AGENCY FOR INTERNATIONAL DEVELOPMENT
WASHINGTON, D. C. 20523

BIBLIOGRAPHIC INPUT SHEET

FOR AID USE ONLY
BATCH 60

1. SUBJECT CLASSIFICATION

A. PRIMARY: Food production and nutrition — AS00-0000-G635

B. SECONDARY: Human nutrition--India

2. TITLE AND SUBTITLE

Technical data and methods: sample design,questionnaire,data technology,data analysis,and systems modeling

3. AUTHOR(S)

(101) Cantor,S.M.,Associates Inc.,Haverford,Pa.

4. DOCUMENT DATE	5. NUMBER OF PAGES	6. ARC NUMBER
1973	149p.	ARC

7. REFERENCE ORGANIZATION NAME AND ADDRESS

Cantor

8. SUPPLEMENTARY NOTES *(Sponsoring Organization, Publishers, Availability)*

(In Tamil Nadu Nutrition Project,v.2,Sect.A)

9. ABSTRACT

10. CONTROL NUMBER

PN-AAD-773

11. PRICE OF DOCUMENT

12. DESCRIPTORS

India
Project planning
Surveys
Tamil Nadu, India?

13. PROJECT NUMBER

14. CONTRACT NUMBER

AID/nesa-399 GTS

15. TYPE OF DOCUMENT

AID 590-1 (4-74)

THE TAMIL NADU NUTRITION STUDY

VOLUME II SECTION A

TECHNICAL DATA AND METHODS

Report to the U. S. Agency for International
Development Contract No. AID/nesa-399
MISSION TO INDIA

SIDNEY M. CANTOR ASSOCIATES INCORPORATED Haverford, Pennsylvania 19041 U. S. A.

SIDNEY M. CANTOR ASSOCIATES INCORPORATED

THE TAMIL NADU NUTRITION STUDY

VOLUME II SECTION A

TECHNICAL DATA AND METHODS

SAMPLE DESIGN; QUESTIONNAIRE;
DATA TECHNOLOGY; DATA ANALYSIS
SYSTEMS MODELING

VOLUME II-A

TECHNICAL DATA AND METHODS

		Page
Chapter I.	Commentary and Introduction.	1
Chapter II.	Sample Design. S. Rajagopalan	3
Chapter III.	Field Operations in the Consumer Food Habits and Expenditure Survey.	8
	A. The Survey Instrument (Questionnaires). S. Rajagopalan	10
	B. Data Gathering Problems.	25
	1. Field Survey Reactions. S. Rajagopalan and T. P. Gopalaswamy	
	2. A Note on Nutrition Related Morbidity Data.	32
Chapter IV.	Data Technology. B. W. Marlow	34
Chapter V.	Data Analysis.	71
Chapter VI.	The Vector Model.	129
Chapter VII.	A Conventional Approach to Systems Modeling for Nutrition Planning. R. Rajagopalan	138

CHAPTER I

COMMENTARY AND INTRODUCTION

This part of the Tamil Nadu Nutrition Project Report is intended to be a guided, step by step passage through the methodological and technical structure of the project, with particular reference to the Consumer Expenditure and Food Habits Survey. This emphasis on the consumer survey will be easily understood with the realization that, not only are consumers the heart and reason for being of the food system, but their attitudes and behavior clearly indicate the probable success or failure of any efforts toward improvement of nutritional status of the population.

Throughout this volume and indeed this project, great pains have been taken to emphasize the need to avoid preconception of findings and the consequent loading and restriction thereby imposed upon design of survey instrument, data gathering, and data processing and analysis. The most compelling reason for this attitude and approach, admittedly an unusual one in the investigation of national or regional food and nutrition problems, is the natural uncertainty of perception of the primary or key questions to be answered, before the fact. It was suspected that some of the most important questions to be asked and answered would not come to light until the investigators were deeply immersed in processing and analysis of original data from the field. Experience proved this actually to be the case. Here, then, is the reason for what many observers felt —and so stated— was an unseemly preoccupation with methods and technique throughout the project. That reason is expressed most effectively, we believe, in Chapters IV and V, which describe and explain the highly flexible and adaptable, ultra utilitarian data processing and analytical methodology which have evolved in the efforts to fulfill the need implied above. While the methodology is carried to a complex stage, it must be emphasized that the degree of complexity utilized at any stage need only be as complicated as that required to resolve the problem. Moreover, as regularly stated, the data are in such condition of storage that any of a number of methodologies may be employed. The choice is in the hands of the analyst.

At the outset of the project, a serious problem at the interface of substantive and technical functions became evident. It concerned the tacit acceptance of economic laws as immutable truth to be invoked in study and analysis of any complex system having an important economic component. Cultural,

social, and political conditions have been ignored in many studies involving nutrition intervention, either on the assumption that they were unimportant, or that it was too difficult a task to produce valid information on the interactions of these factors with theoretical economic behavior. The custom has been to presume the existence of economic man, and to ignore the existence of cultural man. But cultural man is real and dominant, as anthropologists and behaviorists have been telling us much too quietly; he controls where economic man rather consistently breaks most of the rules that the economists' reason says he should follow. Study of a food system probably gives stronger expression to this dilemma than most other areas of human involvement and concern. The Tamil Nadu Nutrition Project has attempted and, we believe, succeeded in a good start on resolution of the problem. The techniques and methods applied will be readily apparent in a reading of Chapters IV and V.

One final point must be made: the technical approach to the study represents a deliberate, conscious effort to break out of the customary mold of technical assistance, which has rather consistently erred in attempting a direct transfer of Western cultural and economic traditions and practices to the radically different cultures of developing nations.

CHAPTER II

SAMPLE DESIGN OF THE CONSUMER SURVEY (FOOD HABITS SURVEY)

Objectives of the Survey. A clear, unequivocal statement of objectives is a prerequisite in reaching a decision on sampling design. It contributes to fixing on the different variables in the system about which data must be gathered. The broad objectives of the TN Consumer Survey was to understand the interplay of variables in the consumer subsystem and their probable impact on the nutritional status of the population in general and the vulnerable groups, preschool children, and pregnant and lactating mothers, in particular. The nutritional status can be gauged in part from the consumption pattern. This formed the backbone of the survey instrument. In order to understand the consumers food behavior to the greatest extent possible, it was deemed necessary to gather as much cultural, social, demographic and economic data as possible for all households. Having concluded on this objective and the consequent variables, a decision on the sample size was made based on:

1. The minimum number of units needed to make reliable estimates,

2. Personnel, time, and cost restraints.

Sample Size. Based on the experience of the National Sample Survey and the 1961 Census Food Habits Survey, a sample size of 2800 households was considered adequate for the purposes of this study. The sample size was biased towards the vulnerable target groups of the population. This was done deliberately to get a better estimate of the condition of the vulnerable groups. How to select these 2800 households then became the burden of the sampling design.

Sampling Design. The design adopted was a two-stage, stratified random sample. The first stage of sampling was village and urban blocks, and the second stage was the household. Stratification was then repeated in two stages.

First - rural and urban division.

Second - each of the strata was again divided as follows:

a. Rural areas were divided into two regions; coastal and inland. The regions contained the following districts:

Coastal Areas Districts	Inland Areas Districts
Kanyakumari	Madurai
Tirunelveli	Coimbatore
Ramanathapuram	Salem
Thanjavur	Dharmapuri
Tiruchirapalli	North Arcot
South Arcot	The Nilgiris
Chingelput	
Madras	

To cover the rural area, 360 villages were selected for the study. The number of villages selected in each district was proportionate to the 1961 district population, subject to a minimum of 16 villages per district. Table 1 shows the number of villages selected in each district. In each district again, the villages were stratified according to those major population groups, in the National Sample Survey Classification.

i. Population less than 2000
ii. Population between 2000 and 5000
iii. Population 5000 and above.

In this case the decision was urged by the subcontractor who was highly oriented to opinion surveys and urged that nutrition questions because of their subjective nature be classified accordingly. It is now felt that continuing work can effectively use the National Sample Survey Classification.

The number of villages allotted to each district was again proportionally allocated according to the number of villages in each population group and category iii was weighted 1 to 4 in this process of allocation. In retrospect, it appears that it would have been better to use the 26 homogeneous strata developed by the National Sample Survey as the base group and then select the villages in each stratum. This procedure appears more rational than attempting stratification at every stage on the basis noted. The 26

TABLE 1

SAMPLE DESIGN, RURAL AREA: DISTRICT/VILLAGE/HOUSEHOLD RELATIONSHIP

Region	Districts	Rural population in '000 (1961 census)	Proportional allocation (Villages)	Adjusted Sample Size Villages	Households
Region 1 (Coastal)	Kanyakumari	847	12	16	64
	Tirunelveli	1882	25	24	96
	Ramanathapuram	1822	25	24	96
	Thanjavur	2584	35	34	136
	Tiruchirapalli	2512	34	34	136
	South Arcot	2656	35	34	136
	Chingleput	1741	24	24	96
	Madras	-	-	-	-
	SUB-TOTAL	14044	190	190	760
Region 2 (Inland)	Madurai	2196	35	32	128
	Coimbatore	2525	40	38	152
	Salem Dharmapuri	3187	51	46	184
	North Arcot	2515	40	38	152
	The Nilgiris	229	4	16	64
	SUB-TOTAL	10652	170	170	680

According to the 1961 census, there were about 13,700 villages in Tamil Nadu.

strata of the National Sample Survey are formed by grouping teshils which are (i) continguous, (ii) have similar cropping patterns, population density and about the same altitude above sea level, and (iii) have good transport and communication facilities among them.

b. Urban areas are grouped into six classes based on population in the census. For the purpose of this survey using comparable census classifications, urban areas were stratified into three zones:

	No. of Towns
Zone I - Madras City	1
Zone II - All class I and II towns, except Madras City	27
Zone III - All classes III to VI towns	311

All of the towns are subdivided into census enumeration blocks of approximately equal size. A sample size of 240 blocks is described based on the experience of the National Sample Survey and the 1961 census Food Habits Survey. The number of sample blocks allocated to each of the zones was according to the population with a weighting of two for Zone I and II compared to Zone III. The number of enumeration blocks selected in each of the zones is shown in Table 2 of this chapter.

Having selected the villages and urban blocks, the sampling frame was constructed by listing all of the households in each of the villages and urban blocks. Five households were selected randomly from the frame, one for each of the five following characteristics (after finally numbering the households having different characteristics in the sample frame).

1. Pregnant women/households
2. Lactating mothers/households
3. Preschool children/households
4. School age children/households
5. Others

The households were selected as two independent sub-samples and four home visits to the households were planned. For the second and third round, fifty percent of the sample was replaced. In the fourth round, a fresh simple random sample of 2800 households was selected by combining all the rural and urban sample households selected in the first and second rounds.

TABLE 2

SAMPLE DESIGN, URBAN AREA: BLOCK/HOUSEHOLD RELATIONSHIP

Stratum	No. of towns (in State)	Urban population in '000 (1961 census)	Proportional allocation (blocks)	Adjusted sample size Blocks	Households
Sector 1	1	1729	46	60	360
Sector 2	27	2939	79	104	624
Sector 3	311	4323	115	76	456
TOTAL	339	8991	240	240	1440

CHAPTER III

FIELD OPERATIONS IN THE CONSUMER SURVEY

CONTENTS

		Page
A.	The Survey Instrument (Questionnaire) - S. Rajagopalan	10
B.	Data Gathering Problems	
	1. Field Survey Reactions - S. Rajagopalan and T. P. Gopalswamy	25
	2. A Note on Nutrition Related Morbidity Data	32

CHAPTER III

A. THE SURVEY INSTRUMENT (QUESTIONNAIRE)

The questionnaire form displayed on the following pages is the one used for the fourth round (i.e., fourth quarterly survey of approximately 2500 families). It is the product of a continuing evolution and refinement initiated with the pilot test which preceded the first round. There is every expectation that there will be further development and modification in future work on the project, particularly with respect to procedures in the final form detailed in Chapter IV.

Taking the family or household as the primary reference unit for the consumer subsystem, the questionnaire was designed to produce the most detailed description and characterization that could be obtained. Note that the family composition is detailed on the first page, and that persons involved in each of the five parts of the questionnaire were keyed to this list by serial number. The last page, supplementary questions on family planning, was used only in the fourth round.

This instrument was planned to contain the maximum amount of objective, factual information that could be supplied in unequivocal, direct answers elicited by the interviewer with the absolute minimum —preferably none— of probing. It was also designed to eliminate the need for any calculations to be made in the field or office by the interviewer, and to remove any need or propensity for any form of aggregating or pre-processing operations in the field, such things being entirely within the domain of the data processing and analytical operations. Information on knowledge and attitudes have an important place in the description of the system; questions in this area were framed with a view to obtaining simple, direct answers that would not require probing by the interviewer.

CONFIDENTIAL

Name of Investigator :

Code No._____

Date_____

THE INDIAN INSTITUTE OF PUBLIC OPINION

2-A, NATIONAL INSURANCE BUILDING, PARLIAMENT STREET, NEW DELHI-I.

SURVEY OF FOOD HABITS IN TAMILNADU

SCHEDULE-2

PART I : HOUSEHOLD PARTICULARS

Quarter			
1	2	3	(4)

Pr.	La./Inf.	0-6	6-15	N

Rural/Urban

I. Family Characteristics and Consumption :

1. Identification of sample household :

   1. Serial Number.....
   2. Sample Village/Block.....
   3. Village Category/Stratum.....
   4. Sub sample.....
   5. Taluk/Town.....
   6. District.....
   7. Street.....
   8. House Number.....

2. Household Particulars :

   (a) Name of the head of the family interviewed.....

   (b) Person interviewed (relation to the head of the family).....

   (c) Household religion : 1□ Hindu 2□ Christian 3□ Muslim 4□ Others (Specify)

   (d) Caste and subcaste of the household : Caste.....Subcaste.....

   (e) Occupation of the head of the family..... (f) Type of family 1□ Nuclear 2□ Joint.....

   (g) Is the family living in rented or owned house ? 1□Rented ; 2□Owned ; If owned, value of house.....

   (h) Land owned (in acres) :.....; Value in Rs.....

3. (a) Please give the details of your household members and their demographic particulars :

Sl. No.	Relation to head of family	Sex Male-1 Female-2	Age in completed years	Marital Status	Age at Marriage	Age at first child birth	General Education (Standard completed)	Occupation (last month)	No. of children alive	No. of children dead	P/L

3. (b) Has any member of the family left or arrived last month ? 1□ Yes 2□ No. If yes, please indicate the sl. no. of the family member and period of stay in each case.

Arrived		Departed	
Sl. No.	No. of days stayed in	Sl. No.	No. of days stayed away

4. Please give details of purchases/receipts of food and non-food articles in your household.

Description of Item	Local Unit	Standard Unit	Conversion factor Local/Std. Units	Time when last purchased/ obtained	Quantity purchased/ obtained	Where and how purchased/ obtained	Price per local unit in retail market	Total Cost	How long the quantity purchased/ obtained will last lasted	Normal frequency of purchases/ receipts, etc.	Estimated consumption for 30 days
1	2	3	4	5	6	7	8	9	10	11	12
Cereals	X	X	X	X	X	X	X	X	X	X	X
Raw Rice		Kg									
Boiled Rice		Kg									
Wheat		Kg									
Rava		Kg									
Maida		Kg									
Jowar		Kg									
Bajra		Kg									
Maize		Kg									
Ragi		Kg									
Tapioca		Kg									
Sago		Kg									
Others (specify)		Kg									
Pulses	X	X	X	X	X	X	X	X	X	X	X
Red gram		Kg									
Black gram		Kg									
Bengal gram		Kg									
Green gram		Kg									
Horse gram		Kg									
Others (specify)		Kg									
Milk & Milk Products	X	X	X	X	X	X	X	X	X	X	X
Milk		Lit									
Ghee		Kg									
Curd		Lit									
Butter		Kg									
Butter Milk		Lit									

1	2	3	4	5	6	7	8	9	10	11	12
Baby food		Kg									
Edible Oil	X	X	X	X	X	X	X	X	X	X	X
Gingelly oil		Lit									
Ground nut oil		Lit									
Coconut oil		Lit									
Vanaspathy		Kg									
Others (specify)											
Meat & Meat Products											
Mutton		Kg									
Chicken		Kg									
Pork		Kg									
Beef		Kg									
Egg		No									
Fish (fresh)		Kg									
Fish (fried)		Kg									
Vegetables	X	X	X	X	X	X	X	X	X	X	X
Fresh Fruits	X	X	X	X	X	X	X	X	X	X	X
Dry Fruits	X	X	X	X	X	X	X	X	X	X	X

1	2	3	4	5	6	7	8	9	10	11	12
Sugar	X	X	X	X	X	X	X	X	X	X	X
Sugar (crystal)		Kg									
Jaggery		Kg									
Khandasari		Kg									
Others (specify)		Kg									
Salt & Spices	X	X	X	X	X	X	X	X	X	X	X
Salt											
Turmeric											
Black pepper											
Dry chillies											
Green chillies											
Garlic											
Tamarind											
Ginger											
Coriander											
Mustard											
Fenugreek											
Cuminseed											
Cloves											
Cardamom											
Beverages & Refreshments	X	X	X	X	X	X	X	X	X	X	X

4. (a) Also please tell us your expenditure on non-food items during the last month.

Non-food Item:	Expenditure last month (in Rs.)		Non-food Items	Expenditure last month (in Rs.)	
	Cash	Credit		Cash	Credit
Pan Supari (133-135)			Durable goods (190-240)		
Tobacco (137-144)			House Rent		
Intoxicants (146-151)			Education (247-251)		
fuel & light (153-164)			Medicine (252-257)		
Clothing (166-182)			Other Services (259-312)		
Foot wear (185-188)			Non-food (sub-total)		

5. (a) Please tell us between which figures your household monthly Income falls :

	% in kind		% in kind		% in kind		% in kind
0-200 ☐		201-400 ☐		401-800 ☐		801 and above ☐	
0-50 ☐		201-300 ☐		401-600 ☐		801 and above ☐	
51-100 ☐		301-400 ☐		601-800 ☐		Unwilling to disclose ☐	
101-150 ☐						Unable to specify ☐	
151-200 ☐							

(b) Interviewer's opinion of household income given above.

Gross underestimate ☐ Apparently accurate ☐ Gross overestimate. ☐

5. (c) If any member of the household borrowed money last month, please indicate the amount and purpose.

Amount_____ Purpose_____

6. (a) Please indicate the source of water supply for the village and the house.

	In the village	In the house
No supply	☐	☐
Well water	☐	☐
Tap water	☐	☐
Tank or river water	☐	☐

6. (b) i) Is the village electrified ? Yes ☐ No ☐

ii) Is your house electrified ? Yes ☐ No ☐

7. Has some one been ill in your house during the previous 30 days ? [1]☐ Yes [2]☐ No

If Yes, please give the following particulars :

Serial No. of the household member	Nature of illness	Duration of illness	ATTENDANCE		Cost (in Rs.)
			Type	Frequency	

8. Has some death occurred in your house during the one year up to the date of interview? [1]☐ Yes [2]☐ No

If Yes, please give the following particulars :

Sl. No.	Age	Sex	Cause of death

9. (a) Institution/Type/Qualification of medical help :

Institution	Type Allo-1, Ayur-2, Homeo-3	Qualification
Hospital		
Dispensary		
Private Doctor		
Midwife		

9. (b) Please indicate the distance of the nearest medical aid from your house (in kms).....

II. DIETARY HABITS

1. Please tell us if the family is :

1 ☐ Vegetarian 2 ☐ Vegetarian but takes egg. 3 ☐ Non-vegetarian.

If answer is (2) or (3) ask the following details for each member of the family :

Source	Food Item	Serial No. of the family members taking different food items									
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Prepared at home	Egg										
	Fish										
	Meat										
Purchased from outside	Egg										
	Fish										
	Meat										

2. (a) If some members of the family are eating outside, please give the following details :

Serial number of the household member	No. of times food taken outside per week	Place of eating	Reason	Foods consumed	Average Purchase Price per week (in Rs.)

2. (b) Do the women and/or children of the household receive any food at Balawadi, School, Maternal and childcare centres, Urban bread distribution centres or at any other place. Please give details :

Sl. No. of household member	No. of times food taken received per week	Distribution centre	Reason	Foods Received	Cost per week (if any) in Rs.

3. Please tell us the foods prepared and consumed in your house the previous day and by whom in the family were these consumed. If not consumed please say it saved for the next meal or destroyed. Time Codes : **0☐ Pre-prepared 1☐ Morning ☐2 Noon 3☐ Evening 4☐ Night**

Prepared at home :

Time of preparation (code)	Food Item Prepared/ Purchased	Quantity Prepared/ Purchased		Principal Ingredients used						Quantity consumed by each member of the family (in units as in col. 3)										Balance units (4-20)	
				Ingredient			Ingredient				Serial No. of the household member (as in Block 3)										
		Unit	Quantity in units as in (3)	Ingredient	Unit	Quantity	Ingredient	Unit	Quantity	Time of consumption (code)	1	2	3	4	5	6	7	8	Total	Saved for the next meal	Destroyed
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)

Q : 3 Cond.

Purchased from outside

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22

4 Was the previous day : 1 ☐ Festival 2 ☐ Celebrations 3 ☐ Fast 4 ☐ Mourning 5 ☐ Any other event 6 ☐ Normal

5 If you are avoiding any foods in the household, please tell us the reasons for avoiding them :

Food Item	Reasons								
	Heaty	Cold	Gas producing	Religious	Traditional	No liking	Not good for health	Not economical	Others (specify)

6 (i) Do you preserve fruits or vegetables at home or buy from outside ?

0 ☐ **Do not preserve at home or buy from outside**
1 ☐ **Preserve at home** 2 ☐ **Buy from outside** 3 ☐ **Both**

(ii) If preserved at home please state the form in which preserved.

1 ☐ **Jams** 2 ☐ **Pickles** 3 ☐ **Dried forms** 4 ☐ **Others (specify)**

(iii) What are the foods you commonly preserve ?

(iv) Also name the preserved foods purchased from outside.

