November): 545–551.

Ayres, C., M. Nerlove, S. A. Vosti, and M. Weeks. 1991. Agricultural change in Brazil's agroecological zones from 1970–1985: A descriptive analysis. Paper prepared in conjunction with EMBRAPA/IFPRI collaborative research contract (SP 2262-001).

Banfield, C. F., and L. C. Bassill. 1977. A transfer algorithm for nonhierarchical classification. *Applied Statistics* 26: 206–210.

Bradley, T. 1990. Technology adoption in Brazil. Ph.D. diss., University of Pennsylvania, Philadelphia.

David, C. C., and K. Otsuka, eds. 1994. *Modern rice technology and income distribution in Asia*. Boulder, Colo., U.S.A.: Lynne Rienner.

de Haen, H. 1991. Environmental consequences of agricultural growth. In *Agricultural sustainability, growth, and poverty alleviation: Issues and policies*, ed. S. A. Vosti, T. Reardon, and W. von Urff. Proceedings of the conference held September 23–27, 1991, Feldafing, Germany. Deutsche Stiftung fur internationale Entwicklung (DSE) and the International Food Policy Research Institute.

Dillon, W. R., and M. Goldstein. 1984. *Multivariate analysis*, New York: John Wiley.

Everitt, B. 1977. *Cluster analysis*. London: Heinemann Educational Books.

Feder, G., R. E. Just, and D. Zilberman. 1985. Adoption of agricultural innovations in developing countries: A survey. *Economic Development and Cultural Change* 33 (no. 2): 255–298.

Fisher, W. 1953. On a pooling problem from the statistical decision viewpoint. *Econometrica* 21 (October): 567–585.

———. 1958. On grouping for maximum homogeneity. *Journal of the American Statistical Association* 53: 789–798.

Forgy, E. S. 1965. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. Paper prepared for the WNAR meetings, University of California, Riverside, Calif., U.S.A.

Friedman, H. P., and J. Rubin. 1967. On some invariant criteria for grouping data. *Journal of the American Statistical Association* 62: 1159–1179.

Haavelmo, T. 1944. The probability approach to econometrics. Supplement to *Economia* 12 (July): 1–115.

Hartigan, J. A. 1975. *Clustering algorithms*. New York: John Wiley.

Hawkins, D. M., M. W. Muller, and J. A. ten Krooden. 1982. Cluster analysis. In *Topics in applied multivariate analysis*, ed. D. M. Hawkins. Cambridge: Cambridge University Press.

Hazell, Peter B. R., and C. Ramasamy. 1991. *The Green Revolution reconsidered: The impact of high-yielding rice varieties in South India*. Baltimore, Md., U.S.A.: Johns Hopkins University Press for the International Food Policy Research Institute.

Jensen, R. E. 1968. A dynamic programming algorithm for cluster analysis. *Journal of Applied Statistics* 12: 1034–1057.

Lee, K. L. 1979. Multivariate tests for clusters. *Journal of the American Statistical Association* 74: 708–714.

Lipton, M., and R. Longhurst. 1989. *New seeds and poor people*. Baltimore, Md., U.S.A.: Johns Hopkins University Press.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1.

McRae, D. J. 1970. MIKCA: A FORTRAN IV iterative *K*-means cluster analysis program. *Behavioral Science* 16: 423–424.

Milligan, G. W. 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika* 45: 325–342.

Milligan, G. W., and M. C. Cooper. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50: 159–179.

Plucknett, D. L. 1993. *Science and agricultural transformation*. Lecture Series No. 1. Washington, D.C.: International Food Policy Research Institute.

Reardon, T., and S. A. Vosti. 1992. Issues in the analysis of the effects of policy on conservation and productivity at the household level in developing countries. *Quarterly Journal of International Agriculture* 31 (4): 380–396.

Reardon, T., P. Matlon, and C. Delgado. 1988. Coping with household-level food insecurity in drought-affected areas of Burkina Faso. *World Development* 16 (9): 1065–1074.

Scott, A. J., and M. J. Symons. 1971. Clustering methods based on likelihood ratio criteria. *Biometrics* 27: 387–397.

University of Viçosa. 1986. Programa de Desenvolvimento Rural Integrado da Zona da Mata, MG (PRODEMATA). Relatório final de avaliaço (produção). Mimeo.