1 ☐ Jams & Jellies 2 ☐ Pickles 3 ☐ Dehydrated or dried vegetables 4 ☐ Tinned vegetarian 5 ☐ Tinned non-vegetarian 6 ☐ Tinned milk 7 ☐ Others (specify)

7 Have you heard the words 'Vitamin,' 'Minerals,' 'Calorie,' 'Proteins,' 'Calcium,' and 'Iron' ? If so please name the source of information and the foods in which they are present

Nutrient	Source of information	Foods in which present
Vitamins		
Minerals		
Calories		

Nutrient	Source of information	Foods in which present
Protein		
Calcium		
Iron		

PART II

PREGNANT WOMEN

★★

Serial Number of household member.....

1. Please tell us the following particulars :

Age at marriage	Age at first child birth	Doing household work			If employed			
		Yes, with help	Yes, without help	No	Occupation	Time of work (hrs. per day)	Place of work from home in kms.)	Whether on leave (Yes, No or N/A

2. Please tell us how many children you have ?

Living : Total number... .....

S. No.	Age	Sex	Remarks

Died : Total number..... .....

S No.	Age	Sex	Remarks

3. Please also tell us the No. (if any) of premature births... .....Still births.....Miscarriages.....

4. Please tell us the foods that you are avoiding in your diet during pregnancy :

Particular foods avoided		
Items	Reasons	Source of advice

5. Please also tell us the special foods that you are at present including in your diet during pregnancy which you normally either do not take or take in lesser quantities.

Particular foods included						
Items	Unit	Quantity per week	Duration in weeks	Reasons	Source of advice	Cost per week in Rs.

6. Please tell us the foods (if any) that you wish to take, since you consider them good for your health during pregnancy but for some reason you are not taking.

Name of the food item	Reasons for not taking				
	Expensive	Prohibited by Doctors	Not approved by elders	Not available	Other (please specify)

7. Please name herbs, medicines or medicinal supplements (if any) that you are at present including in your diet

Item	Unit	Quantity per week	Duration in weeks	Reasons	Source of advice	Cost per week (in Rs.)

8. Please tell us the month of pregnancy.....

9. Please tell us how many times you visited the doctor last month ?.....

10. General health of the pregnant mother (Investigator to record by observation or enquiry)

1☐ Swelling on the body 2☐ Vericose vein 3☐ Hypertension 4☐ Anaemia

5☐ Others.....(Please specify)

11. General Appearance : 1☐ Good 2☐ Average 3☐ Poor

PART III

★★★

INFANTS AND LACTATING MOTHERS

I. Infants :

(i) Serial number of the household member..... (ii) Serial number of informant.....

(iii) Age (in months).....(iv) Sex 1☐ Male 2☐ Female (v) Birth weight (in kgs.).....

1. **What foods did you give your child during first three days after birth ?**

....

2. **At present how many times you breast-feed your child daily ?** .....

3. Please give details of supplementary foods (if any) that you are giving your child at present.

Food Item	Started feeding at the age of (in months)	Frequency of feeding per day	Form in which given	Quantity during the previous week given		Cost per week (in Rs.)
				Unit	Quantity	

4. Please name the foods that you wish to give your child since you consider them good for his/her health but for some reason not giving at present ?

Food item	Reasons for not giving at present			
	Expensive	Prohibited by doctors	Not approved by elders	Others (please specify)

5. Please name any medicines, tonics or medicinal herbs that you are giving your child ?

Item	Amount per week		Duration in weeks	Cost per week (in Rs.)
	Unit	Quantity		

6. Are there any obvious nutritional deficiency symptoms ? (to be recorded by observation or enquiry)
1 ☐ Anaemia 2☐ Pot belly 3☐ Swollen legs 4☐ General weakness 5☐ Any other illness, defect or deformity (Please specify).....

7. General appearance of the infant 1☐ Good 2☐ Average 3☐ Poor.

II. Lactating Mothers **Serial number of the household member.....**

1 Please tell us the following particulars :

Age at marriage	Age at first child birth	Doing household work			If employed					
		Yes, with help	Yes, without help	No	Occupation	Time of work (hrs. per day)	No. of days per week	Place of work from home (in kms.)	Relation (to the child) of person looking after the child when away	Whether on leave (Yes, No or N/A)

2. Please tell us how many children you have ?

Living : Total number.....

Sl. No.	Age	Sex	Remarks

Died : Total number.....

Sl. No.	Age	Sex	Reasons

3. Please also tell us the number (if any) of : Premature births .....Still births..... Miscarriages.....
4. Please tell us what stage of lactation you are (in months).....
5. Also foods you are specially avoiding.

Particular foods avoided		
Item	Reasons	Source of advice

6. Please name the foods that you are specially including during lactation.

Particular foods included						
Item	Unit	Quantity per week	Duration in weeks	Reasons	Source of advice	Cost per week (in Rs.)

7. Please name the foods that you wish to take since you consider them good for health during lactation period but are unable to take for some reason.

Food Item	Reasons for not taking			
	Expensive	Prohibited by doctors	Not approved by elders	Others (please specify)

8. Please name any medicines, medicinal supplements or herbs that you are currently taking?

Item	Unit	Quantity per week	Duration in weeks	Reasons	Source of advice	Cost per week (in Rs.)

9. General health of lactating mother (to be observed and recorded)

1☐ Good 2☐ Average 3☐ Poor

PART IV WEANED CHILDREN 0 - 6 YEARS

★★★★

Sl. No. of household member..... Age..... Sex..... ..... Sl. No. of the informant..... .....

1. Please tell us at what age (name the child).....was weaned (Record age in months).....
2. Please tell us any foods you gave him/her when he she was weaned :

Item	Amount per week		Duration in weeks	Cost per week (in Rs.)
	Unit	Quantity		

3. Please name the foods that you do not wish to give your child at present :

Foods	Reasons	Source of advice

4. Are there any foods that you wish to give your child since you consider them good for his/her health but are, for some reason, not giving at present ?

Food item	Reasons for not giving				
	Expensive	Prohibited by doctors	Not approved by elders	Not available	Others (please specify)

5. Please name any herbs, medicines or medicinal supplements that you are at present giving your child to improve his/her dietary in-take

Particular foods included						
Item	Unit	Quantity per week	Duration in weeks	Reasons	Source of advice	Cost per week (in Rs.)

Investigator : If the mother of the child is employed ask Question 6 below. Otherwise proceed to Question 7,

6. a) Please tell us for how many hours the child is alone when mother goes away for work..... .....
   b) Who looks after the child when the mother is away for work.. .....
7. Are there any obvious nutritional deficiency symptoms ? (Investigator to record by observation or enquiry)

1☐ Angular stomatitis
2☐ Bitot spot
3☐ Rough skin
4☐ Rickets
5☐ Bleeding gums
6☐ Dental carries
7☐ Stunted growth
8☐ Under weig...
9☐ Over weight
10☐ Oedema
11☐ Others (please specify).....

PART V CHILDREN BETWEEN 6—15 YEARS

★★★★★

Sl. No. of the household member..... Age..... Sex..... Sl. No. of informant..... .. ..

1. Please name the food items that you do not wish to give to your child at present.

Food item	Reasons for not giving	Source of advice

2. Please name any herbs, medicines, or medicinal supplements that you are at present giving your child to improve his/her dietary in-take.

Item	Unit	Quantity per week	Duration in weeks	Reasons	Source of advice	Cost per week(in Rs.)

3. Please name the foods that you wish to give to your child since you consider them good for his/her health but are, for some reason, not giving at present.

Food item	Reasons for not giving			
	Expensive	Prohibited by doctors	Not approved by elders	Others (please specify)

4. What was the work status of your child last month ?
   1☐ Going to school or on leave from school 2☐ Going out for work
   3☐ Working at home 4☐ Neither working nor going to school

5. Does your child take snacks/meals regularly at the place of work/study ?
   If ' Yes ' let us know the source and cost ?

Source	Cost per week (in Rs.)	If not, please give reasons
1☐ Prepared at home	.....	1☐ Considered not necessary
2☐ Purchased from outside	.....	2☐ Unable to afford
3☐ From school	.....	3☐ Others (please specify)
4☐ Others (please specify)	.....	.....
.....	.....	.....

6. Has the child failed in any class ? If so give details.
   1) Class.....Age..... 2) Class.....Age..... 3) Class.....Age.....

7. If discontinued studies, please give reasons :
   1☐ Poor health 2☐ Continuous illness 3☐ Illness in the family 4☐ Failure in class
   5☐ Others (please specify)

8. Are there any obvious nutritional deficiency symptoms ? (Investigator to record by observation or enquiry)
   1☐ Angular stomatitis 4☐ Rickets 7☐ Stunted growth 10☐ Oedema
   2☐ Bitot spot 5☐ Bleeding gums 8☐ Underweight 11☐ Others (please specify)
   3☐ Rough Skin 6☐ Dental carries 9☐ Overweight .....

SUPPLEMENTARY QUESTIONS

1. (a) What is your opinion about the family planning programme ?

☐ Good ☐ Bad ☐ Indifferent

(b) [If response is good or indifferent] Are you in favour of the programme ?

☐ Yes ☐ No.

(c) If yes, do you adopt any technique ?

☐ Yes ☐ No.

(d) If yes, what technique you adopt ?

(e) In your opinion, how many children should you have ?

2. Please indicate the contribution of the earning members of the household to the total income (given in Q 5.)

S. No. of Member (Same as in Q 3)	Percentage Contribution

Name of Supervisor.....

Whether checked ☐ Yes ☐ No. Date.....

Whether back checked ☐ Yes ☐ No. Date.....

CHAPTER III

B. DATA GATHERING PROBLEMS: FIELD SURVEY REACTIONS

Introduction. Not the least important part of a survey is the actual interviewing of representatives of the selected households by field interviewers. Biased or tactless interviews, badly phrased or slanted questions, inaccurate reporting or language difficulties may defeat the purpose of the survey, no matter how good the design.

Surveys in India have almost become a way of life in the sense that for the last two decades, continuous surveys have been conducted. For example, the National Sample Survey undertakes a yearly country-wide survey of the socioeconomic conditions of the population. Thus, problems encountered in past surveys in India become pertinent to any new survey. The common field problem of many surveys in India is the identification of sampling units and the construction of a sampling frame. Lists of all households or individuals are generally not available. This makes the selection and identification of sampling units very difficult. The problems of identifying households or individuals in rural India is compounded by the non-existence of household numbers, street names or any other means of identification. It is only after inquiring with village panchayats or other village authorities as to the location of specific households (or individuals) that one is able to pin down the sampling unit. Such a procedure reduces the efficiency of the field work, especially in those surveys such as the TN Study, where the time frame represents a significant constraint.

Field interviewers in India have also to wrestle with infrastructure problems. Many villages or taluks are remote and reachable only via poor roads. If the survey is carried out during a rainy season (which in India is almost inevitable), some of the villages are inaccessible. Besides infrastructure and location difficulties, the variable cultures and languages add to field problems. In any one Indian state there are several dialects of the local language, whether it is Hindi, Tamil or Malayalam. In constructing a questionnaire, one is normally concerned with the phraseology of the various questions presented. But it is impossible to construct a survey instrument in such a way as to cater to all dialect variations. It is the interviewer's responsibility to interpret questions in a manner which will satisfy all dialects in a given state. Obviously, such a requirement suffers severe limitations. They range from prolonged and costly training

of interviewers, to the fundamental intelligence level of the field worker himself. Here again, variations exist and there is no statistical methodology to correct for them.

Despite all these problems, however, surveys are conducted in India, and with a reasonable degree of success. After all, research and quantitive analysis do not stipulate ultimate accuracy. Knowing the quality of the data is equally important as is the consistency and sufficiency of information. This survey has encountered all of these problems.

The major Food Habits Survey was contracted to an Indian consulting firm which has had previous survey experience. The 20 or so interviewers hired by the subcontractor were trained and retrained with participation on our part. Prior to the field team's departure, we, as well as the subcontractor, were confident (within realistic margins) that the team was adequately prepared to undertake the task. Consistent and frequent monitoring of the field team was undertaken by the Madras staff throughout each round of the survey. This was the only innovation in the TN survey technique as compared to previous surveys carried out by other institutions, including USAID.

The continuous monitoring paid off, when it proved possible to identify and subsequently rectify inconsistencies from one interviewer to another, dishonesty by some (resulting in dismissals); and perhaps more important, to obtain a real "feel" for the information being collected in the field. There is no substitute for having a firsthand acquaintance with the raw data collected which is subsequently to be used in analysis. It facilitates the judgements that researchers must make in the course of analysis.

Revisit Problems - As noted, the survey design called for four quarter surveys of the same households. The intention was to observe seasonal variations in household consumption patterns. Interviewers, during the second and subsequent rounds, encountered various difficulties in re-interviewing the same households. Many of the respondents were reluctant to be interviewed in the first place and their reluctance intensified with subsequent visits. This was not totally unexpected, since the survey instrument required over two hours of interviewing time.

Few families, especially those in the low income groups, can spare this amount of time on a continuous basis for answering what appeared to them, at times, to be "silly" questions. This occurred despite initial good rapport developed by the field workers during the first quarter survey.

Non-Response Problems. Apart from the reluctant to respond, there was encountered in a number of villages and urban blocks, what might be classified "not at home" respondents. This did not influence the sample design significantly since allowances were made for non-response at the rate of 20%. Nevertheless, in statistical terms, the non-response rate introduces a bias into the sample which has to be contended with, i.e., although the general confidence-level of the sample may not have been reduced due to size, bias has crept in and appropriate methodology for dealing with this bias has to be applied.

A case in point: in one of the rural villages, Villipuram in South Arcot District, of five households studied during the second quarter, only one could be studied during the third quarter because two households were found locked and the other two refused entry. Instructions were given to the field supervisor to revisit such households and complete the questionnaire, if at all possible. In some non-response cases, our persistence paid off. In others, it did not.

Inaccuracies in Responses. Respondents who cooperated did so willingly. However, many answers involved inaccurate statements. The reasons varied from respondent to respondent. One common thread was the sensitivity to questions on family size and income. In the first quarter, the survey instrument contained a question on family planning. As a result many respondents mistook our investigators to be from the Family Planning Department. They were reluctant to report the correct number of children. For example, in Kuzhathooval village in Ramnadapuram District, a household reported, during the second quarter survey, two children and when this household was revisited in the third quarter survey, the family size grew to five children. Having identified this discrepancy, our staff visited the household and established that, in fact, eight children were to be found! We proceeded to ask why the respondent didn't give the correct number of the children in the first instance. The reluctance proved to be due to the suspicion that the investigator was from the Family Planning Department

and that they (the household) might be identified and later asked to adopt family planning measures. In order to eliminate this difficulty, we removed the question from the first quarter survey, and included it in the last (fourth) round.

Although our questionnaire provided ranges of income, thus eliminating the need for reporting a specific income figure, it still remained an extremely sensitive question. Many respondents were unhappy and unwilling to disclose their correct income because of fears that the information would be used for taxation purposes. It was only after persuasive discussions that household members were willing to indicate a range of income.

The reluctance of respondents to disclose their income, is not a new phenomenon in surveys in India. In fact, even today, there are no reliable income statistics. The National Sample Survey abandoned the income approach in their surveys and concentrated on expenditures. They impute expenditures as disposable income. The main thrust of our questionnaire was also on expenditure rather than income. This included expenditure on food and non-food items. The question on income was included primarily as an indicator of internal consistency of the instrument as well as an estimate of the respondent's honesty and reliability. Income figures did generally check the expenditure data to establish acceptable credibility.

Field Monitoring and Supervision Problems. The Food Habits Survey subcontractor's field team consisted of 20 interviewers. This group was later divided into four teams, each headed by one supervisor. A 4:1 ratio in surveys of this kind is very favorable. The National Sample Survey uses a 6:1 ratio, i.e., one supervisor to every six interviewers. Each supervisor provided the subcontractor's headquarters with their respective schedule indicating the dates and location of field work. Although the interview programs were known to all concerned, it was often difficult to locate the team at the exact village on the appointed date. At first this appeared to be a very disconcerting aspect of the survey work. Upon further investigation, it was established that the field team frequently had to spend more time in a given urban block or village because of situations like "not at home" or "not free to answer on the date of the visit." Difficulties in reaching the villages because of bad weather, investigators falling ill or having to leave due to unforeseen circumstances, further complicated the field work and organization.

Thus, prescheduling of interviews, in a sense, became academic. The program and schedule had to be flexible and change as conditions warranted. Having learned this, it was insisted that field supervisors report to headquarters any revision in their planned programs. At times, the supervisor had difficulties in locating his interviewers on short notice. This was not necessarily detrimental. Interviewers exercised initiative and proceeded with their work, later returning to those households not covered.

Field problems were more pronounced in the first quarter than during the second and third quarters. As experience was gained, field plans were designed in a way as to provide greater flexibility. Communications from the field worker to the field supervisor and back to headquarters were also much more efficient and timely.

Elements of Caution When Interpreting Certain Responses. Having reviewed a substantial number of field questionnaires, it became apparent that careful analysis and judgment would have to be exercised when examining specific answers received from the survey. Such synthesis is justified on the basis of what was learned from observations in the field and from our discussions with interviewers. Critical items requiring such attention follow:

(a) The person interviewed in the same household in each quarter may be different. Therefore inconsistencies may result with regard to age of household members, income and consumption. The same person was not interviewed every time for various reasons, among them, leaving the household, taking a job in a different location, or simply disappearing.

(b) The survey instrument was printed in English and occasionally investigators found it difficult to interpret Tamil equivalents in the appropriate dialect when translating specific questions. This undoubtedly created some inaccuracies in responses which were very difficult to measure and identify. This is an important reason for storing raw data. Debugging procedures described later in this volume are pertinent.

(c) Among the questions framed was the value of house and land. Generally, the investigator was able to obtain an estimate for the house value. For the value of land, he had to rely completely on the respondent. Here clearly, there can be a wide variation of estimates. Regional consistency, however, can be tested by appropriate programming.

(d) In a multi-family dwelling consisting of many households, it was particularly difficult for both the respondent and the investigator to obtain a good estimate of the value of the portion of the structure occupied by the surveyed household. Arbitrary division had to be resorted to in such cases.

(e) With respect to intrafamily food distribution and consumption pattern, housewives generally did not know what their husbands or other members of the family were eating outside of the house. We encountered difficulties with the respondent's ability to conceptualize quantities. For example, how much 200 milliliters or a tablespoon or a teaspoon measured. It was only with repeated probing and physical demonstrations that estimates were obtained. Invariably the interviewer had to measure the quantity of food indicated by the respondent with one of the three vessels he carried with him. Respondents in general were reluctant to tell the exact quantities of food consumed because of fear of the "evil eye."

(f) It is practically impossible to critically examine responses on ingredients. Hardly any of the respondents measured or weighed the ingredients before preparing a meal. Answers in this part of the questionnaire are thus subject to wide variations. However, this procedure is not unusual in household operations as any person practiced in food preparation knows.

(g) Most respondents in rural Tamil Nadu were uncertain about weight of infants at birth since most births take place outside hospitals. The cultural practices, however, of reducing the food intake of the mother in the last trimester of pregnancy so as to reduce the size of the infant was evident.

(h) In the case of households where cultivation or agricultural labor was the main occupation, it was difficult to assess the proportion of their cash and produce incomes. The investigator had to use his own judgement by asking the quantities of various items received in kind and imputing values to them. He subsequently calculated the appropriate proportion.

(i) The responses to the question "what foods do you wish to give to your children and your reasons for not giving them" were very vague. Not

infrequently, the interviewer got an answer by suggesting certain food items. Thus the interviewer in effect provided the answer. It should be understood that the respondents were unused to thinking about food outside of their daily routine. New foods particularly were a strange concept, hence thinking about them in relationship to children was a strange exercise.

(j) Demographic events such as births, deaths, cause of death, still births or age, suffer from incompleteness and gaps in reporting. The chief reasons for this are variously, no record keeping, ignorance and superstitious fear of reciting such intimate details, (the evil eye). See in this volume the "Note on Nutrition Related Morbidity".

(k) In some instances responses to the "most recent purchases of food articles" did not cover all market channels through which all of the food articles were obtained and consumed during a month. Best estimates were made in some cases by reference to the general data on market structure obtained from the ORG Food Processing Survey (see Commentary Vol. II E).

Summary The field and operational problems highlighted in this paper are not uncommon in any sense of the word. On the contrary, when discussing survey difficulties with representatives of other Indian institutions who perform surveys in India, they were surprised at the low magnitude of our problems. This, of course, was not necessarily accepted as a conciliatory note on the part of this group. This group strived to accomplish the best possible results, under Indian survey conditions.

In retrospect (as always), a better survey could have performed. It would have required more intensive training of interviewers, more interviewers, a shorter questionnaire and perhaps a more efficient one, more frequent and intensive monitoring by the staff (requiring a greater number of TN project staff) and finally more time for adequate preparation and execution. In the absence of such "luxuries", the field survey work is considered satisfactory and sufficient to the objectives of the project — particularly as they relate to food expenditures, identity and consumption matters in Tamil Nadu.

CHAPTER III

B.2 NUTRITION AS A FUNCTION OF PUBLIC HEALTH: A NOTE ON NUTRITION RELATED MORBIDITY IN TAMIL NADU

In the course of the Tamil Nadu Nutrition Project, it was proposed that data be collected relative to the health and hygiene of the people within the State. During the analytical phase of the project it was intended that these data be correlated with the results from the Consumer subsystem such that nutritional status could be integrated with the health and hygiene of the population.

Data collection was initiated by submitting to over 600 medical installations with beds throughout the State a questionnaire designed to obtain information on nutrition related morbidity. Such information was to include relationships of nutrition based diseases to age, sex, location and type of medical installation, etc. Responses to the questionnaire were slow in coming in, despite vigorous assistance by the Tamil Nadu State Health Secretary and the Directorate of Health Services and Family Planning. In the end only 132 (20.7%) of 638 units responded.

There are explanations for such a low return —including the main one of non-availability of data— but even within the returns the only consistent relations were age and disease. With a small sample and two variables, a pattern of sorts might have been arrived at. However, records from most installations used the age of 12 as a dividing line between children and adults with little attention paid to sex, single or repeated admission, etc. Further, as in so many countries, little or no subjective diagnostic tests were used to verify or confirm the disease designated. The prevalent association of doshams (sins) and nutrition related diseases, as noted in volume IIB, is another reason for mistrusting the clinical data.

These problems are not those of India alone; all one has to do within the country is to spend a few days at a few medical installations to appreciate the problems involved. The number of patients per qualified physician, for instance, is overwhelmingly high; and equipment for carrying out proper diagnostic work is frequently not available. Moreover, even in situations where equipment was found, there was a dearth of qualified paramedical technicians and maintenance personnel.

In the face of a minimum of data, it was impossible to gain a realistic picture of nutrition based diseases in the State. The available data were, however, converted by mathematical manipulations to give a projected profile for Tamil Nadu. However, because of its theoretical character, the profile is not included in the data bank.

Among the missing elements constituting medical data are medical records and statistics of an adequate, uniform and statewide understandable nature. This phase of the morbidity situation within Tamil Nadu has been examined and potential corrective measures have been recommended by S. Rajagopalan. Mr. Rajagopalan was Senior Statistician with the Tamil Nadu Nutrition Project and has been designated as the director of the group which will continue the project throughout India.

As part of the data gathering exercise, some of the project members visited several Primary Health Centers within the State. Many medical officers attached to the Centers were interviewed regarding the status of operations at the sites. The findings verified those contained in the report of the Tamil Nadu Planning Commission Task Force on Health, Family Planning, Nutrition and Sanitation. Thus the team recordings are not repeated here but they are available at the Contractor's office.

The absence of information bearing on nutritionally related diseases as a natural consequence did not allow the execution of what was to be a pertinent part of the Consumer subsystem study; namely, the correlation of medical data with other characteristics of families participating in the various consumer surveys. It may well take much too long to set up functional medical and evaluation systems within the whole State. Still, such could and should be done on a small but significantly large enough sector of people who will become involved in subsequent surveys as the projected nutrition program expands under local leadership.

CHAPTER IV

DATA TECHNOLOGY

CONTENTS

	Page No.
INTRODUCTION	35
DATA ENTRY	36
EDITING	49
RESTRUCTURING OF THE DATA BASE	51
PRELIMINARY PROCESSING	53
COORDINATION AND INTEGRATION OF THE SYSTEM	56
ESTIMATE OF PROGRAMMING REQUIREMENTS	63
GENERALIZED DATA BASES	67

INTRODUCTION

This section covers the mechanics of converting data as gathered by an interviewer to data which can be statistically analyzed in a meaningful way. The section is aptly titled "Data Technology" since a technology of data handling has developed in recent years complete with sophisticated techniques and tailormade hardware. The recent advances in this technology have opened entirely new possibilities for detailed data handling and storage, and, therefore, more thorough data analysis. The development of this advanced Data Technology is especially relevant to the analysis of complex, multivariable systems, such as food and nutrition systems which require enormous amounts of detailed data for legitimate analysis. Many of the methods discussed here are either impossible or economically unfeasible on the computer equipment of the past. For this reason, special attention is paid to pointing out why many current data methods were originally adopted and why most of these methods and the limitations that they place on data analysis are now out of place and should be discarded. With these new developments and their many inherent possibilities for improving or destroying analytical efforts, the mechanics of data conversion must be understood by every project analyst. He must be qualified to make a number of decisions concerning data conversion and be fully aware that his decisions can significantly effect the cost of the project and ultimately the type, number and precision of conclusions that can be drawn from the data gathered.

With the foregoing in mind, this section will review the basic concepts of data handling for computer operations and the various stages that data must pass through before it is analyzed. In addition to reviewing the mechanics of data conversion, this section presents an outline for processing survey data on a continuous basis. The processing system described is a highly cost/effective one which reassigns the traditional division of labor among interviewer, coder, keypuncher, and computer to the division of labor most compatible with the new technology and the new economics. The processing system presented here is a combination of ideas of the computer subcontractor* and the experience gained in the Tamil Nadu project.

A guiding theme throughout this section is the belief that the data gathered by a large-scale survey should be stored in a data "bank" in unaggregated form. For the purpose of analysis, a particular subset of the data in the data "bank" can be identified, assembled to the desired level of aggregation

*Symbionics, Inc.

and copies extracted, leaving the "bank" intact and ready for additional data requests. The extracted subset of the data bank can then be analyzed efficiently and, therefore, at low cost by the computer. At the close of the initial project, the data bank can pass into the public domain. It can then be utilized and added to by other workers who may be continuing the initial analysis of the data or working from an entirely different perspective. The data bank would, of course, be available for the detailed planning and evaluation of follow-up surveys. All of this can be accomplished at little or no additional expense to the original project. The key to the success of the data bank approach is that (1) the data is unaggregated, and (2) that an easy to use, well documented program is provided which can retrieve and aggregate data as required.

The various steps and topics discussed below are: Data Entry, Editing, Restructuring the Data Base, Preliminary Processing, Coordination and Integration of the System, Estimate of Programming Requirements and Generalized Data Bases.

DATA ENTRY

Coding design - The decision to analyze data with a computer is made because the volume of data is too great and the calculations too complex to perform accurately by hand within an acceptable period of time and at reasonable cost. Working with a computer should not be thought of as being different in any substantial way from directing humans to perform the same tasks except that the computer will work much faster, and consequently can perform more work if properly instructed and provided with the necessary data. It follows, then, that the first principle in coding design is to see that the cards, which represent the computer version of the data, are an exact duplicate of the survey questionnaire down to the name of the interviewer, the name and address of the person interviewed and any remarks that the interviewer has made concerning the interview. It must be possible to reproduce the questionnaire in all of its detail from the computer cards alone. Any bit of data which is not transferred to the cards, and therefore to the computer, cannot be analyzed.

The basic decision to use a computer was made because the data mass was too voluminous to handle manually; it, therefore, is a mistake to assume that any data not coded can later be analyzed by hand. Data not coded is data lost. It is a waste of interviewer time as well as that of the respondent to collect data that will not be coded. Time wasted by an interviewer reduces the number of interviews conducted per day and, therefore, increases the cost of the project. This is a very serious problem since the cost of interviewing (or "data gathering") is usually a large item in the budget of the project. Information not coded from the questionnaire may also be vital to a full analysis and, therefore, can conceivably limit or distort the conclusions to be drawn from the survey. Errors of outright omission of data to be transferred from a questionnaire to the computer via cards or other means (see below) are to be avoided at all costs.

False economies - While errors of omission are the most obvious type of coding error, there are several other types of errors which also can both add substantially to the costs of the project and cripple the analytical work. It should be noted that the errors of omission of data and other coding errors are usually made in some ill-conceived attempt to "cut corners" and reduce minor costs, such as the number of cards to be punched. At all times, decisions concerning coding and every other facet of the project must be made in light of the overall impact of the decision on the project and its cost effectiveness. In addition to the omission of data, there are other common errors of false economizing in the use of card columns and the aggregation of data.

In general, little or no attempt should be made to economize on the number of card columns used by a particular data item or the number of cards used per questionnaire. When coding numeric data such as the cost of a food purchase, one should allow enough columns for the largest expected purchase (plus columns for the decimal precision required)* plus an additional column. The number of cards required by the questionnaire should be determined by the sum of columns required to code each data item adequately. Cost of cards and keypunching is relatively small when compared to cost of data gathering and the value of the analytical work. Economizing in this area is counterproductive in most cases.

---

* It is not necessary to actually code the decimal place, though it is necessary that all values be coded with the same precision if no decimal is coded.

Frequently, attempts are made to conserve columns by coding data in special schemes, as in Figure 1. While the code format designer may have had a particular chart in mind that would use this scheme, and assumed he would be saving computer time and card columns by coding these data items together, he has generated more problems than he has solved. Such a scheme requires that a coder interpret the two answers, compare them to a scale and produce a third number. This takes time and there is considerable room for error. The alternative would be to give the questionnaire directly to a keypuncher and have him punch the two original values without interpretation. This latter approach is faster and more economical since it eliminates the need for the coder. In addition to the coding problem, if in the course of the analytical work a different grouping of these two data items is required, a program would have to be written to decode the number and then to reassemble it in the new aggregate. If the two values are coded separately, they can quickly and easily be combined in any way, without complication, by the programmer.

The questionnaires should be numbered sequentially as they are received at the keypunch facility or prenumbered when printed. No attempt should be made to code interviewer or sample strata information in this number. The questionnaire number should use as few columns as possible since it is repeated on each card punched. The use of hexidecimal digits* to conserve the number of columns required for these housekeeping identifiers is advisable. Information concerning the interviewer and sample strata, etc., can be entered as separate data items on one of the input cards.

When assigning codes for all questions, a code should always be provided for the case in which there was no response by the respondent or if the question was not applicable. If the card columns are left blank it is not possible to ascertain for sure if the coder, keypuncher, interviewer or respondent has or has not made an error of some kind. However, the use of these null codes should be tightly controlled. Interviewers should not be satisfied with checking "no response" instead of probing for the answer to a sensitive question. Likewise, coders should not code illegible responses as "no response", the questionnaire should be returned to the interviewer for clarification in such cases. A null code means a null response, and no additional meanings should become attached.

Coding schemes - The coding scheme of Figure 1 is a very simple one. However, these schemes can reach great heights of complexity and become destructive. For example, data for each individual in a household is gathered by the interviewer, but only the number of "adults" or "preschool" children is coded. Specific data relating to each individual are

* Numbers to the base 16.

Figure 1

CODING SCHEMES

SPECIAL SCHEME

(Not Recommended)

Questionnaire:

Q. Is Family Living in Rented or Owned House?

1. ☐ Rented 2. ☑ Owned

Q. Family Type

1. ☐ Nuclear 2. ☑ Joint

Coding Instructions:

Rented or owned house and type of family:

Rented - Nuclear	-1
Rented - Joint	-2
Owned - Nuclear	-3
Owned - Joint	-4
Not mentioned - Nuclear	-5
Not mentioned - Joint	-6
Not mentioned - Rented	-7
Not mentioned - Owned	-8

Code "4" in Column 10

DIRECT CODING

(Recommended)

Questionnaire:

Q. Is Family Living in Rented or Owned House?

2. ☐ Rented 3. ☑ Owned 4. ☐ No response

Q. Family Type

2. ☐ Nuclear 3. ☑ Joint 4. ☐ No response

Coding Instructions:

Rented or Owned House	Code "3" in Column 10
Type of Family	Code "3" in Column 11

totally lost. This can become a serious problem since the analysis is limited from the start to the groups defined by the designer of the questionnaire. Besides not knowing the exact composition of the various predefined groups, the groupings may actually obscure the phenomenon being investigated. The use of these schemes may have been justifiable when computers were slow and cumbersome and the intent of these schemes was to minimize the amount of computer processing. However, modern large scale computers can move and recombine large amounts of data very efficiently so there is no need to try to encode data into special schemes to reduce the work of the computer. The computer can combine the original data into any elaborate scheme desired more quickly, accurately, and for less money than the lowest paid coder. Furthermore, survey data currently being collected should be preserved in their original form to allow workers on other projects to tap this store of information in their own ways and for their own purposes. To insure the maximum utility of data, they should be gathered and stored in their most simple, unaggregated form. If data are not stored in simplest form, wholesale recoding and repunching or even additional surveys may be required. Such expense can be avoided. Encoded responses to questions with a fixed number of answers into special space-saving schemes ranges from being foolish, as in Figure 1, to being destructive. In all cases, the responses to questions with a fixed number of answers should be coded separately using the codes of the questionnaire. The coding of questions which can have any of a large number of responses is handled differently and is covered below.

Glossaries and categories - A glossary is a list of terms, their definitions, and the symbols used to represent them. Glossaries are necessary to code the answers to questions which have a large number of possible answers, such as the occupation or village of an individual or the foods he consumed for breakfast. It is common to group glossary terms into categories and have the interviewer record the category he believes that a given item belongs in, or for the interviewer to record the exact response and have a coder assign the category. This method has many shortcomings. First, the various categories must be defined before the coding work is begun. This requires a great deal of work by a member of the project staff who must anticipate every possible response. More critical still is the fact that the categories must be selected in some associated way that will be useful to the work of the project. In nutrition work especially, a single set of groupings of food items is not possible. By what attribute are the foods to be grouped: protein content? calorie content? vitamin content? price? mineral content? processed or non-processed? "hot" or "cold"? etc. Few of these attributes correlate well with more than one or two

other attributes in the list, making groups defined by one attribute practically meaningless when another attribute is examined. Trying to group several attributes together yields groups of virtually no utility for the complex type of analysis required to understand nutrition problems. The logical impossibility of creating a single set of food groups of even marginal utility should be obvious even to the lay reader. Categories can be assigned but they will only limit the precision with which the analytical work can be done, since at best only an average value can be assigned to each category.

Other drawbacks to using categories are that either the interviewers or the coders must learn the definition and dimensions of each category and make a judgment as to which category an item belongs. Not only is it unlikely that a group of individuals will have the same understanding of the definitions of categories, but individual judgment will vary and be affected by the degree of boredom and fatigue of the individual. Furthermore, when new items appear which were not originally anticipated, their placement in the categories is likely to be erratic. The coders and interviewers are not specialists and difficult judgments should not be expected of them. This will result in miscodings which will not be detected by normal editing checks and will eventually result in incorrect statistics. A certain amount of time is required for each judgment,adding extra costs to the project budget.

The simple solution to this problem is to code each item and allow the computer and a specialist to group the items as is required for a given problem during the analytical phase of the project. This allows for rapid coding since the name need only be located in a list and no judgment is required; there is also the added advantage of allowing the item to be grouped and regrouped as the need arises during the analytical stage. Also, it is not necessary for a member of the staff to attempt to define various categories before the survey is conducted, when the nature of the data to be grouped is not always known.

A few words of caution must be recorded concerning the method of building and maintaining glossaries. A glossary should not, with few exceptions, be "preloaded" with terms that are expected to be encountered in the survey. Rather, as each questionnaire is coded, new terms should be added to a master glossary and an appropriate code assigned. At the end of each day, each coder's glossary should be updated with the new terms encountered during the day. The glossary items and their definitions and codes should

be kept on punched cards so that updated glossaries can be quickly printed by the computer. The initial set of terms in the glossary can come from the pretest of the questionnaire and the test coding of those questionnaires by the coding and card punching supervisors. With a glossary assembled in this manner, glossary terms will be in the language of the respondents. The coder has then only to find the exact term in the glossary to assign the proper code. The glossaries can become very large, expecially in the case of foods and occupations, but this is not cause for concern. If the terms have been properly defined (see below), a specialist can identify synonyms and compress the list during the analytical phase. It is very advisable to assign the task of identifying synonyms to a specialist rather than to other members (interviewers, coders) of the project who are much less likely to have an objective understanding of the situation, and of the needs of the project as a whole. In some cases, the use of various language synonyms, for example, by different groups of individuals reveals important group attitudes which would not otherwise be detected if the synonyms for a given food, or occupation, etc., were grouped before the coding and punching of the computer cards.

An exception to the idea that glossaries should be built up as the data are coded, not be preloaded with standard terms, is in defining units of measure. The definition of common units of measure is obviously important in a food survey. If the quantities of food purchased and eaten are not recorded accurately, any calculation on those quantities will give misleading results. There are two approaches to insure that this type of error does not arise. One is for the interviewer to convert the quantities reported to him by the respondent to standard units such as cups, litres (liters), and kilograms, and to record the number of standard units. This requires that the interviewer carry implements to measure the standard units and that he perform a calculation on every quantity reported to him in nonstandard units. The other method is for the interviewer to record accurately the number of nonstandard units per appropriate standard unit in terms of weight (grams or kilograms) per unit of volume (cup or liter) and record the quantities in the nonstandard units. This relieves the interviewer of several calculations and the errors which will inevitably occur when an individual performs many calculations. However, he must still carry the measures of the standard units. With accurate conversion factors the computer can convert the quantities later.

This second method is complicated by the fact that in most developing countries a plethora of units of measure exists, many of which vary in definition from place to place. Most governments are attempting to standardize units but efforts are only beginning to have an impact. The

solution to this difficult problem must be some combination of the two approaches, minimizing the number of calculations performed by the interviewer while maintaining the necessary accuracy in the data, based on the local situation as interpreted by a local specialist. In general, routine calculations such as summing various expenditure values to give total expenditure should not be required of interviewers and coders. These calculations can be done by the computer more efficiently when they are required during the analytical phase of the project.

Problems in maintaining the integrity of other glossary terms can also be expected. In this case, the judgment of the interviewer must be relied upon. When ascertaining occupation, the interviewer must probe the respondent when the occupation reported is vague, such as "businessman" or when the interviewer suspects that a term is being used in an uncommon way. In such cases, the interviewer should record both the responses given to him initially by the respondent and the more common term which he thinks more correctly describes the occupation of the respondent. The term coded will be the answer given by the respondent, while the observation of the interviewer will be included in the definition of the term. Definitions of all glossary terms should be recorded, preferably on the same punched card which holds the term and its symbolic code which is used to make copies of the glossary. This is important, for these definitions will assist the specialist in his work of identifying synonyms and otherwise grouping the terms. The definitions are to be a collection of interviewer remarks gained from interpretations of specific responses as described above and from general debriefing of interviewers conducted at various times during the survey or when the survey is concluded.

Types of codes - While coding two or more items together using special schemes is to be discouraged as costly and unnecessary, the use of alphabetic and mnemonic codes have many cost effective advantages. Such codes frequently can be abbreviations which assist in the recall of the term. A tendency to avoid alpha and therefore mnemonic codes is a result of the limitations of earlier computers and data entry equipment. These limitations, including inadequate computer memory size, and the lack of drum cards on keypunch equipment (to name two reasons) have long since been overcome. If numeric codes are used, only 11 possible states can be described by one column (0, 1-9 and blank). Using alphabetic codes, 27 possible states can be described by one column (A - Z, and blank). If three columns are used together 11^3 or 1331 states are describable using numeric codes. With alphabetic codes 27^3 or 19683, nearly 15 times as many unique states, can be described. If alphabetic and numeric codes are combined, the number of states described by three columns is even

greater. However, in actual practice it is advisable to avoid using the zero and the blank in numeric codes and the zero, the letter "O", the number one, the letter "I", and the blank when alpha and numeric codes are used. Coders and keypunchers tend to confuse and misinterpret these codes with great regularity, and IBM Fortran under OS with the Extended Error Handling Facility will convert blanks in numeric fields to zeros. This reduces the number of unique states described by three numeric fields to 9^3 or 729, and three alphanumeric fields to 32^3, or 23,168.

Mnemonic codes reduce the time required for coding as well as reduce the number of errors. Within a short time, the coders will learn the mnemonic codes for frequently used glossary terms (such as RRW for raw rice, FSH for fish, EGG for egg, etc.), reducing the time spent in looking up codes in the glossary lists. Coding years, months and days as "Y", "M", "D", will result in more rapid, error free coding than if they were coded as "1", "2", and "3". Mnemonic codes in a data base also allow for greater readability. This is important when irregularities of some kind are detected in the data during the analytical work. The analyst can consult the computer listing and quickly read the line of data and determine the nature of the problem (see Figure 2). This "readability" of the data base is a result of the mixture of alpha and numeric codes. All alpha coding would be as unreadable as all numeric coding.

It should be noted that in spite of the addition to keypunch machines of programmable drum cards which automatically shift the machine back and forth between alpha and numeric type, keypunch operators can still punch numeric data faster than alpha data. This is a result of the fact that the keyboard of the keypunch machine has the numeric keys grouped like the keys of a calculator instead of like the keys on a typewriter. This allows for numeric data to be entered with one hand while the other hand marks the place on the data sheet. Alpha data requires two hands thus making it more difficult to keep track of place and making the operator somewhat slower in handling the data sheets. But for experienced keypunch operators using drum cards, mixed alphanumeric data results in only a minor detriment in speed when compared to pure numeric data. A good mixture of alpha and numeric data will normally come from coding questions according to the number and type of responses expected. All numeric data such as prices, quantities, ages, land owned, income, etc., should, of course, be coded numerically with adequate space provided as mentioned above. When there are less than eight possible responses to a question, the responses should be coded numerically except when a meaningful mnemonic code exists, such as "M" and "F" for male and female or "Y", "M", "D" for years, months and days. In all other cases alphabetic codes are indicated.