Vosti, S. A., and J. Witcover. 1996. Slash-and-burn agriculture: Household perspectives. *Agriculture Ecosystems and Environment* 58 (June): 23–38.

Wanmali, S. 1992. *Rural infrastructure, the settlement system, and development of the regional economy in Southern India.* Research Report 91. Washington, D.C.: International Food Policy Research Institute.

Ward, J. H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58: 236–244.

Wolfe, J. H. 1970. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research* 5: 329–350.

World Bank. 1992. *World development report: Development and the environment.* Washington, D.C.: World Bank.

World Resources Institute. 1994. *World resources 1994–95: Data base.* Washington, D.C.: World Resources Institute.

Marc Nerlove is a professor in the Department of Agricultural and Resource Economics at the University of Maryland. Stephen Vosti is a research fellow in the Environment and Production Technology Division of the International Food Policy Research Institute. Wesley Basel was formerly in the Department of Economics at the University of Pennsylvania.

RECENT IFPRI RESEARCH REPORTS (continued)

86 *THE EFFECTS OF INTERNATIONAL REMITTANCES ON POVERTY, INEQUALITY, AND DEVELOPMENT IN RURAL EGYPT,* by Richard Adams, Jr., 1991

85 *COMMERCIALIZATION OF AGRICULTURE UNDER POPULATION PRESSURE: EFFECTS ON PRODUCTION, CONSUMPTION, AND NUTRITION IN RWANDA, by Joachim von Braun, Hartwig de Haen, and Juergen Blanken, 1991*

84 *EFFECTS OF EXCHANGE RATE AND TRADE POLICIES ON AGRICULTURE IN PAKISTAN, by Paul Dorosh and Alberto Valdés, 1990*

83 *DEVELOPMENTAL IMPACT OF RURAL INFRASTRUCTURE IN BANGLADESH, by Raisuddin Ahmed and Mahabub Hossain, 1990*

82 *AGRICULTURAL GROWTH AND STRUCTURAL CHANGES IN THE PUNJAB ECONOMY: AN INPUT-OUTPUT ANALYSIS,* by G. S. Bhalla, B. K. Chadha, S. P. Kashyap, and R. K. Sharma, 1990

81 *PRODUCTION AND CONSUMPTION OF FOODGRAINS IN INDIA: IMPLICATIONS OF ACCELERATED ECONOMIC GROWTH AND POVERTY ALLEVIATION,* by J. S. Sarma and Vasant P. Gandhi, 1990

80 *HORTICULTURAL EXPORTS OF DEVELOPING COUNTRIES: PAST PERFORMANCES, FUTURE PROSPECTS, AND POLICY ISSUES,* by Nurul Islam, 1990

79 *EFFECTS OF AGRICULTURAL COMMERCIALIZATION ON LAND TENURE, HOUSEHOLD RESOURCE ALLOCATION, AND NUTRITION IN THE PHILIPPINES,* by Howarth E. Bouis and Lawrence J. Haddad, 1990

78 *THE EFFECTS OF SUGARCANE PRODUCTION ON FOOD SECURITY, HEALTH, AND NUTRITION IN KENYA: A LONGITUDINAL ANALYSIS,* by Eileen Kennedy, 1989

77 *THE DEMAND FOR PUBLIC STORAGE OF WHEAT IN PAKISTAN,* by Thomas C. Pinckney, 1989

76 *AGRICULTURAL AND ECONOMIC GROWTH IN ARGENTINA, 1913-84,* by Yair Mundlak, Domingo Cavallo, and Roberto Domenech, 1989

The International Food Policy Research Institute was established in 1975 to identify and analyze alternative national and international strategies and policies for meeting food needs of the developing world on a sustainable basis, with particular emphasis on low-income countries and on the poorer groups in those countries. While the research effort is geared to the precise objective of contributing to the reduction of hunger and malnutrition, the factors involved are many and wide-ranging, requiring analysis of underlying processes and extending beyond a narrowly defined food sector. The Institute's research program reflects worldwide collaboration with governments and private and public institutions interested in increasing food production and improving the equity of its distribution. Research results are disseminated to policymakers, opinion formers, administrators, policy analysts, researchers, and others concerned with national and international food and agricultural policy.