FIGURE 2

EXAMPLES OF NUMERIC AND ALPHANUMERIC CODING

Numeric Codes only*

```
·23628117310035531000624441531101311122000000000300106401050000000000300000001
·23628700360100330400000000370000000615000010950422000000000231202051320412803002432
·23628900360100211000226104190123000200000200000000112401000000000000000004042000003
·23628700360100211000021000400124010200000201011000522000000000000000000202022000003
·23628900360100211000422040000000000000000020 000000622100000000000100000302022000003
23628241184272400600000000366000000000000000000000000000000000000000000000000000005
23628000000000000000000000000000000000012010100000000000000000000000000000000000006
--------------------------------------------------------------------------------
23629117010576031000541415212000000000000000000000100106112000000000000300000001
23629700350056700400000000730000000823075011230422000000000180700601310300328021952
23629700350100211000000261671000223110100000201010000111801060000040000000902020000003
23629700350100211000000410300021500300000000201010000111051900000010000000202010000003
23629700350100211000000251021001230100000000000000000000000000000000000000000000003
23629700350100211000000012110000000000000402020000000000000000000000000000000000003
23629820000000000000000000000000000081000011000011000000000000000000000000000000006
--------------------------------------------------------------------------------
23630146000300000232100000444156111000000000000000000000000000000000000000000000001
23630000327000000000000000000000000000000000000000000000000000000000000000000000002
23630000327200000000000000000000000000000000000000000000000000000000000000000000003
23630000327200000000000000000000000000000000000000000000000000000000000000000000003
23630000327200000000000000000000000000000000000000000000000000000000000000000000003
23630000327200000000000000000000000000000000000000000000000000000000000000000000003
23630000327200000000000000000000000000000000000000000000000000000000000000000000003
23630000327200000000000000000000000000000000000000000000000000000000000000000000003
23630000327200000000000000000000000000000000000000000000000000000000000000000000003
23630000327200000000000000000000000000000000000000000000000000000000000000000000003
```

Alpha and Numeric Codes combined**

```
3  33 1 2 3 7 8 01621 02223 0242526 0 027 0 028293031323334353637383940 41 042A 0
3  33AKPALLKR011RDIPD0  CC5000500300250500 211 1003 10271                 05A 1
3  33 1EFY51WAHSW  0 84 90 5839 7710510007 12  3    4  2  1    0 85  360     A 2
3  33 2AMY32MGCVP  0100 79 4109101 83 6951 10  2    3  1  1    0 95 1575     B 2
3  33 3B3Y22MAHSW  0124 61 1730141 57 2556  8  1    1  4  2    0107 3839     C 2
3  33 4CMM54SUCHI  0 3515111444 5111010593  4  5    5  1  4    0137 9069     D 2
3  33 5CJM23SUCHI  0 2812710004 43 38 7860  3  4    4  3  4    014410131     E 2
3  33002   002          0C8                                     004   0016  A 4
3  33PDYPD01Y  256  373   8  RVAKG01N    5    5 100  RGIPD01N  101  1011013  A 5
3  33BJAPDQ1N   50   601013  AHRPDQ1N   21   65 434  BLGDKQ1N   10    3 101  B 5
3  33RDDKQ1N   10    5 100  MLKLTW1N    1    1 100  GDLLT 1N    0    0   8  C 5
3  33MTNKGL1N    5   25 101  FSDKGW1N    5   15 101  EGGNRJ1Y   86    6 217  D 5
3  33BJLKGV1N   25   601013  DNNKGV1N   21   13 434  TMTKGV1N   10    8 434  E 5
3  33DSKNRV1N  173   34 434  LYVKTI1Y   85    6 434  PLTNRG1N   60    3 101  F 5
3  33CFNRQ1N   20    7 100  SGCKGQ1N    2    5 100  DYCKGQ1N   10   30 101  G 5
3  33TDCKGQ1N    2    5 101  CNRKGQ1N   10   16 101  SLTPDQ1N   32    6 217  H 5
3  330GRLDK120                          1      3030302010                  A 7
3  332CDRDK12DRRWDKD4C                  2      3030302010                  B 7
3  332SPRDKDCDDRDDKCDFBJLKGDC3CRDGM25C2    011015101005                     C 7
3  334CDRDK24DRRWDKD8C                  4120   3030302010                  D 7
3  334KSMDK3B3DRDDKCD5TMTGM5DDCRDGM2504030  1510101005                     E 7
3  331111111111111                      2MGDLMN                            A 8
3  330317181          2    3  2                                            A20
3  3301Y041  02Y011                                                        A21
3  3305032310655WTSWTMPF05  20317181          2    2                       A30
3  3301Y041  02Y011                                                        A31
3  33CD22C3CDRDK021020                                                     A32
3  33044810218      PCTPTC2D0802DIDLNR14024006                             A40
3  33                                                                      A60
3  33104  0  0  01C0  C                                                    A70
3  22  371   3 861023  4131286 821296    3   0  3   0  114   0  0   1104 457 A71
3  33112851241259 3125631268 3 931176 31331466 3  0   0 0  0   0 0  0   0 0 0A72
3  33    81    16    0    97    19    23  325 1  203 1  253 1    0 0    0 0 A73
3  33 2RDI 4CVR 3G 7 23  203 2 60 1056 3 55 1124 3 72 1260 3                 A80
```

Cards as punched for 3 families in the 2nd round of the TN Food Habits Survey (dashes added to improve readability).

** Excerpt from the data base entry for 1 family in the 3rd round of the TN Food Habits Survey.

When names of items are being coded such as foods, occupations, castes, diseases, etc., which belong to glossaries, three letter mnemonic codes have been found to be optimum. Three letters make a good mnemonic (recall RRW, FSH, and EGG from above) and provide for 13,824 possible items to be coded uniquely, using only alpha codes. Given that there are so many unique three letter combinations (plus 18,944 more if the eight acceptable numeric digits are added) and that computer processing and data display are greatly simplified when terms have a common format, all glossary terms should be drawn from one set of three letter codes so that every glossary term in every glossary, i.e., the food glossary, caste glossary, etc., has its own unique three letter code. If each separate glossary is numbered and the glossary terms are kept on punched cards, the cards can be sorted by glossary number for listing purposes and sorted by alphabetic code to check for duplicate codes. The TN study, which has quite large village, occupation, caste and subcaste glossaries, still used only a small fraction of the 13,824 possible combinations.

Questionnaire format - Coding and keypunching can be simplified by using a proper format for the questionnaire. If one question is listed per line and the response is entered in the margin of the sheet on the same line, the keypunch operator can quickly scan down the margin and punch the card. Codes such as "Y", "M", "D" for years, months, and days, "M", "F", "P", and "L" for males, females, pregnant and lactating, respectively, will be listed on the questionnaire and the appropriate code entered in the margin by the interviewer. Terms to be coded from glossaries are recorded next to the margin and a coder will look up the term and enter the appropriate code in the margin on the original questionnaire.

While questionnaires in this format will require more pages than usual, it will not be necessary for coders to prepare "load" or "coding" sheets from each questionnaire to be passed on for punching. With this intermediate step eliminated, much time and money is saved in the overall process. Certain types of questions, such as those concerning the demographic particulars of household members and food inventories, are best in a grid format with the data for each individual (or item) being recorded on a separate line. Adequate spacing must be provided for the interviewer to record his responses and special codes to be assigned so that the keypunch operator can quickly pick out the data to be punched. Depending on the type and number of questions to be asked, the dimensions of the questionnaire page might be varied to use the space on the page efficiently. Square and normally dimensioned pages are best used for inventory type questions which require grids, while more narrow sheets are favored for direct, one answer "opinion" type of questions.

Also printed on the questionnaire will be spaces for the household number to be entered for each card punched as well as the number of each card type. Questions should be arranged to give the most logical flow for the interview without consideration of the order in which the responses will be punched onto the cards (see below). Formats for the cards at this stage are determined solely by the space required by the responses entered in the margin of the questionnaire. When 80 columns have been used or a complete data item cannot be entered in the remaining number of columns (including the numbers required for questionnaire identification and card type), the card is finished and a new card type is begun.

Nutrition surveys, especially, ask questions which can receive a large range in the number of responses. To the question "what foods do you avoid in this house?" or "what foods did you purchase today?", the answer can be none or a dozen. The method for coding this type of question is to allow for a complete card type for each kind of "open ended" question. The question is laid out in the first available spaces after the questionnaire identifier, the card number and a column to keep track of the number of cards of this type that have been punched. (This should be an alpha code to begin with, followed by numeric codes if required and included on every card if it can have multiples or not.) The format for the question is repeated as many times as possible until all 80 columns are used or until a complete response cannot be coded in the columns remaining. The responses are coded serially on the card; if additional cards are required, they are designated "B","C", etc., the first card of every type being "A".

To minimize the number of these card types, questions of the same general type should be standardized (this is the only point in which processing requirements should be taken into consideration when planning the questions to be asked by the survey). For example, for <u>all</u> persons taking meals outside the home the following information should be obtained: the place where the meals were taken, the foods eaten, the cost per day (or meal), the frequency of taking meals and the reason for taking meals out. This will include workers who eat in factory canteens as well as children and mothers who take food in balwadies. These questions may be asked at different points during the interview, but they will be coded together using the same format. The same would be true for all household purchases, whether food or nonfood . The quantity, price, place of purchase, etc., is recorded for each purchase and coded on the same card type. At least one card for each card type should be coded whether or not there is data for that card type. If one card of each type is not coded even when there are no data, it would not be possible to determine at a later time if the card had been lost, mispunched, not coded or should in fact not have been coded.

Coding one card of each type can be avoided by using other techniques, but if the data are standardized so that several questions use a single card type, it is unlikely that a card with no data will often be required. Since these questions may be scattered throughout the questionnaire it is expeditious to assign one card type to a single keypunch operator so that he or she will become familiar with the exact location of these questions and can punch them with a minimum of training time. In fact, when punching alphanumeric data, it is expedient to divide up the questionnaire by card type so that it is not necessary to continually change the drum cards in each machine. The number of operators assigned to each card type can be adjusted so that the rate of punching yields as many completed punched questionnaries per day as possible. The cards can later be sorted into questionnaire sets.

At a later stage (see Restructing the Data Base) the data items are reorganized to facilitate computer processing. But at this current state, (the Data Entry phase) the intent is to transfer the data from the questionnaire to the computer cards (or other media) as rapidly as possible with as few errors as possible. Until this is done, no analysis can be performed, no feedback to the interviewers can be provided, and no future planning or alterations in the questionnaire are possible.

To remove as many mispunchings from the completed cards as possible, every punched card should be verified using the original questionnaire a second time. The verifying equipment (similar to regular keypunch equipment) will put an error notch over any card column in which the data being keyed into the verifier do not correspond exactly to the data originally punched on the card. An error notch on a card will prevent the card from being read by the computer. The same keypunch operator should not punch and verify the same set of cards since one of the errors that can be caught by verifying is the misinterpretation of handwritten alpha and numeric codes. Cards with errors are corrected, and then repunched and reverified. The punched and verified cards are then passed on to the next step in the process, the Edit phase.

EDITING

Objectives - The edit phase checks for gross data entry and interviewer error. The purpose of the edit and some of the later work (also see Preliminary Processing, below) is to remove all errors of the data entry type and as many obvious interviewer errors as possible. The analysis must contend with interviewer bias and misleading or inadequate information from the respondents. These problems need not be compounded by purely mechanical errors such as mispunchings. There is much redundancy in the Edit phase. This is to insure that as many errors as possible are found.

An Edit program is prepared which checks every data field of every input card. This Edit program performs five types of checks:

Check for completeness of card set -- This is a simple check that all card types are present for each questionnaire.

Check for correct type (either alpha or numeric) -- If drum cards are used and every card is verified, keypunch errors of this type should be minimal, but a final computer check is required. Errors are returned to data entry.

Check for range -- In columns where responses are coded with only a fixed number of codes (such as 2-6) the number coded can be tested to see whether it is within the range of acceptable codes for the response.

Check the validity of numbers, such as whether the quantity and price of an item are within reason, occurs at a latter stage (see Preliminary Processing). Errors are returned to data entry and back to the interviewer if the data recorded were outside the acceptable range for that question.

Check that codes are glossary terms -- This is an important check to confirm that the glossaries are up to date and to prevent some miscoded data from entering the system. As mentioned above, there should be a common glossary of standardized format, but subdivisions of that common glossary can be maintained. The coded answer for a given question is checked against the list of the subset of glossary terms which are valid responses to that question (if a miscoding is also a valid code, the error is not detected.)

A useful output of this checking process can be a series of tables which list the number of times each glossary term has been used by a given question, by a given subsample of the survey, etc.

Especially useful are tables which total the number of times each glossary term is used for each question for each interviewer. These "bias" tables can be inspected for unusual patterns occurring for an interviewer, such as reporting that a large number of families were, for example, avoiding a certain food for a certain reason while the balance of the interviewers reported only a few families avoiding the same food for any reason (patterns of this type have been identified in the TN Food Habits Survey data). Statistical tests can be performed routinely to establish the significance of these patterns. If a pattern is found to be significant, the interviewer's supervisor can be instructed to observe the behavior of the interviewer more closely. Interviewers should not be threatened with this monitoring of their performance. It should be presented as a helpful tool in maintaining the interviewer's professional objectivity, which is essential to a successful survey and accurate analysis.

Internal consistency - When —as is normally the case in nutrition surveys— a complicated questionnaire is used, internal consistency checks should be performed. Such a check is needed, for example, when there are many "junction" type questions, those which if answered in a desired way require that a number of additional questions be asked before the next question in the main sequence of the questionnaire is asked. The Edit program works like an accountant's balance sheet: noting responses which indicate that additional questions were to be asked, that additional punched data can be expected and noting if they are found.

When errors are found, the original questionnaire is examined and returned to either Data Entry or to the interviewer. Additional cards are punched or the interviewer returns to the respondent to gather the missing data. As was the case in bias checking above, computer checking for internal consistency should not be used as a threat to the interviewer, but merely a check that they understand the questionnaire and are fulfilling the job they are being paid to do. Consistency checking will also make falsification of questionnaires by interviewers and deception by respondents less likely to succeed. The programmer who prepares the Edit program must have a good understanding of the questionnaire so that important consistency checks are identified and performed.

For monitoring the interviewer's work via bias tables and consistency checks to provide timely feedback, the editing must take place as soon as possible after the interview, certainly within two weeks of the interview (see the Coordination and Integration section below for a more thorough discussion of timing requirements).

When a questionnaire has passed all of the Edit checks without error, it is passed on in card, tape or disk file form to the next phase in the process, Restructuring of the Data Base.

RESTRUCTURING OF THE DATA BASE

Purpose - The purpose of the Data Entry and Edit phases is to transfer data from the questionnaire to computer cards easily, quickly and error free. This should be done while maintaining the integrity of the questionnaire down to the smallest detail, so as to hold transfer costs to a minimum and to allow analytical work to begin at the earliest possible moment, but in no way limiting the ultimate scope and precision of the analytical work. The purpose of the restructuring phase is to convert the computer cards from the form most efficiently processed by humans to a form most efficiently processed by the computer. Fundamental decisions concerning the uses to which data are to be put must be made before the restructuring program can be written (see the Introduction and the Method of Data Handling in Preliminary Processing section).

Storage of relationships - The most important consideration in the restructuring program is to provide adequate space for calculations and groupings. As emphasized above, calculations are best left to the computer and glossary items are better left unaggregated, to be later grouped and regrouped as required. When the input data are put in a new format by the restructuring program, adequate space must be allocated to store various calculations such as total expenditure, per capita expenditure, nutritional needs of the individual, etc., to be calculated from the data and various groupings to be applied to the glossary items and calculations. Calculations should be stored in their most concise and useful form. For example, nutritional status in relation to protein, calories, and vitamins is most easily expressed as a percent of needs met instead of grams, calories, or units received. The quantity received can always be recalculated by multiplying the percent need met by the need. Furthermore, nutritional status is most correctly understood as a percentage of need met, since need varies greatly with age, sex, activity, etc. Glossary

terms will probably be categorized several times for different purposes, so space must be left for a number of categories. The group membership of of various mnemonic codes does not necessarily have to be stored though it is usually more efficient to do so if the grouping is to be used more than once. The mnemonic codes should be retained in the data base for the readability they give to computer listings (discussed above) even though later analytical programs are likely to refer to the groupings assigned to a given mnemonic code more often than to the code itself. The ability to store calculations, groupings, etc., is basic to the development of a data base of high utility which can be analyzed at low cost.

In the TN study, it became obvious that a full analysis of the Food Habits Survey data would require easily as much space to store calculations, groupings, etc., as was required by the original data. If the data are stored on disk and a direct or "random" access method is employed, the amount of space used has little effect on the speed of the processing (see discussion on Data Handling), so adequate space can be allowed for, within reason. The data base can reside on mountable disk packs which are mounted only when required and are fairly inexpensive. To maintain simplicity in the data base, the space for calculations on an individual should be grouped with the basic data on the individual, the space for various groupings of an item purchased should be adjacent to the mnemonic code for the item as well as the purchase information, and similar provisions should be made for other data handling needs, as appropriate.

New cards required - After restructuring, the number of records per questionnaire is likely to be twice the number of the input cards. This may result in only one individual, for example, per card instead of four or five as on the input cards. Whereas the input cards were 80 characters long, the data records in the restructured data base can be of any length. Entirely new card types must be created to record the family totals for various indexes. Each household in the TN Food Habits Survey data base has, in addition to the above, several records (card storage) of certain data. The restructuring program must also create whatever index cards or indicators are required by the access method used to read the data base.

PRELIMINARY PROCESSING

Preanalytic calculations - Preliminary processing is a preanalytical phase in which indexes such as the monthly expenditure, nutritional needs and nutritional status are calculated. It is distinguished from the analytical phase by the fact that no statistical or other analysis is performed on the data base as a whole. Each questionnaire is processed independently. The writing and debugging of computer programs to calculate these indexes is much simplified if the instructions for the preceding three phases have been adhered to judiciously. The calculations from each run are stored in the data base to be used as data by subsequent runs. For the TN Food Habits Survey, the Preliminary Processing included the following:

1. Calculation of protein and calories requirements (in terms of adult equivalents)* per individual based on the age, sex and reproductive status of the individual. The occupation and therefore the additional protein and calorie needs required or substracted for the level of activity was not taken into account since the primary emphasis in the study was on infants and pregnant and lactating women, not working adults.

   In addition to storing the protein and calorie requirements for each individual, the requirements were summed to yield a daily family requirement; this daily family average was multiplied by 30.42** to give average monthly requirements for protein and calories. Both daily and monthly requirements for protein and calories were stored in the data base.

2. Normalization of food purchases to a 30.42 day month. In rounds 3 and 4 of the Food Habits Survey, food purchase information was obtained on the most recent purchases and included an estimate of the frequency of

---

* 1.00 Adult Equivalent (AEQ) of calories = 2500 calories
1.00 Adult Equivalent (AEQ) of protein = 45.0 grams of reference protein.

** Note that (30.42 x 12=365.0) whereas (30x12 = 360.0) and (31x12 = 372).

purchase of that food. Based on the frequency of purchase, the quantity and cost of each food item obtained for an average month was calculated.***

3. From the normalized food purchase data, monthly food expenditure was calculated and added to nonfood expenditure, which was obtained for an average month, to give total family expenditure and food expenditure per month. The per capita expenditure (total monthly expenditure divided by number of persons in the household) and the total monthly expenditure per adult equivalent of calories required (monthly total expenditure/total number of calorie adult equivalents required per day) were also calculated in rupees and stored for each family.

4. Family protein and calorie status was calculated and stored as percent of monthly family protein and calorie needs (in terms of AEQs) met, based on the normalized food purchase data. Standard reference protein and calorie values were used for each food item.

5. Individual protein and calorie status was calculated from the 24 hour food recall question. This question reported the food eaten, the time prepared and time eaten, the quantity prepared, the quantity of ingredient used and the quantity taken by each individual. Standard food values were used for each food item except when the ratio of ingredients was known to affect the nutrients in the food prepared. In such cases, the quantity and type of ingredients were taken into consideration. The supplemental effect of amino acids on a meal basis was not taken into account although the quantity of reference protein received was adjusted by the protein/calorie ratio of food taken by each individual for each meal. Nutritional status of the individual was stored as the percent of protein and calorie needs met. Family averages were calculated by dividing the total nutrients ingested by each individual by the total family protein and calorie requirements per day.

*** This is in contrast to the first and second rounds when the respondent was required to estimate the quantity and cost of each food item in an average month.

Data Handling Program - The Food Habits Survey Data Base was maintained as a disk file and was accessed via the Hermes data handling subroutine (written in IBM Fortran IV, Level G). The access method employed is IBM Basic Direct Access Method, more commonly known as 'BDAM' or "direct" accessing. Direct accessing allows for records to be selectively read and written in any order throughout the entire disk file. This method has many advantages which far outweigh its few disadvantages when compared to conventional, sequential processing using tapes or disks. Among the reasons for selecting a random access method was that when working with the file, only a small fraction of the data is read during any one pass, that the passes occur frequently, that a small number of records is usually written in each pass, and that the file is very large (250,000 card images).

Hermes is an extremely flexible program which will read and write data required for a given program, allowing for the majority of data records to be skipped during each pass. Data are passed to the calling program via labeled common statements which are included in the calling program as required. This relieves the programmer of constantly spending his time on input-output problems. With modification of the Hermes format and I/O statements, the Hermes subroutine can be used to access survey data in any format, provided an index card is created for each card set indicating the number of each card type present. The Hermes subroutine can also be used to extract a set of detail records from the master data base. This set of "file" of detail records can then be analyzed in the sequential mode and results written back into the master data base via Hermes.

With the conclusion of the Preliminary Processing, analytical work of a substantial nature can begin. This is not to deny that great familiarity and confidence in the data comes from observing and monitoring the data as it is worked up in the above manner, or that opinions concerning the nature of the conclusions to be found haven't been formed. But all of those impressions and opinions are only personal intuitions which must be confirmed or denied by the more objective statistical techniques of the analytical phase.

COORDINATION AND INTEGRATION OF THE SYSTEM

"Process" identity - Several concepts and operating procedures have been discussed in the above sections. Occasionally, the importance of proper timing was alluded to. Proper timing and sequencing are at the very roots of a successful survey project. This section is, and rightly so, the most important in this chapter since the order and timing in which the above steps must be performed is discussed.

One may have gathered from the preceding that what was being described was not a set of tasks that would be completed one after another, but a highly integrated ongoing system of continuous data gathering and analysis. Critical feedback loops, as a part of the system, would allow the survey to be adjusted and redirected if required to achieve the desired results. However, before this continuous process begins, there must be a long period of planning and development. The overall objectives of the study and the survey questionnaire, sample design, preliminary processing and analytical techniques required to reach that objective must be defined.

Format organization - The questionnaire can be put in the proper format for easy processing at the stage of pretesting. The necessary editing, restructuring, preliminary processing and analytical programs must then be created or modified from existing programs and tested (on the questionnaire pretest data) before the survey work can begin. To minimize the time required for this process, several programmers should be assigned to the task (see Estimates section below). During this time the analytical staff of the project must necessarily work closely with the programmers. However, they would also have time to identify and collect secondary data in the form of government and private statistics that may be useful, monitor (or supervise) the selection and training of interviewers and perform any necessary presurveys. In spite of the high level of activity that would be occurring in all areas of the project at this time, it is essential that a group of individuals on the main staff be responsible for understanding every aspect of the project so that all aspects will function as one smooth system. This is especially important if the work is done on a subcontract basis. Subcontracts must be flexible enough to accomodate modifications of instructions as the survey progresses, since adjusting the survey process is expected to occur as a matter of course. When the computer programs have been written and successfully tested, the sample design and questionnaire perfected, the process of survey and analysis can begin.

Processing system - Figure 3 is a diagram of the processing system in operation. The data flows and feedback loops are indicated for one cycle only, though they would occur continuously throughout the duration of the project. All but one of the elements and feedback loops on the diagram have been discussed previously, but they will be reviewed again here briefly.

Survey Work - Refers to the interviewers in the field collecting data.

Data Entry - Includes minor coding, keypunching, and key verifying onto card, tape or disk with the original questionnaire.

Editing - Checks all of the data fields on the input cards for various errors. Scans card sets (questionnaires) for internal consistency. Interviewer bias and other tables maintained.

Restructuring - The input cards or card images are converted (reorganized) from the form most suitable for rapid and easy Data Entry to the form which facilitates computer processing.

Preliminary Processing - Calculation of various indexes such as nutritional status, expenditure, etc., for each card set (questionnaire).

The Editing, Restructuring and Preliminary Processing should occur as a concerted process, the data passing from one stage to the next within the same program rather than being submitted to three separate programs. Various economies can be achieved if these three programs are chained together so that the input cards are read only once (assuming no errors) and the output is the completed data base entry.

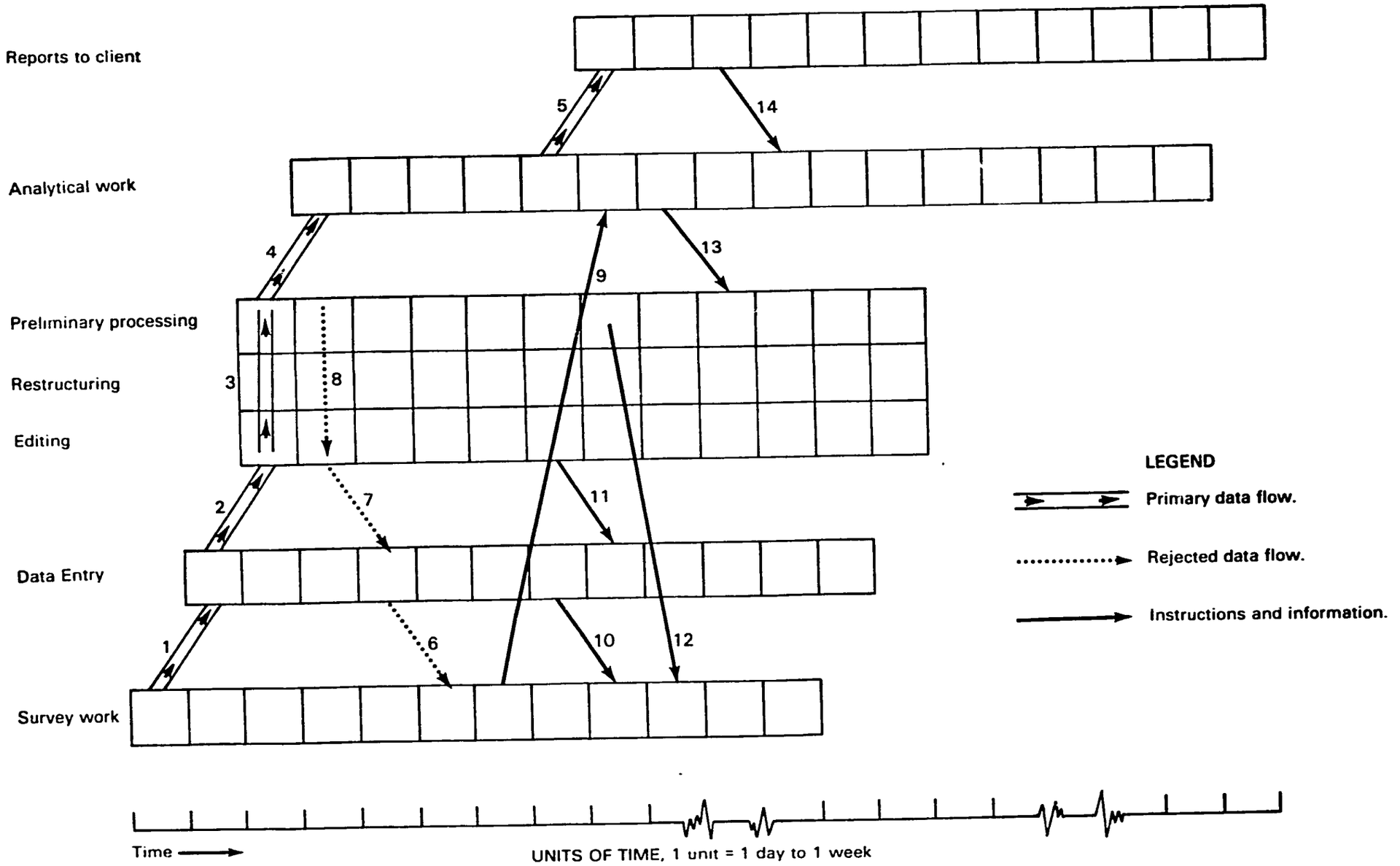
Analytical Work - Statistical and other analysis of the available data base.

Reporting to the Client - This element has not been previously discussed.

The client is defined as an individual or office outside of the project that the project is either intended to serve or is funded by. (The project staff should be exhorted to appreciate who their "client" is, early in project.) A feature of the ongoing survey process technique is that reporting with

FIGURE 3

THE ONGOING SURVEY AND ANALYTICAL PROCESS



See text for commentary and identification of numbered pathways.

meaningful content to the client can begin very shortly after the beginning of field work if the system is functioning properly. This will allow the client to better understand the nature of the study and to justify its continuance. If reports with hard content do not come to the client, he can rightly conduct an investigation and alter the staffing pattern if deficiencies are found.

Primary data flow - Arrows number 1, 2, 3, 4, and 5 indicate the movement of error free data through the system. Vertical arrow 3 is indicative of the concerted process at that point. At arrow 1, the data are in the form of survey questionnaires, arrow 2 as either cards, tapes or disk files, at arrows 3 and 4, it is a random access disk file and at arrow 5, the data are in the form of a written report.

Rejected data flow - Arrows number 6, 7, and 8 indicate the movement of data that has been found to be in error. The nature of errors is either a mechanical Data Entry type error or interviewer error of some kind (distortion, omission, etc.). These are gross feedback loops.

Flow of instructions and information - Arrows 9 through 14 are sensitive feedback loops which can adjust the direction and emphasis of the entire system. Arrow 9 indicates interviewer reaction to the use of the questionnaire. Arrows 10 and 11 indicate the feedback on interviewer performance from the bias tables and other Edit and Data Entry checks. Arrow 12 indicates a change in the data gathered by the survey that is required for the calculation of additional indexes or more precise calculation of existing indexes. Arrow 13 indicates changes in the Preliminary Processing and possibly in the Restructuring program to provide for new or expanded analysis (this may also be carried down arrow 12 depending on the requirements). Arrow 14 is feedback from the client in terms of approval or disapproval, the meeting or not meeting of the client's needs, or the alteration of the project to the client's changing needs. A small amount of activity in loop 14 alone can produce considerable activity throughout the rest of the system.

Rejected data traveling down arrows 8, 7, and 6 eventually results in new activity in the primary data flow, arrows 1 through 5. An especially reactive loop can be initiated by minor activity of arrow 13, a change in the requirements of the analytical work. This results immediately in the alteration of the Preliminary Processing program and possibly the Restructuring program. Depending on the nature of the requirements,

arrow 12 may also be activated, represented as a change in the questionnaire. In turn, arrows 6, 7, 8, 10, and 11 may be activated if the interviewers do not fully understand the new requirements or arrow 9 if the new question is too difficult to administer. The point of this is that the processing system described here is a dynamic one that operates with a very delicate balance. Keeping the system in balance requires constant vigilance. If not properly managed, the system will bring itself to a halt automatically. It must be pointed out that this system does not create problems. It only identifies them as they develop and allows them to be analyzed and dealt with appropriately. In the conventional survey situation, where the data are not analyzed until long after the field team has disbanded, various shortcomings will have gone undetected and uncorrected during the survey phase and allowed to invade the data base. The only solution in that case is to purge the defective questionnaires (as many as half of the total questionnaires) or just learn to live with the distortions and other shortcomings that may or may or may not be identified in the data. The processing system can be thought of as a built-in evaluation system of the entire survey-analytical process.

Time scale - The linear scale along the bottom of Fig. 3 is the time scale. The period of time assigned to each unit depends on the level of technology used (see below) and the amount of coordination that can be engineered by the project staff. Assuming that the interviewers move as teams and the data entry medium is punched cards, the unit of time on the scale can reasonably be expected to equal one week. This would require that questionnaires be collected from the field and returned to the central processing office, either by a staff car and driver or registered, special delivery postal service at the end of each week (in the TN Food Habits Survey, about 200 questionnaires were completed per week). At the beginning of the following week, Data Entry work could begin and a complete set of verified cards would be ready by the end of this second week. An appropriate number of coders, keypunchers and verifiers would be hired to maintain this pace as well as correct errors identified by later steps.

During the third week, the cards would be passed through the Editing, Restructuring and Preliminary Processing programs and have all Data Entry errors corrected and then added to the survey data base. (Questionnaires with interviewer errors would be returned to the field with the weekly trip of the driver or by mail to be resubmitted when corrected.) Analytical work could then start in the fourth week after the survey work began. Depending on the nature of the sample design and the order in which interviews are conducted, entire geographic subsamples could be collected in a week or two and analyzed before the entire survey was completed. Significant reports could then be sent to the client on a regular basis.

Rate of the process and speedup techniques - The rate determining step in process is the frequency with which questionnaires are gathered and are passed to the Edit and subsequent steps via Data Entry. Increases in efficiency in this area will be felt throughout the system as more responsive feedback loops. For example, interviewer bias is not reviewed until the third week during the Edit phase (about two weeks after the interview). If the interviewer is found to be at fault, he has conducted two weeks worth of faulty interviews which must be resurveyed as well as purged from the data base and be reentered with the more accurate responses. One method of increasing efficiency in this area is to use key-to-tape equipment which allows for the keypunch operator to key directly onto a computer tape (which can be read by the Edit program). This eliminates the cost and burden of handling paper cards but requires, of course, that reliable and adequate numbers of key-to-tape machines are available.

Another method of reducing the time required by data entry is to enter data via remote terminals over phone lines to a central computer facility. Using this approach, coders and keypunchers would travel with the interviewing team in the field, carring with them glossaries and portable computer terminals. Each day, the coder/keypunchers would type in the questionnaires from the previous day, and the data would be stored directly on a disk file, eliminating the intermediate storing of data on entry cards or tape. Each night, the Edit, Restructuring and Preliminary Processing programs would be run to process the new data received and reports indicating the items found to be in error would be printed. These error reports along with the updated glossaries could be listed at the remote terminal at the start of each day and corrective action taken immediately on errors found in survey work only two days old. Other communications between the project staff and the field team supervisor could also occur regularly via the remote terminal. This method reduces feedback time from two weeks to two days, and allows the project staff to remain in constant contact with the field teams. The method requires that the central computer facility have a reliable time sharing system, that portable terminals be purchased (which can be used like the more conventional IBM typewriter type of computer terminal when the field work is completed), and that adequate telephone and power sources are available in the field.

The third and most exciting method of reducing feedback time and maximizing output of the survey is for all computer processing to be performed on a minicomputer which would, along with the entire project staff, be mobile and move with the survey team. A minicomputer system would allow

the keypunch operator to be connected directly to the computer which would perform the Editing, Restructuring and Preliminary Processing as the data are keyed in. Errors would be identified in microseconds instead of in days. Analytical work could be performed by the mini or the data easily transferred to tape and analyzed by a large scale computer if additional resources were required. The processor speed of most minicomputers is equal to, if not greater than, most large scale computers. Minicomputer systems with adequate memory disks, tape drive, terminal and printer cost approximately US $50,000. The system would be self-contained in an air-conditioned vehicle (in the light truck category) with its own power source. The advantage of using a mobile minicomputer system is that errors of all types can be identified and corrected within 24 hours, insuring as accurate a record of the respondents' attitudes as they can be related. In a mobile situation, the project staff would travel with the unit, making observations and collecting additional data at each sample site. It would also permit them to direct the work of the interviewers and any social scientists employed in the survey. This closeness with the world to be analyzed would undoubtedly also pay off in keener insights into the problems being investigated.

Another important advantage in using a minicomputer, either as a mobile unit or in a more conventional stationary setting, is the cost control that it offers. Estimating the computer resources (and therefore the cost of computer processing) required for Data Entry, Preliminary Processing and Analytical work is extremely difficult. Cost overruns in computer operations are not uncommon. With the purchase of a minicomputer system, costs are fixed at one price (the price of purchase). It is not likely that additional computer resources would be required since the mini can essentially be run 24 hours a day if need be. Another important consideration is that a minicomputer makes the project independent of a large scale computer facility and its procedures, problems and operating schedule. This alone can be the major consideration in selecting a minicomputer system over a remote data entry system using the time sharing system of a central computer. If processing delays are encountered, the feedback loops are thrown out of schedule, the processing system will begin to deteriorate and the schedule of the entire project must necessarily be revised.

The single significant disadvantage of using a minicomputer system, either as a mobile or stationary unit, is the problem of servicing. Both routine service and emergency repair can be a problem outside the United States and Western Europe, especially when the computer is required on a daily basis. Interruption of processing on the minicomputer would have the same

bad consequences as the interruption of service on a large scale computer that was relied upon. Minicomputers have, however, shown themselves to be highly reliable. Also to be considered is the fact that the minicomputer can be used for followup and additional survey work, using the same programs with only minor modifications. This fact makes the acquisition of a minicomputer system especially attractive to government agencies and companies charged with survey work on a regular basis.

Regardless of whether a mini or a large scale computer is employed, one and only one computer facility should be used. Coordinating work done at a number of computer centers, especially if the computers are not of the same type (or worse still, not of the same generation, such as trying to combine the work of an IBM 1401 and an IBM 360) is not likely to result in any saving of time or money.

ESTIMATE OF PROGRAMMING REQUIREMENTS

The data processing activities of the Tamil Nadu project were not a model of the process described above in action. Many events prevented the smooth process that was originally planned by the computer subcontractor from developing, and many additional insights were gained as the project progressed which have shaped the process system presented here. Inspite of the delays and unusual circumstances under which the data processing for the TN project was conducted, reliable estimates of the amount of time required for computer programmers to develop and operate a system as described here can be made.

The man months required to develop the programs required to process survey data is presented in Fig. 4. The time estimates include the time required for the consultations and planning for each phase in addition to generating the actual computer programs. The various tasks outlined in Fig. 4 must be accomplished at some time during the course of a large scale computer analysis. No step can be avoided without considering the cost of the project, and allowing data of questionable reliability from entering the system for analysis. It can only be highly recommended that the programming effort be concentrated at the beginning of the project to insure that the data gathered are of the highest quality and reports concerning the project will appear in a timely fashion. A staff

FIGURE 4

MANPOWER REQUIREMENTS FOR PROGRAM DEVELOPMENT

Task	Time Requirement	Remarks
Data Entry	15 man days	Layout questionnaire and establish coding and punching procedures.
Editing	30 man days	Edit program including creation, retention and analysis of "bias" type tables.
Restructuring	30 man days	Restructuring program and adaptation of Hermes Data Handling Program.
Preliminary Processing	60 man days *	Programs to calculate various indexes.
Analytical Processing	60 man days *	Selection and adaptation of existing statistical and display programs, and development of some new techniques.
Sub-total	195 man days	
Orientation	15 man days per programmer	Orientation to project and meetings to coordinate various program segments.
Total Program Development Time Required	255 man days (8½ man months) or about 2½ to 3 calendar months	Using 4 programmers, 2 programmer analysts and 2 programmers with no special experience.

* Assuming (1) questionnaire similar to the one used for TN Food Habits Survey, (2) that a project analyst works with the programmers at all times, and (3) that computer time is available on demand.

of four programmers is suggested for the development of the program. Two would be experienced programmer-analysts who would remain with the project after the development of the program system. One would oversee the Data Entry, Editing, Restructuring, and Preliminary Processing of the survey data while the second would work closely with the staff analysts in performing the computer analysis of the data. The remaining two programmers could be hired on a temporary basis only unless there were other parts of the project which required additional computer work after the development of the survey processing programs.

A team of four programmers could be expected to complete and successfully test all aspects of the system within three calendar months. The work would begin as soon as the questionnaire for the survey was in final form, presumably after the pretest of the questionnaire. Due to the interlocking nature of the processing system, the Data Entry activity must be completed first,followed by the Restructuring program. The Analytical, Preliminary Processing and Editing programs can then be worked on simultaneously with the questionnaire pretest data being used as test data. It is desirable that the programmer who will ultimately be responsible for the analytical operations work on the Data Entry layouts. This will familiarize him with the nature of the data to be analyzed and enable him to discuss the project with the analysts without an additional orientation period. The estimates presented here assume that the amount and complexity of data to be collected is similar to that collected by the TN Food Habits Survey, that a project analyst works with the programmers at all times and that computer time is available on demand. In the case of the TN project, computer processing was allowed on a "time available" basis at an AID computer facility in Washington, D. C., (that is, TN work could be processed only after most of the regular AID work had been processed). The unpredictable nature of the processing time that would be available made it virtually impossible to plan and meet deadlines or to work at a normal pace.* The same would be true of any other project requiring a computer if the availability of the computer was also restricted.

* It should be noted that the staff of the computer facility attempted to optimize the working conditions as much as was possible under the circumstances. For their efforts, the project staff is grateful.

Estimating the amount of computer time required is, as was mentioned above, basically impossible without examining the specifics of the project. The amount of data to be processed, the computer language, the skill of the programmer and the type of computer tabs used are all major sources of variation in the total time required. For the sake of a rough approximation only, for a typical month during the course of the TN work in which all phases of the process were being performed, including program development, the computer resources required were as follows:

Number of Jobs	180
Elapsed Time	29 hours
CPU Time	12 hours
I/O Requests (90% disk)	1.5 million
Device Mounts (Tapes and Disks)	350
Language	IBM FORTRAN IV, Level G
Computer	IBM 360 Model 50, OS

The above is for processing the complete Food Habits Survey data base of approximately 20 million bytes. If the programs were completely developed before the survey work began and the survey data processed on a continuous basis as outlined above, the computer resources required on a monthly basis would be considerably less. To process the entire TN Food Habits Survey, using the system discussed here, the total computer resources required would be equal to about five or six of the above months.

With some additional work, the TN Preliminary Processing and Analytical programs can be assembled into an easy to use program package for nutritional analysis. Such a package would reduce the development time required for the Preliminary Processing and Analytical programs from 120 man days to about 30 man days. The 30 man days would be required for the programmer to familiarize himself with the package and select the options required for his project. The time required for the development of the specific Data Entry, Editing, and Restructuring programs would remain approximately the same, though the programmers

would be assisted by examples in the package. Overall, the use of a program package for nutritional analysis would reduce the programming support required for each project that employed the package by about two man months. This would mean reducing the programming staff from four to two and still developing the entire system in less than three calendar months. The wide use of a package would also insure that directly comparable results would be generated by different projects working on food and nutrition problems.

GENERALIZED DATA BASES

The discussion up to this point has centered upon the efficient handling of survey type data. However, almost every sample survey requires additional data for the purpose of validating, analyzing and making projections based on the survey data. These additional data are often referred to as "secondary data" and include population, economic and agricultural statistics.

Whenever a large amount of secondary data is required and found to be available, steps should be taken to facilitate the comparison of the survey data with the appropriate secondary data. This can be accomplished by adding a code identifier to each questionnaire which places it within the scheme of the secondary data. For example, in India there is a large body of very good population and economic data collected by the National Sample Survey. To improve the value of the data collected by the N.S.S., the entire country has been divided into homogeneous zones, many of which cut across the lower political subdivisions. By including a code for the N.S.S. zone, the survey data can be collected for each zone and analyzed independently or in groups of zones and then compared and augmented with the N.S.S. data for that zone or group of zones.

As a matter of course, the name of the town or village in which the questionnaire was completed should be coded. The larger political division to which the town or village belongs should also be coded. Even if it is unlikely that the first analysis will utilize a particular subdivision, it is likely that at some later time, when additions are made to the data and entirely different groups with different interests than the first analytical group use the data, certain subdivision will become important. In any event, encoding the town or village name insures that the data can later be grouped into N.S.S. zones, (if they were not originally coded) or any other grouping that may be found to be valuable.

Although secondary data are usually derived from surveys such as the Food Habits Survey conducted by the TN project, only summary data are usually required. For example, the totals for the production of various crops by taluk for the past 20 years along with data relating to rainfall, amount of arable land, the use of fertilizers, etc., would be valuable secondary agricultural data for analyzing the food and nutrition system in Tamil Nadu. This summary type of data must be specially coded and accessed by sophisticated computer programs to be useful on a large and meaningful scale. Many of the same principles discussed above, such as storing the data in the smallest possible aggregate, using glossary terms and standardized data formats which are consistent throughout the data base, and maintaining the data base as a randomly accessed disk file, apply equally well here. It should be noted, however, that the creation and maintenance of a generally useful nonsequential data base is too complicated a topic to be discussed fully here, and that if the creation of a secondary data base is contemplated, it is strongly advisable to consult with an experienced computer specialist.

As part of the TN project, a sophisticated random access computer program to service a large secondary data base was developed.* However, due to unavoidable delays and the plan of the project, it was only at the close of this phase of the project that the analytical work had reached the point at which this program would have been useful.

An inventory of the tapes comprising the data bank of the Tamil Nadu Nutrition Project is contained in Tables 1 and 2, on the following pages.

* Symbionics, Inc., FAT development.

TABLE 1

TAMIL NADU FOOD HABITS SURVEY DATA TAPE

(PROCESSED DATA)

Three tape copies are available, created on IBM 36Ø equipment:

VOL = SER = TAMIL3
STATE3
SMCØØ3

9 Track, 1600 BPI

DCB = (RECFM = FB, LRECL = 8Ø, BLKSIZE = 12ØØØ)

Documentation available from contractor's home office on request.

LABEL = (X, SL)	DSN =	DESCRIPTION
1	TAMIL31 STATE31 SMCØØ31	3rd Round of Food Habits Survey (approx. 1Ø7, 194 records)
2	TAMIL32 STATE32 SMCØØ32	4th Round of Food Habits Survey 1st and 2nd Round Replicates Mini Food Habits Survey, Coimbatore (approx. 13Ø, 767 records)

TABLE 2

TAMIL NADU PROJECT DATA AND SAMPLE PROGRAMS

Three tape copies are available, created on IBM 360 equipment:

VOL = SER = TAMIL2
STATE2
SMC002

9 Track, 1600 BPI

DCB = (RECFM = FB, LRECL = 80,* BLKSIZE = 12000)

Documentation available from contractor's home office on request.