IFPRI is a member of the Consultative Group on International Agricultural Research and receives support from Australia, Belgium, Canada, Centre de coopération internationale en recherche agronomique pour le développement (CIRAD), China, Denmark, Food and Agriculture Organization of the United Nations, Ford Foundation, France, German Agency for Technical Cooperation (GTZ), German Federal Ministry for Economic Cooperation (BMZ), India, Inter-American Development Bank, International Development Research Centre (Canada), International Fund for Agricultural Development, Japan, Netherlands, Norway, Overseas Development Institute, Philippines, Rockefeller Foundation, Spain, Sweden, Switzerland, United Kingdom, United Nations Development Programme, United Nations Children's Fund, United States, and the World Bank.

RECENT IFPRI RESEARCH REPORTS

103 *REGIONAL TRADING ARRANGEMENTS AMONG DEVELOPING COUNTRIES: THE ASEAN EXAMPLE,* by Dean A. DeRosa, 1995

102 *SOURCES OF INCOME INEQUALITY AND POVERTY IN RURAL PAKISTAN,* by Richard H. Adams, Jr. and Jane J. He, 1995

101 *PRICING BEHAVIOR IN PHILIPPINE CORN MARKETS: IMPLICATIONS FOR MARKET EFFICIENCY,* by Meyra Sebello Mendoza and Mark W. Rosegrant, 1995

100 *ADOPTION OF HYBRID MAIZE IN ZAMBIA: EFFECTS ON GENDER ROLES, FOOD CONSUMPTION, AND NUTRITION,* by Shubh K. Kumar, 1994

99 *ECONOMIC REFORM IN EUROPE AND THE FORMER SOVIET UNION: IMPLICATIONS FOR INTERNATIONAL FOOD MARKETS,* by Rod Tyers, 1994

98 *THE CHANGING PUBLIC ROLE IN A RICE ECONOMY APPROACHING SELF-SUFFICIENCY: THE CASE OF BANGLADESH,* by Francesco Goletti, 1994

97 *TRADE PESSIMISM AND REGIONALISM IN AFRICAN COUNTRIES: THE CASE OF GROUNDNUT EXPORTERS,* by Ousmane Badiane and Sambouh Kinteh, 1994

96 *POVERTY, HOUSEHOLD FOOD SECURITY, AND NUTRITION IN RURAL PAKISTAN,* by Harold Alderman and Marito Garcia, 1993

95 *SECTORAL GROWTH IN CHILE: 1962-82,* by Juan Eduardo Coeymans and Yair Mundlak, 1993

94 *FERTILIZER USE ON SMALLHOLDER FARMS IN EASTERN PROVINCE, ZAMBIA,* by Dayanatha Jha and Behjat Hojjati, 1993

93 *ECONOMIC INCENTIVES AND COMPARATIVE ADVANTAGE IN INDONESIAN FOOD CROP PRODUCTION,* by Leonardo A. Gonzales, Faisal Kasryno, Nicostrato D. Perez, and Mark W. Rosegrant, 1993

92 *FAMINE IN ETHIOPIA: POLICY IMPLICATIONS OF COPING FAILURE AT NATIONAL AND HOUSEHOLD LEVELS,* Patrick Webb, Joachim von Braun, and Yisehac Yohannes, 1992

91 *RURAL INFRASTRUCTURE, THE SETTLEMENT SYSTEM, AND DEVELOPMENT OF THE REGIONAL ECONOMY IN SOUTHERN INDIA,* by Sudhir Wanmali, 1992

90 *LABOR IN THE RURAL HOUSEHOLD ECONOMY OF THE ZAIRIAN BASIN,* by Tshikala B. Tshibaka, 1992

89 *GROWTH IN JAPAN'S HORTICULTURAL TRADE WITH DEVELOPING COUNTRIES: AN ECONOMIC ANALYSIS OF THE MARKET,* by Masayoshi Honma, 1991

88 *DROUGHT AND FAMINE RELATIONSHIPS IN SUDAN: POLICY IMPLICATIONS,* by Tesfaye Teklu, Joachim von Braun, and Elsayed Zaki, 1991

87 *INCENTIVES AND CONSTRAINTS IN THE TRANSFORMATION OF PUNJAB AGRICULTURE,* by Anya McGuirk and Yair Mundlak, 1991

(continued on inside back cover)

International Food Policy Research Institute
1200 Seventeenth Street, N.W.
Washington, D.C. 20036-3006 U.S.A.