LABEL = (X, SL)	DSN =	DESCRIPTION
1	TAMIL21 STATE21 SMC0021	Take Home Test Growth Data
2	TAMIL22 STATE22 SMC0022	# 3 Day Weightments Survey Data
3	TAMIL23 STATE23 SMC0023	Morbidity Data
4	TAMIL24 STATE24 SMC0024	ORG Food Processing and Distribution Survey (80 character records only)
5	TAMIL25 STATE25 SMC0025	ORG Food Processing and Distribution Survey (800 character records only) * LRECL - 800 this data set only
6	TAMIL26 STATE26 SMC0026	Secondary Data * LRECL - 20 this data set only

7	TAMIL27 STATE27 SMCØØ27	Protein, Calorie, and Amino Acid values required for 24-hour food recall program
8	TAMIL28 STATE28 SMCØØ28	Protein, Calorie and Food Group values required for 30 day food inventory program
9	TAMIL29 STATE29 SMCØØ29	Interunit Conversion Factors required by unit conversion subroutine (CONVRT)
10	TAMI.12Ø STAT 12Ø SMCØ12Ø	Hermes Data Handling Subroutine
11	TAMI1.21 STAT121 SMCØ121	Sample Driver and MULTD subroutine for multi-dimensional displays
12	TAMI122 STAT122 SMCØ122	HERMES common and type statements / program to calculate individual and family protein and calorie requirements / program to calculate individual and family nutritional status from 24 hour food recall / CONVRT interunit conversion subroutine
13	TAMI123 STAT123 SMCØ123	Program to normalize food purchases / program to calculate family nutritional status from 30 day normalized food inventory data
14	TAMI124 STAT124 SMCØ124	Program to assign rank order and pentile values / jobstream JCL for assigning rank order values

INTRODUCTION

The analytical system and methodology used in this project has a number of innovative aspects. It begins with a data bank of processed original data that are completely unaggregated.* Data are made accessable by programs that permit ordering and reordering in any desired way for computations data displays, and analytical operations. A detailed description of the technical process applied to development of the data bank was presented in Chapter IV. The resulting data base made it possible to meet the fundamental requirement set up for the analytical program: that there be no preconception of findings or relationships and, as a consequence, no presumption of prior knowledge of causes and effects. This deterministic requirement, that the relationships emerge from the data, is a direct outcome of this innovative approach to systems development; later, when the system is adequately characterized, it leads to construction of simulation models. The approach uses the data themselves and resulting analytical findings to determine the fine structure of the system. Only broad, more obvious components of the system are identified and interrelated in advance (for example, division of the Tamil Nadu food and nutrition system into three component subsystems: agricultural production, food processing and distribution, and consumer practices and behavior), and even this broad construction is maintained with a degree of flexibility over the interacting system components.

Quality of the data base and, in turn, analytical findings of a systems investigation depend upon how nearly representative the sample is of the population under study. Ideally, a completely random sample of the entire population would be used. But for practical reasons, most compelling of which is cost, compromises are made. The sample design adopted for this project, described in Chapter II, compromised the ideal by two practical distortions. First, in keeping with the Government's primary nutrition

---

* The data bank also contains much secondary data, which are made available at a variety of levels of aggregation and stratification. It then becomes of greatest importance to select the forms and degrees of aggregation of secondary data that are most useful and applicable to the work to be done with the original data. For example, the geographic sampling zones of the National Sample Survey are the most appropriate basis for stratification of a consumer survey sample.

objective of improving the survival and quality of children, the sample frame included a modest positive disproportion in the numbers of pre-school children and pregnant and lactating women. Second, for practical reasons which are somewhat arguable after the fact, a higher proportion of households classified as urban than occurs in the actual population of the State was allowed in the sample. These distortions have been taken into account in the analytical work and evaluation of findings. Based on the results to date, elimination of the rural/urban distortion is strongly recommended in the continuing work of the project.

The analytical methodology, based upon rank ordering, partile distribution, and vector analysis, is highly innovative in the sense that, to the best of our knowledge, this is the first time that these techniques have ever been applied to a nutrition survey or a broad scale study of a food and nutrition system. The preservation of the raw data bank and computer programs for calling out data and ordering them for analysis provide equal access to any more familiar, conventional approach to data analysis and systems modeling. The work described in Chapter VI is a case in point. The primary motive underlying the innovative efforts that have gone into the Tamil Nadu Nutrition Project has been recognition of the controlling influence of locally indigenous, ordinarily non-quantifiable cultural factors on the acceptance of nutrition improvement measures by designated beneficiaries of a program.

Existing methodology in the field did not offer a satisfactory means of introducing the cultural component into the analysis, but the data technology and analytical system developed in the present work have provided a definite start toward that objective. An essential part of the work has been the development of multidimensional data displays to facilitate analysis of complex interrelationships, including rank ordered or categorized social and cultural factors. The most commonly used analytical technique, linear regression analysis, was felt to be inappropriate; in a developing society such as that of a village in Tamil Nadu, the interactions among all phases of the economic society are incomplete, sometimes tenuous, sometimes lacking. Too often the numbers obtained in such analyses, proposed as representing degrees of association in a collection of two factor relationships and multifactorial associations with single "dependent" variables, are accepted as measuring cause and effect relationships and used as the basis of planning decisions. Too often, wrong decisions are made on such bases. The complex food and nutrition system of Tamil Nadu, culturally controlled and economically ill-defined, required an analytical tool more revealing than an equation derived for a developed society and more compact and workable than the raw data file.

The following discussion traces the development of such a concept from the original input/output data collection through the development of linear data "tree" or relevance associations from food production to final consumption. A methodology for presenting the data in five variable arrays is presented and finally the analytical tools for assigning numbers to visually perceived target areas.

ADHERENCE TO INPUT/OUTPUT PRINCIPLE

The key to any analytical process is the identification of a base line or set of reference points from which measurements can be made. In the analytical discussions to follow, every effort is directed to characterization of input-output nodes through which base line data will be identified and compared. A node is any point in the system at which a set or subset may be characterized, described, and specifically identified so that input to it and output from it can be measured. Nodes are the points of connection of the component parts of a "data tree" (or relevance tree);* position of a node on the data tree determines whether the data set it connects is dominant or subordinate to other sets or branches of the tree. The origin and direction of propagation of the data trees may be varied at will; cause and effect may be inverted. This allows a fully recursive study of the system. What may appear as a supply cause of limited availability of a commodity, can easily explode into a demand pull inflation for the same commodity which in time, as prices rise, will cause further limitation of supply to the poor.**

---

* "The point where a branch is subdivided is known as a node. While there must be at least two branches emerging downward ("depending") from a node, there is no upper limit on the number. In addition, there is no requirement that each node in a specific relevance tree have the same number of branches depending from it....The branches depending from a node must be a closed set...." --J. P. Martino, Technological Forecasting for Decisionmaking. Elsevier, N.Y. 1972. Chap. 9 - Normative Methods. 2.Relevance Trees, p. 288.

** The milk supply in India in 1969 was a case in point. Price advanced dramatically with supply.

One configuration of the Tamil Nadu food system is diagrammed in Figure 1, with key components expanded in Figures 2, 3, and 4. The diagram as shown is directed toward an understanding of the consumer subsystem as the dominant subsystem in the State. The State itself, as the entity embracing all of the systems, subsystems, and components under consideration, is indicated only by a label at points of connection. The graphic outline is designed to show the major input-output junctures (nodes) and interfaces between components of the overall system. Letter codes for identifying positions on the figures are in an expanding sequence from Figure 1 to 4. Some examples of trees or branches develop at dominant nodes such as: families Ⓑ, Districts Ⓐ, a business or industry Ⓒ, a Government ministry. Subordinate data sets are characterized by the next higher set to which they belong and are not capable of separate identity within the system. Some examples of such sets, connected at subordinate nodes, are: individuals within a family, expanded as in Figure 4 P Ⓓ; taluk or town as a political subunit of a panchayat union or District (Figure 2); products or services of a business Ⓒ, expanded in Figure 3; divisions or departments of a Government ministry, of an industry or a company.

Items in the data sets which tie families or individuals to the food system may be either identification characteristics or measurable characteristics include names, locations, social status (jati), economic status (income and property), and job, position, or profession as an index of both social and economic status; also sex and reproductive status, including nursing and weaned infants, pregnant and lactating women. Measurable characteristics include age, size (height and weight), morbidity (whether healthy or acute or chronically ill), nutritional status as indicated by measured extent to which needs are met, educational level, and those identification characteristics which can be ranked or directly quantified (income and jati rank, for example).

One of the valuable characteristics of the analytical methodology is the flexibility and ease with which portions of the system under analysis can be revamped to change subset relationships for the purpose of examining specific interrelationships at a different level of aggregation. An example is the comparative determination of degree of calorie need met on an individual basis, on a family basis by summation of individual member's needs, and on a family aggregate basis. In the first instance, a data subset decribing the individual is treated as independent, not as a subordinate subset

FIGURE 1 Food System Integration: Tamil Nadu Component Interrelations.

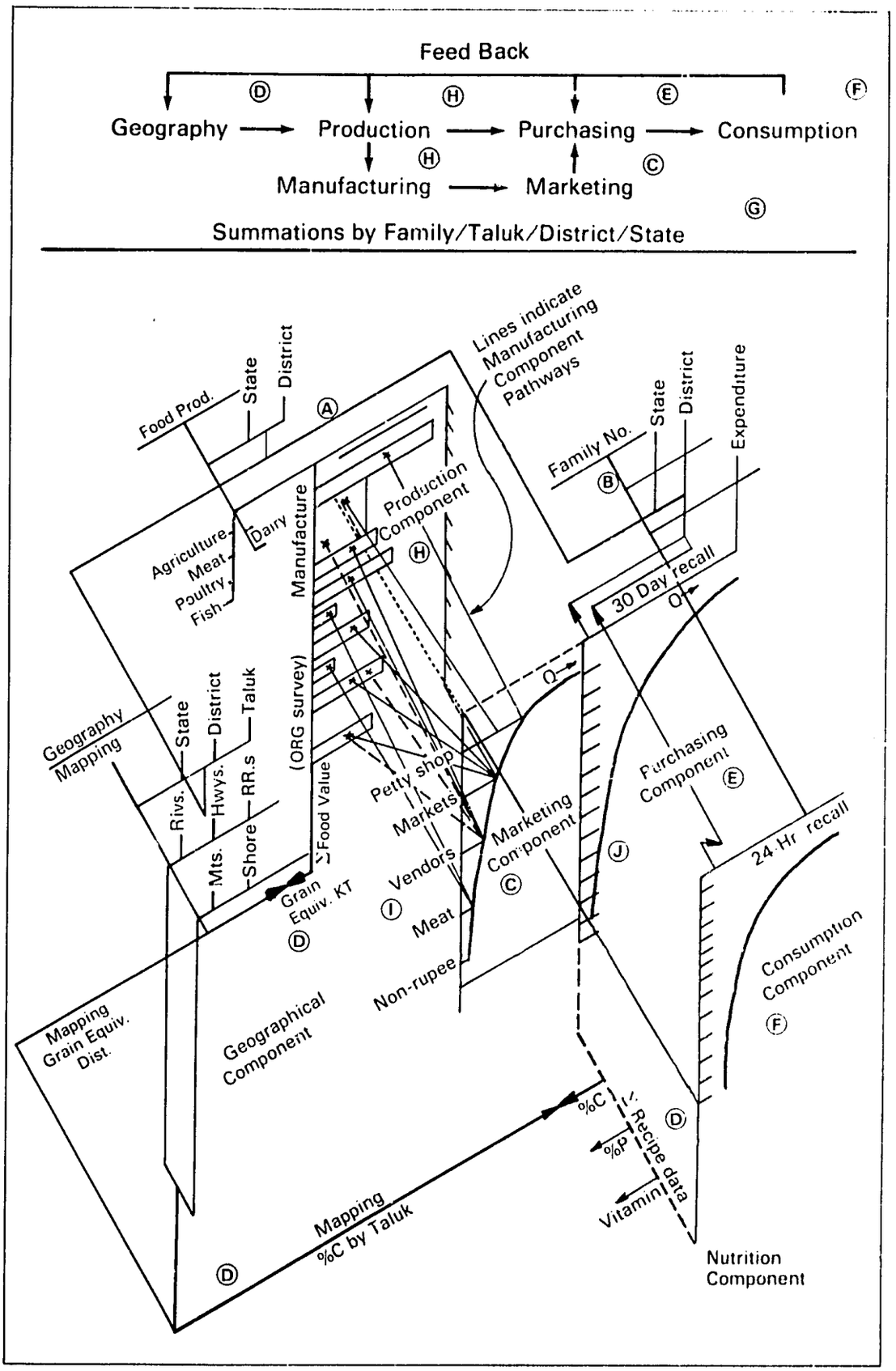
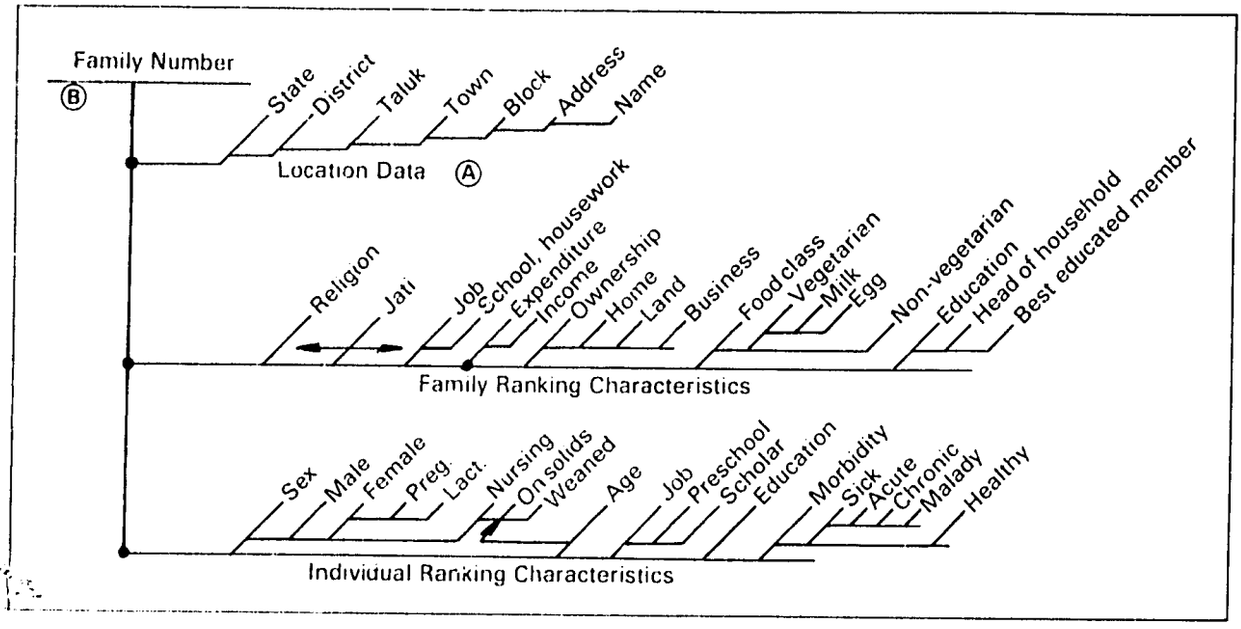


FIGURE 2 Identify Data Tree: Family and Individual.



of the family data set, in the second it clearly is, and in the third, the individual's data do not enter into the determination, which is based only on the total number of adult equivalents per family without regard to demographic composition.

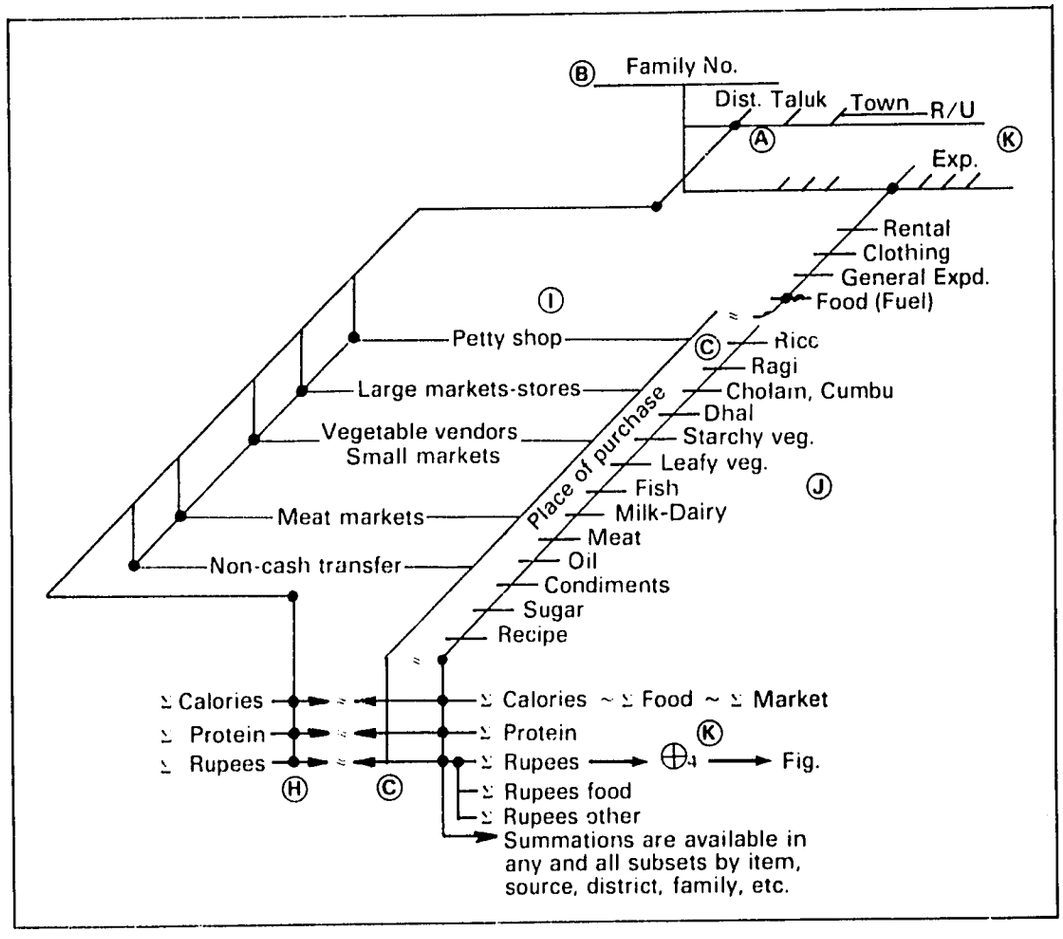
Example of Input and Output (I/0) - In a transit of the food system from agricultural production to the ultimate consumers there are many fully recursive cycles and subcycles. The output of each stage in the integrated system from the land to the consumer is the input to a succeeding stage.

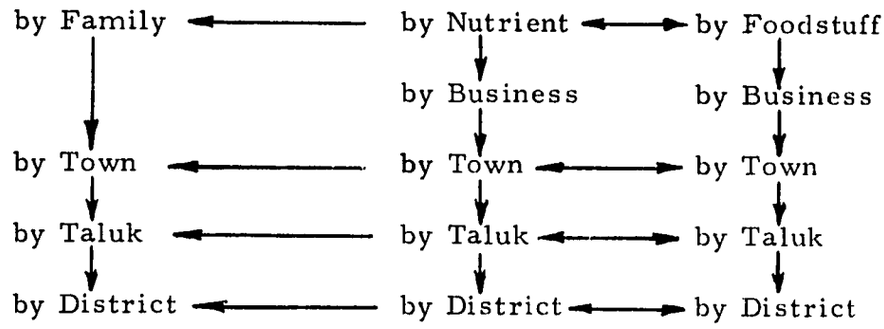
The elemental subcycle of the system is family food input/output. Food purchases by a family is input. See E in Figure 1. Expenditure for the food is an output (see C). Family food input and output is a summation of inputs and outputs for the individual members. Food consumed is an individual input F . The degree to which the food meets the individual's nutritional need is an output D . Total family consumption and average nutritional status are secondary inputs and outputs for the family (Σ F , Σ J and secondarily Σ H).

The summation of all family input/output within a given taluk or District is a measure of net consumption for that taluk or District. This may be compared to the net input/output from raw materials to products. See D , Figure 1, and the amplification in Figure 4. In Figure 4, the process from F to D follows foods eaten L , compared to recipe M , summed at N and then distributed to individuals P with final collection of data at D . In each figure, the Ds are comparable but they are at different levels of aggregation. Summation of net consumption for the State then provides a measure of the major input to the agricultural production subsystem, in terms of demand, thereby completing the overall input/output cycle for food production and consumption.

Summation of the food procurement/consumption system, up to the District level, may be diagrammed as below. At each level of aggregation the summation process may be extended (vertical arrows) or compared (horizontal arrows).

FIGURE 3 Food Purchase Patterns (30 Day Recall) by Family.





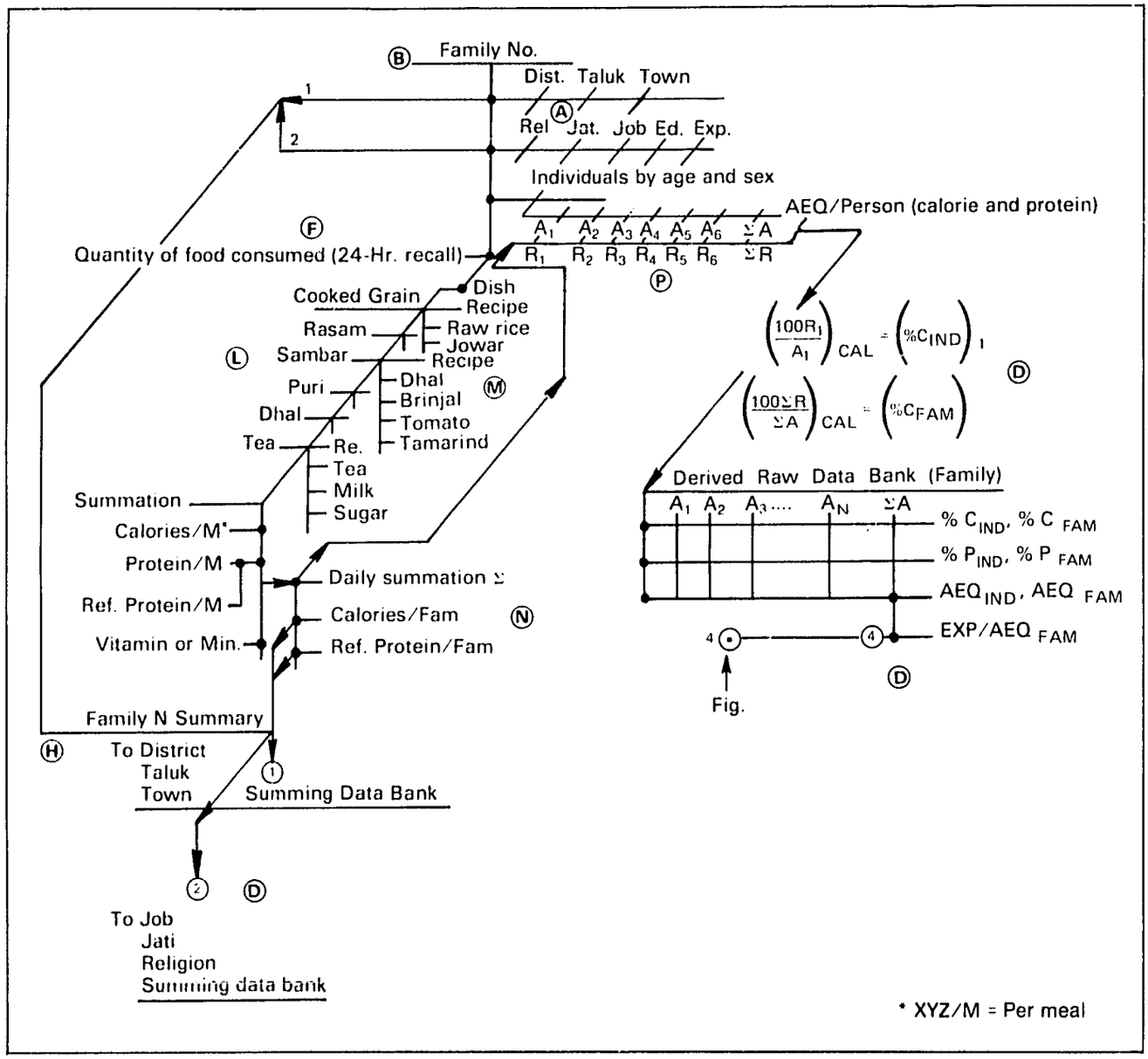
This expandable input/output cycle is only one of a large number that can be developed from the data bank and computer programs providing access to it. The data technology that has been developed permits each analyst to investigate those comparisons and pathways that are of interest to him.

SOURCES AND PREPARATION OF DATA

A substantial body of secondary data (i.e., derived from existing compilations of statistics) on the agricultural subsystem was available. Also an original, large-scale survey of food processing and distribution businesses was carried out in the course of the project. However, the primary objective —provision of guidance and a basis for decision-making in the planning of nutrition intervention programs— required a large mass of detailed, highly specific original data on the consumer subsystem. Consequently, most effort has been applied to data gathering, development of methodology and analysis of this one of the three major subsystems. Principal source of original data is the Consumer Expenditure and Food Habits Survey (usually referred to as simply the "food habits survey"), supplemented with data from a trial of a take-home food distribution scheme as a means of meeting nutritional goals for weaning children and younger preschoolers, those up to three years of age. Reference data trees for these two sources are those defining families in the food habits survey and the target children in the take-home test.

Critical Data Items in the Consumer Survey - To reach the goal of a food system analysis and evaluation based upon knowledge of food behavior and practices at the household and individual level, it was necessary to take special measures to insure accuracy and completeness of the data sets in three special categories: household identity, household food procurement and consumption, and food intakes of individuals.

FIGURE 4 Food Consumption Patterns (24 Hr. Recall) by Individual and Family.



1. Household identity, as illustrated on the first page of the consumer questionnaire, is a record of many data items, including location, demographic makeup, economic and social status, and cultural characteristics. Within the family group each member is identified by a serial number and sex, age, and reproductive status in the case of females. The identity data tree is diagrammed in Figure 2.

2. Food procurement and consumption were determined from a 30 day recall inventory of expenditures and goods procured, with emphasis on foodstuffs and how and where transfers took place. These data served as a check on family income (defined by ranges only) and also allowed correlation with the food distribution system (H). Questions were designed to provide a maximum of input/output data as, for example, in this series about the procurement of rice: When was the last time rice was obtained? How much was obtained? How much did it cost? From whom was it obtained? How long will it last? How often is such a purchase made? As a consequence of this line of questioning, it was possible to identify 27 procurement sources for foods in five main groupings ((I), Figure 1) and over 100 purchased products. The purchased products were sorted into 13 food groups, as shown in the extended data tree in Figure 3 (J). Summation by amounts of foods (deduced from price and total amount spent when necessary) and frequency of purchase provided a check of household food procurement against local production. This becomes one step toward confirming farm production statistics obtained as secondary data inputs.

3. Food intakes of individuals, determined by recall of kinds and amounts of food consumed by each member of the household during a preceding 24 hour period, provided the primary measure of family and individual nutritional status, and the only means of determining intrafamily food distribution. The recall was based upon foods as eaten ((L) Figure 4) with a notation on the quantities of the three or four most important ingredients (from the point of view of calorie and protein contributions) used in preparing the foods (M). Total quantity of each recipe prepared (N) and distribution of it to individual household members (P) was recorded. The data tree extension is shown in Figure 4.

From the problem aspect of data on food consumption, the most significant information gathered here was the recipe knowledge. A dish such as gruel varied from mixtures of one part grain and 20 parts of water (kangee) to three parts of grain, mostly ragi, and one part of water (kali). This is extreme variation but shows that many dishes having

common names may vary considerably in food value. Even coffee, a major or only source of milk in diets of many of the poor, varied from one part of milk to two parts of coffee (as served) to one part of milk to 20 parts of coffee. Usually as much milk as can be afforded is used, since coffee is a customary drink for children. Such points as the recipe approach represent a major difference between this survey and most other other food consumption surveys on a broad scale. The emphasis here on intrafamily food distribution reinforces utility of the data for identifying target groups. It was clearly shown, for example, that almost no one, rich or poor, educated or illiterate, feeds a weaning child well.

Additional household data gathered covered the subjects of morbidity, mortality (infant mortality in particular), and effects of circumstance, religious and cultural influences, and hearsay knowledge of nutrition on family food practices. In the final quarter of the four quarter survey, questions regarding attitudes toward family size and family planning were included. Much of these data have yet to be incorporated in the analysis.

Data Preparation - Satisfactory preparation and preliminary processing of field data for the data bank and subsequent analytical operations has been one of the most time consuming and intellectually demanding tasks of the entire project. The problems, procedures, and highly efficient and flexible system finally developed have been reported in detail in Chapter IV.

DATA ANALYSIS METHODS AND OPERATIONS

Normalization of Critical Data

This is the first analytical step. In this project, normalization meant conversion of foods to nutrient equivalent, and conversion of persons into units of nutrient need. In order to establish a basis from which to measure, a typical adult male is defined as the need unit requiring one adult equivalent of protein (AEQP) or calories (AEQC)* (AEQC = AEQ). A standard

* Although calories and protein are the primary determinants of quantity and quality of diet, the AEQ normalization concept could also be applied to other problem nutrients, vitamin A for example, if the needed data were procured.

protein requirement, defined according to height and weight by established European norms, was adopted.* Data on stature from the Nutrition Atlas of India** were used, and a stress factor was applied. Purpose of the stress factor is to compensate for low birth weight (and also stresses of chronic malnutrition and disease), based on indications that persons born underweight tend throughout life to utilize food somewhat less efficiently than others who were not similarly stressed. For Tamil Nadu, a stress factor 1.2 was assumed, resulting in a set of protein requirements equal to 90 percent of the European norms. Standard calorie requirements were defined in a similar manner.

Next, data on foods were analyzed to determine the nutritional value of each item. Estimation of quality of the protein in a food was based upon the modified FAO Amino Acid Reference Pattern for utilizable protein (WHO Tech. Rept. Ser., No. 301), which specifies 4.2% lysine, 3.1% sulfur containing amino acids, 2.8% threonine, and 1.0% trytophan. (In the order shown, these four essential amino acids are most likely to be limiting to utilization of protein from Indian foods. Using data from "Nutritive Value of Indian Foods" (Nat. Inst. Nutrition, Hyderabad, 1971), content of calories and reference protein in each listed food or recipe was calculated. In using these data to calculate nutritional status of persons an important assumption was stipulated: that protein would be utilized as protein, not as respiratory substrate, to the extent of the root mean between the actual level and 7 percent reference protein calories as the standard. This level of utilizable protein is considered adequate for the needs of 98 percent of all individuals receiving an adequate number of calories.***

Nutritional status of families was calculated on two bases, from summation of calories and protein in the 24 hour recall of food intakes by the individual family members, and as a confirmatory check, from food availabilities as determined from the 30 day recall inventory of household procurement. Individual nutritional status was calculated only from

---

* WHO Technical Report Series, No. 301.

** C. Gopalan and K. Vijava Raghavan, National Institute of Nutrition, Indian Council of Medical Research, Hyderabad, 1971.

*** WHO Technical Report Series, No. 301.

the 24 hour recall record. Three indexes of nutritional status of families and individuals were calculated: percentage of calorie need fulfilled (%C), percentage of protein need fulfilled (%P), and as a measure of nutritional quality of the food eaten (independent of dietary adequacy in terms of consumption of enough food to meet calorie needs), the ratio %P/%C.*

Based upon family size as defined by the calorie need in adult equivalents (AEQ), two additional indexes were calculated: (1) family expenditure per AEQ and (2) AEQ/Family (used as the reciprocal).

The calculated indexes of nutritional status appear as the final output of the food consumption data tree, Figure 4.

Rank Ordering

In the introduction to this chapter, reference was made to the analyst's need for "something more revealing than an equation derived for a developed society and more compact than the raw data file." This need has been met by development of a procedure which still permits regression analysis but also furnishes the analyst and decision maker a visualization of the data dispersed to a degree far beyond the conventional three dimensional representation. In addition, simple tools to test level of significance and degree of correlation that were already available could be extended from two dimensions to more than two. First step in the process requires rank ordering of all variables to be considered. With some variables the task is easy: (a) income or expenditure: rich to poor; (b) family size: small to large; (c) education: much to none; (d) extent to which nutrition requirements are met: % P high to low, %C high to low.

Other variables are not amenable to any such simple rule (a) District; (b) rural versus urban location; (c) Caste; (d) occupation; (e) others such as age, sex, reproductive status, family position. In each of these, indirect measures are needed to permit the data to be rank ordered.

* The ratio was almost always above one, confirming the growing general opinion that, for South India, if calorie needs are met by consumption of enough of the customary foods, protein needs, on average, will be adequately fulfilled.

(a) District, first ranked alphabetically by name, was reordered on the basis of median family nutritional level attained in the sample population in the District. Values ranked were the averages of the median indexes (partile position) for %C and %P. The net result is a ranking of the quantity of utilizable protein calories, a measure of quality in the diet. This index, for example, shows that at 90 percent of calorie needs met, the average person's need for protein is fully met. Districts were also ranked according to total amounts of food calories produced within each one.

(b) Rural/urban split is similarly ranked. (In this case, the unique position of Madras City, 100% urban, distorts the data generally).

(c) Caste (jati) was ranked independently by two scholars having intimate knowledge of the social structure of Tamil Nadu. Each jati was positioned on an arbitrary 1 to 5 scale (Table 1). The ranking was complex, since questions about caste or jati also evoked responses that reflected social or economic status but were outside of the traditional social system. For example, some respondents answered with the name of a non-Hindu religion, such as Protestant, Jain, Catholic, Muslim; others gave place of origin, e.g., Maharashtra, and some responded with with their family names. As a result, an attempt was made to define a rank based upon a broader understanding of relative social and economic position. Final Caste (jati) rankings correlated well with District rankings based upon production or consumption patterns. Correlation was inverse, indicating that the quantity of production was a function mainly of the number of labor and service jatis, as might be expected in a labor intensive society.

(d) Occupation ranked in much the same manner as caste.

(e) Combined socioeconomic ranking: On the assumption that one of socially higher jati would probably have a more prestigious job, receive a better education and perhaps earn a higher income, the ranking was devised using jati, occupation, education of head of household, and income as parameters. The final socioeconomic index was the geometric mean of the four indexes:

$$SI = \sqrt[4]{(\text{jati rank}) \times (\text{job rank}) \times (\text{education rank}) \times (\text{expenditure rank})}$$

JATI CODE LIST FOR TABLE 1 *

MDL	Mudaliyar	KSU	Kosava
NDR	Nadar	KNG	Karunigar
BHM	Brahmin	RJU	Raja
SWT	Sowrashtra	VLM	Vallunar
VAR	Vetaikaran	KMR	Kammavar
KKP	Kanakkar	APU	Annupu
RAO	Rao	NTM	Nathanmar
VSA	Vaishnava, Iyengar	AWL	Agammal
IYR	Iyer	PSI	Poosari
BDH	Badaga	MRV	Maravan, Kallar
MRC	Maraikar	RUT	Rawuthus
SGH	Sengutha	LBI	Labbai
AMR	Agamudaiar	MUR	Muthuraja
PNK	Panicker	DVM	Devanmar
LYT	Lingayat	ABL	Ambalakar
GDR	Gounder	BBR	Barber
NDU	Naidu	SEV	Servai
VLR	Vellalar	DBI	Dhobi
KUL	Kallar	BYR	Boyar
PTR	Pathan	SLR	Saliar
NNR	Nainar	CHL	Chakkali
JAN	Jain	KDR	Kundumbar
PAT	Pattar	MYI	Malayali
PTN	Pattani	DGR	Devangar
KMM	Kollar, Kammar	SFI	Safi
KMA	Kamma Reddi	PNR	Pannaiar
GWA	Gowda	MSY	Maistry
NCK	Naicker	VYN	Vaidyan
CHY	Chetty	MEN	Menon
PDC	Padayatchi	SAT	Sait
PLI	Pillai	SVR	Servar
ASR	Acari	SRM	Sairam
EDY	Idaiyar	YOW	Yokeswarar
UYR	Uddaiyar	HJN	Harijan
RDI	Reddi	ADR	Adi-Dravidar
ROC	Roman Catholic	PLR	Pallar
NAR	Nair	ART	Arunthathi
SHK	Sheik	PYR	Parayar
VKL	Vokkaliga	SCT	Scheduled Tribe
PST	Protestant	ODR	Odder
RJT	Rajput	SMN	Semman
KHY	Kshatriya		

* "Ranking of Social Status and Occupation in TN" Michael Moffatt, TN Nutrition Study Vol. II-B, pp 113-152 (1973).

TABLE 1

JATI RANK ORDER INDEX

1		2		3		4		5	
No.	Code	No.	Code	No.	Code	No.	Code	No.	Code
148	MDL	236	GDR	127	NCK	105	MRV	310	HJN
127	NDR	97	NDU	125	CHY	31	ROU	50	ADR
133	BHM (101)	74	VLR	100	PDC	29	LBI	10	PLR
	SWT 13	10	KUL	97	PLI	26	MUR	6	KMB
	VAR 7	5	PTR	85	ASR	26	DDM	5	ART
	KKP 4	5	NNR	72	EDY	25	ABL=AMB	4	PYR
	RAO 3	5	JAN	49	UYR	20	BBR	3	SCT, OOR
	VSA 3	4	PAT	38	RDI	19	SEV	2	SMN
	IYR 2	2	PTN	36	ROC	18	DBI	1	KVS, OUD, BIY, GDO, KML
38	BDH	1	KMM	22	NAR	17	BYR		
12	MRC	1	KMA	10	SHK	14	SLR		
12	SGH	2	GWA	10	VKL	11	CHL		
12	AMR			6	PST	7	KDR		
3	PNK			4	RJT	6	MYI		
1	LYT			4	KHY	5	GPG, DGR		
				3	KSU	4	SFI, DUM, PNR		
				2	KNG	3	DVR, GMI		
				2	RJU	2	MSY, VYN, MEN, PIK, SAT, SVR		
				1	VLM, KMR	1	KNA, SMY, PTU, PTV, MNV, SNI, SRM, YOW, EZL, KTU, MTI, YGY, YRL, DDL, BRK		
				1	APU, NTM				
				1	AWL, PSI				
Total 486		442		798		429		398	

The Probability Grid

After the various rank order indexes have been obtained they may be assembled in ordered arrays of two or more levels of complexity. Although any boundary values may be used, it has been found expedient to use definite partile boundaries wherever possible. In practice, we have found pentile distribution (division of data into five ranked groups, each containing 20 percent of the total sample) to be most useful.

As an example, the distribution of individuals in a pentile grid of ($\%C_I$) percentage of individual calorie need met versus ($\%C_F$) percentage of family calorie need met is shown in Figure 5. The grid appears similar to a contingency table in analysis of variance except that totals for all rows and all columns are equal.* Frequency is constant, group interval is permitted to float.

The singular properties of such a grid make analysis simple and straightforward:

1. Data may only be symmetrically distributed about the diagonals. The requirement that each row and column total the same percentage of the grand total precludes any but symmetrical distribution.

2. Correlation of any two variables is therefore along the axis of maximum data concentration. A modified tetrachoric** correlation coefficient will usually define the degree of correlation adequately, but with low values. (For even numbered partile distributions the coefficient would be precise.) Refer to Figure 6, which shows Figure 5

---

* Approximately. The need to estimate the partile boundaries to rank families by summation of individuals caused some error in the exact pentile definitions of the $\%C_F$ boundaries.

** Two variables dichotomized, i.e., each separated into just two classifications. (see Introduction to Statistical Data Processing, T. D. Sterling and S. V. Pollack, Prentice-Hall, Inc., 1968.. pp. 424-426.

FIGURE 5

DISTRIBUTION GRID FOR INDIVIDUALS
Percentage Family Calorie Need Met
versus
Percentage Individual Calorie Need Met

$\%C_{FI}$ \ $\%C_I$ Pentile	1	2	3	4	5	Total
5	49	61	170	620	1776	2676
4	111	271	652	1009	565	2608
3	247	631	851	622	158	2509
2	580	1003	675	246	57	2561
1	1598	624	241	93	33	2589
Total	2585	2590	2589	2590	2589	12943

with the four cells in each corner summed. The modified tetrachoric correlation coefficient is

$$r = \sin\left[90^{\circ}\ \frac{(T_2 - T_1) - (B_2 - B_1)}{(\text{Total} - C)}\right]$$

$$r = \sin\left[90^{\circ}\ \frac{(3970 - 492) - (429 - 3805)}{12943 - 851}\right] = 0.78$$

3. The degree of significance of the data may also be determined using the chi squared (χ^2) test. This statistic is defined as

$$\chi^2 = \Sigma\left(\frac{|A - P| - 0.5}{P}\right)^2$$

where 0.5 is the Yates correction factor, A the actual data value, and P the probable data value.

In the example in Figure 5, there are (12943 total events)/25 cells or 518 chances per cell for a purely random distribution. Then

$$\chi^2 = \Sigma\left(\frac{|A - 518| - 0.5}{518}\right)^2 = 9908$$

There are (5 minus 1) x (5 minus 1), or 16 degrees of freedom. Using this value to enter a χ^2 table (Table 2) we find that there is no reasonable chance that these data are random, since 9908 lies about 135 standard deviations from the mean.

4. The degree of dispersion may be determined using the standard error of the estimate or second moment about the major axis.* Again, the simplicity of mapping makes it possible to define the index easily.

---

* This is moment of inertia about a line. The concept may be extended to N space using vector mathematics to project all data to the major axis in N space.

FIGURE 6

DATA FROM FIGURE 5 OUTLINED FOR MODIFIED TETRACHORIC CORRELATION COEFFICIENT CALCULATION

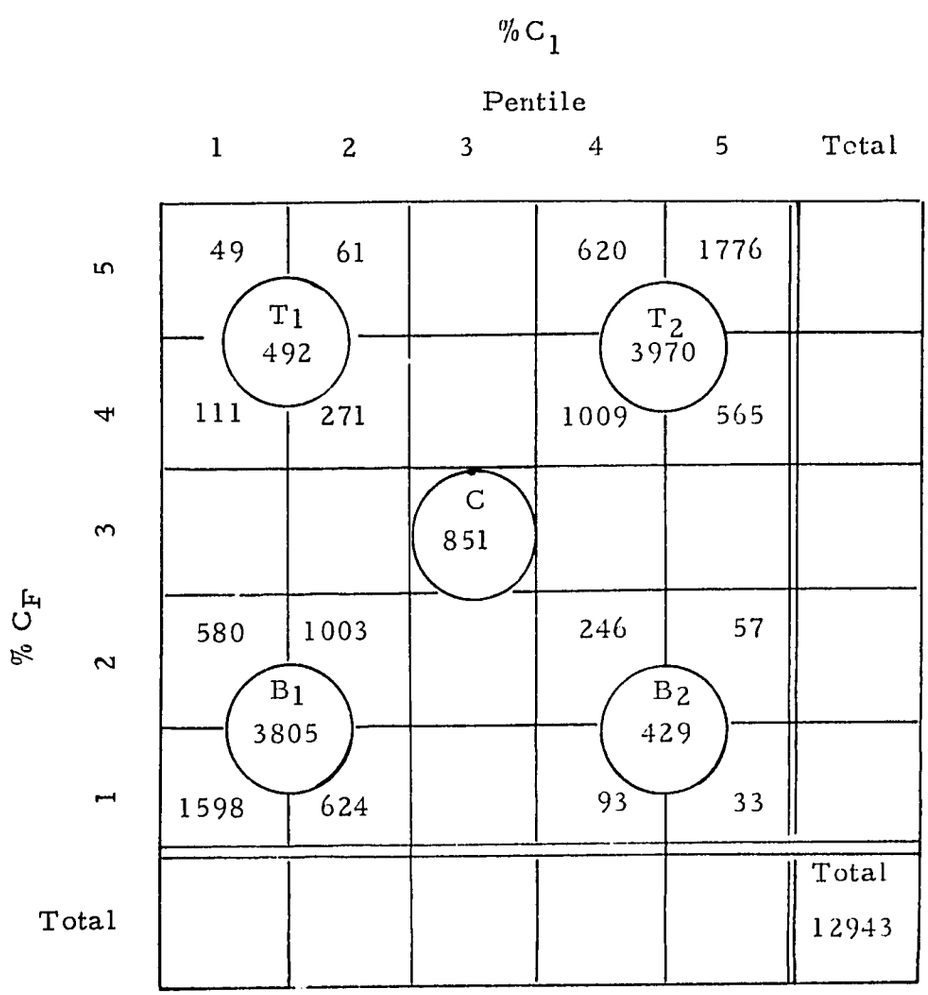


TABLE 2

SELECTED PERCENTILE VALUES OF THE CHI-SQUARE DISTRIBUTION

f	Percent Larger Than 0.1	0.05	0.01
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.73
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
20	28.41	31.41	37.57
30	40.26	43.77	50.89
40	51.81	55.76	63.69
50	63.17	67.50	76.15
60	74.40	79.08	88.38
70	85.53	90.53	100.42
80	96.58	101.88	112.35
90	107.57	113.14	124.12
100	118.50	124.34	135.81

For 30 degrees of freedom or more the number of standard deviations from the mean at which a Chi Square (χ^2) lies may be estimated by the formula:

$$Z = \sqrt{2\chi^2} - \sqrt{2f - 1}$$

Where Z = Number of standard deviations from the mean

χ^2 = Chi Squared statistic

f = Degrees of freedom

e.g., at .01% and 70 = f $\chi^2 = 100.42$

$$Z = \sqrt{2x100.42} - \sqrt{2x70-1} = 14.17 - 11.79 = 2.38$$

and $Z = 2.326$ is correct.

$$M_t S^2 = M_{ij} d_{ij}^2 \;/\; \Sigma M_{ij} ,$$

where S = standard error of the estimate to grid diagonal

$d_{ij} = i - j$

= Spearman's "D" if all variables are in partiles

M_{ij} = number of points in box (i, j)

M_t = total number of points where i, j are partile coordinates.

For pentile distributions the resulting value of S^2 will vary from 0.0 to 8.0, where 0.0 indicates perfect correlation or no dispersion along the chosen axis, 4.0 indicates total randomness or complete dispersion, and 8.0 indicates perfect correlation along the orthogonal axis.* In this example,

$$S^2 = \frac{14496}{12943} = 1.12$$

Standard Deviation: In a partile system each variable is distributed randomly along its major axis. Each is isonormal or collectively homonormal. As a consequence,

$$\sigma_x = \sigma_y = \sigma_n \text{ and } \sigma_x^2 = \sigma_y^2 = \sigma_x \sigma_y, \text{ etc.}$$

---

*For other partile distributions, maximum value = $\frac{(N - 1)(N + 1)}{3}$.

Mid Level = random = $\frac{(N - 1)(N + 1)}{6}$, where N = N'tile.

Each variable pair is calculated then as the net random distribution across the grid in two dimensions.

$$M_t \sigma^2 = \sum M_{ij} d_{ij}^2 = \sum \frac{M_i M_j}{M_t} d_{ij}^2$$

where $M_i = M_j = M_t/p$,

M_t = total number in grid,

M_i = number in row i,

M_j = number in column j,

$d_{ij} = i - j$

p = number of partiles.

Thus, $$\sigma^2 = \sum \frac{(M_t/p)^2}{M_t^2} d_{ij}^2 = \sum d_{ij}^2 / p^2$$

A consequence of this relationship is that the regression coefficients (r), or better, the correlation indexes of X on Y, Y on X, or X on Z, etc., are identical. One coefficient fits all variables used in a single regression, provided all are partiled similarly.

The Correlation Index: $\rho^2 = 1 - S^2/\sigma^2$

In the example, $\rho = \sqrt{1 - 1.12/4.0} = 0.85$. The index ρ will be $> r$. It will approach the correlation coefficient as the functions approach linearity. Here ρ is in good agreement with the tetrachoric correlation coefficient calculated earlier.

The Regression Equation: Each function analyzed along the probability grid is of the form F (x), F (y), etc. --

$$F(x) = m_x(f(x) - f(x)_{50}),$$

where f(x) = function of distribution of x along a probability axis,

m_x = slope of f(x)

$f(x)_{50}$ = the value of f(x) at the 50th percentile,

and $F(x) = \rho(\sigma_x/\sigma_y)F(y) = \rho F(y)$. (note $\sigma_x/\sigma_y = 1$)

$$m_x(f(x) - f(x)_{50}) = \rho m_x(f(y) - f(y)_{50})$$

and $f(x) = = \rho(m_y/m_x)f(y) - \left[\rho(m_y/m_x)f(y)_{50} - f(x)_{50}\right]$.

Similarly, $f(y) = \rho(m/m)f(x) - \left[\rho(m/m)f(x)_{50} - f(y)_{50}\right]$.

In an ordinary nonlinear regression where f(x) = bf(y) + a, the regression coefficients would be

$$r_{x,y} = \rho(m_y/m_x)/b, \quad a_{x,y} = f(x)_{50} - \rho(m_y/m_x)f(x)_{50},$$

or if $\sigma x \neq \sigma y$,

$$r_{x,y} = \rho(m_y\sigma_y/m_x\sigma_x)/b, \quad a_{x,y} = f(x)_{50} - \rho(m_y\sigma_y/m_x\sigma_x)f(y)_{50}.$$

The Probability Function of a Variable:

This operation will be reminiscent of the use of the ogive or cumulative distribution curve coupled with probit analysis. It is during this phase of analysis that rectification of the data takes place. It is the only time that each variable is considered singly, unconfounded by anything else. This

is a singularly useful function of one-variable distribution analysis. Quantification of a function (problem) by this tool alone as a means to decision-making will almost always define the impossible option since it defines the limits, and thereby totally ignores modification by other variables.

The two most common probability plotting grids available are linear versus probability (linear normal) and logarithmic versus probability (log normal). Together they fit a vast number of natural functions. Following are representative functions from the third round consumer survey data, plotted on log normal charts:

1. Percentage of Calorie Needs Met - The basic reference function for the consumer survey is the percentage of nutritional need met as a function of calories consumed, divided by need (AEQ) of the individual. The data (Table 3) are plotted in Figure 7. The accumulated vigintile (20 partiles) boundary values are used. The data are substantially log normal except at the very low end. This dip is most likely due to persons who missed meals with the family and did not account for them elsewhere.* Ignoring the dip, a calculated value for P_c in terms of f(%C) is

$$P_c = m_c \text{ Log } (\%C) - m_c \text{ Log } (\%C)_{50} = 3.06 \text{ (Log}(\%C) - 1.89),$$

where P_c = probability function of % C,

m_c = slope or tangent α = 3.06

$\text{Log}(\%C)_{50}$ = value of Log(%C) at the intersection of the 50th percentile axis

= 1.89, and

%C = percentage fulfillment of individual calorie need.

---

* A default value was considered to circumvent this type of error, but has not yet been included in the program. The default value would be the assignment, to the person who missed a meal, of an amount (%C) equivalent to the average family consumption at the missed meal.

TABLE 3

PERCENTAGE CALORIE NEED MET (%C), INDIVIDUAL

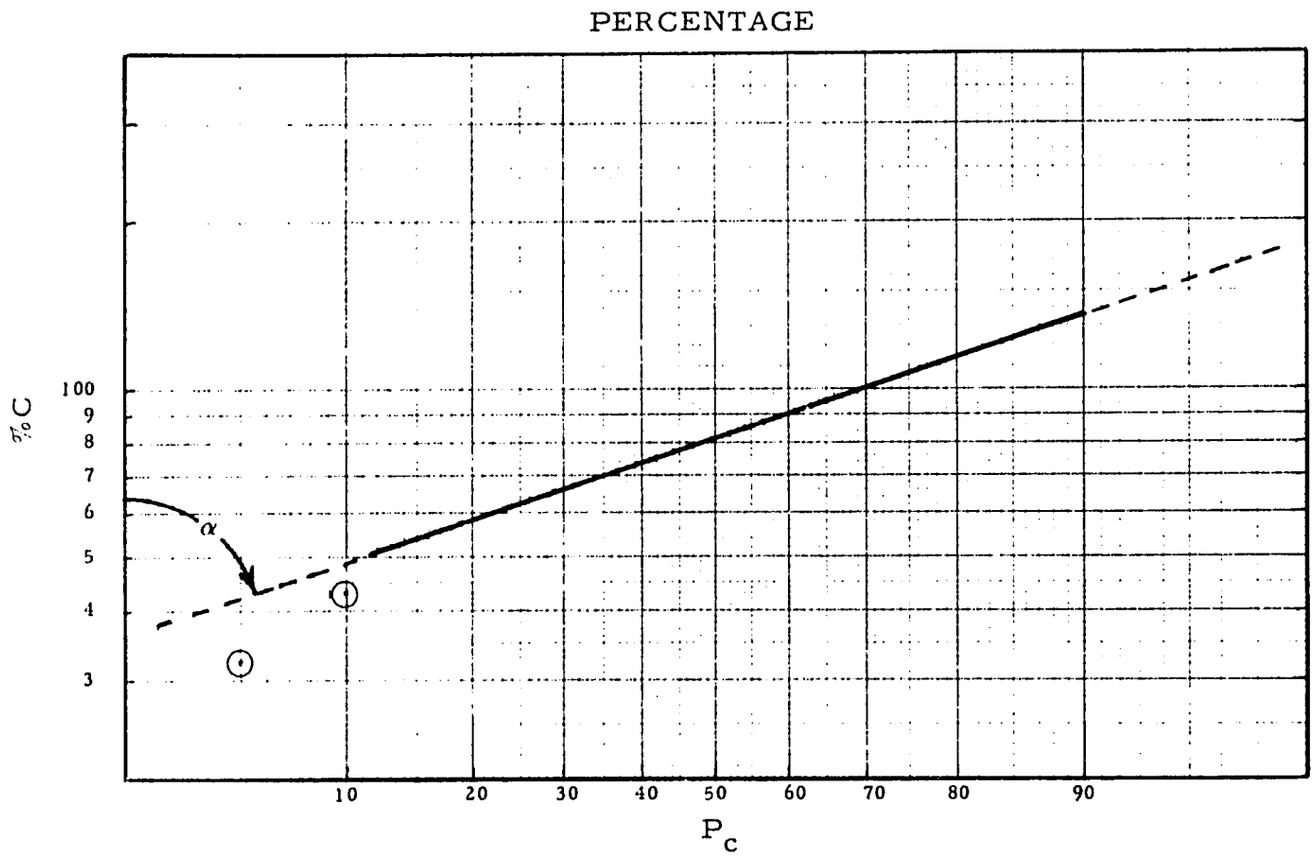
% C *	Rank **	Pentile
33	694	1
44	1388	1
51	2082	1
56	2777	1
61	3471	2
65	4165	2
69	4860	2
73	5554	2
76	6248	3
80	6943	3
84	7637	3
88	8331	3
93	9025	4
97	9720	4
103	10414	4
109	11108	4
117	11803	5
128	12497	5
147	13191	5

* %C - is the percentage of calorie need met by individuals based upon a 24 hour recall of food consumption.

** Rank - is the number of individuals at or below the designated level.

FIGURE 7

PERCENTAGE CALORIE NEED MET (%C), INDIVIDUAL : PROBABILITY FUNCTION (P_c)



2. Degree of Fulfillment of Protein Need by Individual, %P - This function is plotted similarly to %C and is shown in Table 4 and Figure 8. The same low level dip occurs.

$$P_p = m_p \text{ Log } (\%P) - m_p \text{ Log}(\%P)_{50} = 2.96 \text{ (Log}(\%P) - 1.97),$$

where P_p = probability function of %P, and

%P = percentage fulfillment of individual reference protein need.

3. Family Size Defined on a Nutritional Need Basis - It was assumed that the factor, calorie consumption versus requirement, most broadly represented the nutritional status of individuals, could be related to social and economic indexes, and was a useful index for the reflection of other nutritional data. Therefore, family size is defined as the number of calorie adult equivalents per family (A = AEQ/FAM). The data from Table 5 are plotted in Figure 9. The sharp break just below the 15 percent level in curves 1 and 2 indicates a step downward from more than one adult per family to just one adult per family. Ignoring the break, the function for curve 1 is defined as:

$$P_a = m_a \text{ Log } A - m_a \text{ Log } A_{50} = 2.85(\text{Log } A - 0.600),$$

where P_a = probability function for family size, based on ordering rank number, and

A = AEQ/FAM = calorie adult equivalents per family.

For curve 2 the function is

$$P_{ai} = m_{ai} \text{ Log } A - m_{ai} \text{ Log } A_{50} = 2.87 \text{ (Log } A - 0.708),$$

where P_{ai} is the probability function for family size based upon ordering by individual. It takes many more small families than larger ones to add up to 20 percent of the population. It is necessary to use this ranking any time the data are analyzed by individual.

4. Expenditure - In order to place expenditure (used as a substitute for income level) on a basis to reflect probable nutrition level, the function expenditure per AEQ per family (on a family basis (EXP/AEQ) was defined and rank ordered. Data for this function are illustrated in Table 6 and Figure 10. At high income (expenditure) levels, the data are starting to diverge rapidly (indicator of discretionary income).

TABLE 4

PERCENTAGE PROTEIN NEED MET (% P), INDIVIDUAL

% P *	Rank **	Pentile
42	694	1
54	1388	1
61	2082	1
67	2777	1
72	3471	2
77	4165	2
81	4860	2
86	5554	2
90	6248	3
94	6943	3
99	7637	3
104	8331	3
109	9025	4
115	9720	4
122	10414	4
131	11108	4
141	11803	5
157	12497	5
184	13191	5

* % P - is the percentage of protein need met of individuals based upon a 24 hour recall of food consumption.

** Rank - is the number of individuals at or below the designated level.

FIGURE 8

PERCENTAGE PROTEIN NEED MET (%P), INDIVIDUAL: PROBABILITY FUNCTION (P_p)

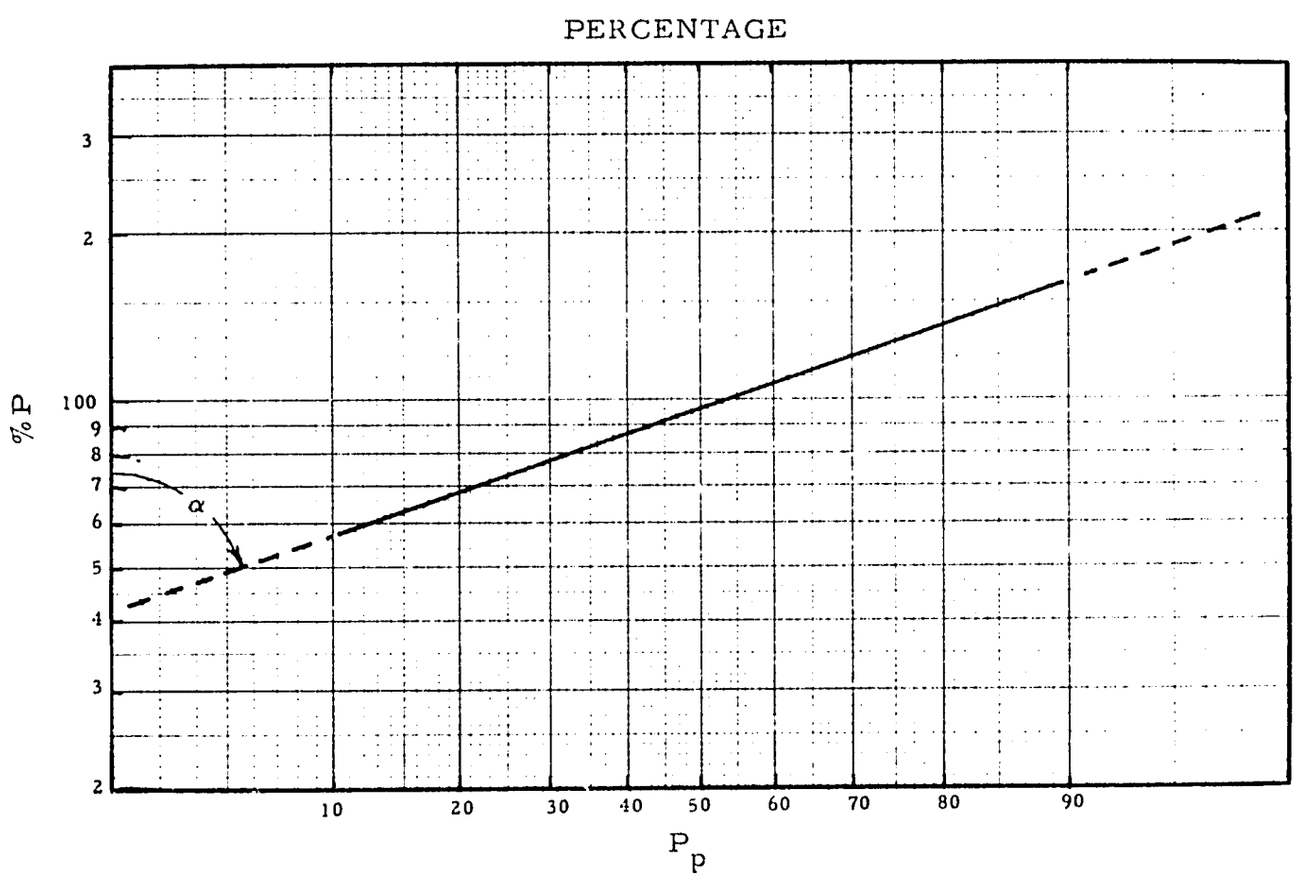


TABLE 5

FAMILY SIZE AS DEFINED BY NUMBER OF CALORIE ADULT EQUIVALENTS PER FAMILY AEQ/FAM

AEQ/FAM*	Rank[1]	Partile	Rank[2] Family No.	By Individual
1.70	124	1		
2.05	248	1		
2.58	372	1		
2.80	496	1		
3.01	621	2		
3.21	745	2	847	20% of individuals
3.52	869	2		
3.70	993	2		
3.84	1117	3		
4.04	1242	3		
4.25	1366	3	1391	40% of individuals
4.50	1490	3		
4.76	1614	4		
5.05	1738	4	1833	60% of individuals
5.34	1863	4		
5.72	1987	4		
6.19	2111	5	2198	80% of individuals
6.83	2235	5		
7.71	2359	5		

* AEQ/FAM - is the net calorie AEQ per family for a 24 hour period. This is the basis of family size used throughout the report. (in AEQ)

Rank[1] - Rank order by 20ths (vigintiles) based upon family rank.

Rank[2] - Rank of family which cumulatively includes the designated percentage of individuals.

FIGURE 9

NUMBER OF ADULT EQUIVALENTS PER FAMILY (AEQ/FAM)
PROBABILITY FUNCTION (P_a)

AEQ/FAM vs P_a FAMILY BASIS (1)
and
AEQ/FAM vs P_{ai} INDIVIDUAL BASIS (2)

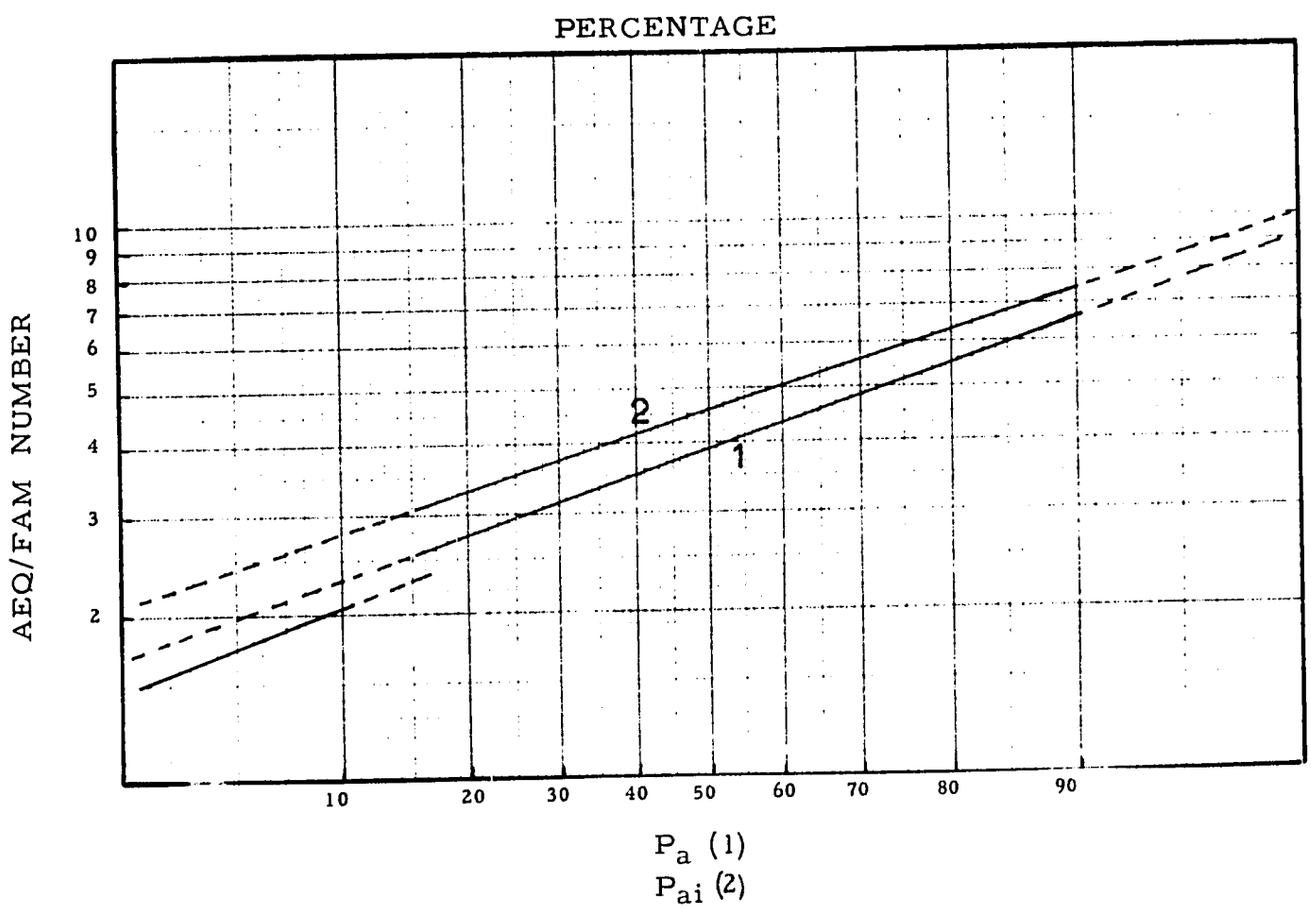


TABLE 6

MONTHLY EXPENDITURE PER ADULT EQUIVALENT
EXP/AEQ

*EXP/AEQ	Rank[1]	Pentile	Rank[2] Family No.	By Individual
20	124	1		
24	248	1		
27	372	1	423	20% of individuals
30	496	1		
32	621	2		
35	745	2		
37	869	2	896	40% of individuals
39	993	2		
42	1117	3		
45	1242	3		
48	1366	3	1395	60% of individuals
51	1490	3		
56	1614	4		
60	1738	4		
65	1863	4	1905	80% of individuals
71	1987	4		
81	2111	5		
97	2235	5		
128	2359	5		

* EXP/AEQ - is the ratio of monthly family expenditure to daily family calorie AEQ (in rupees month/AEQ)

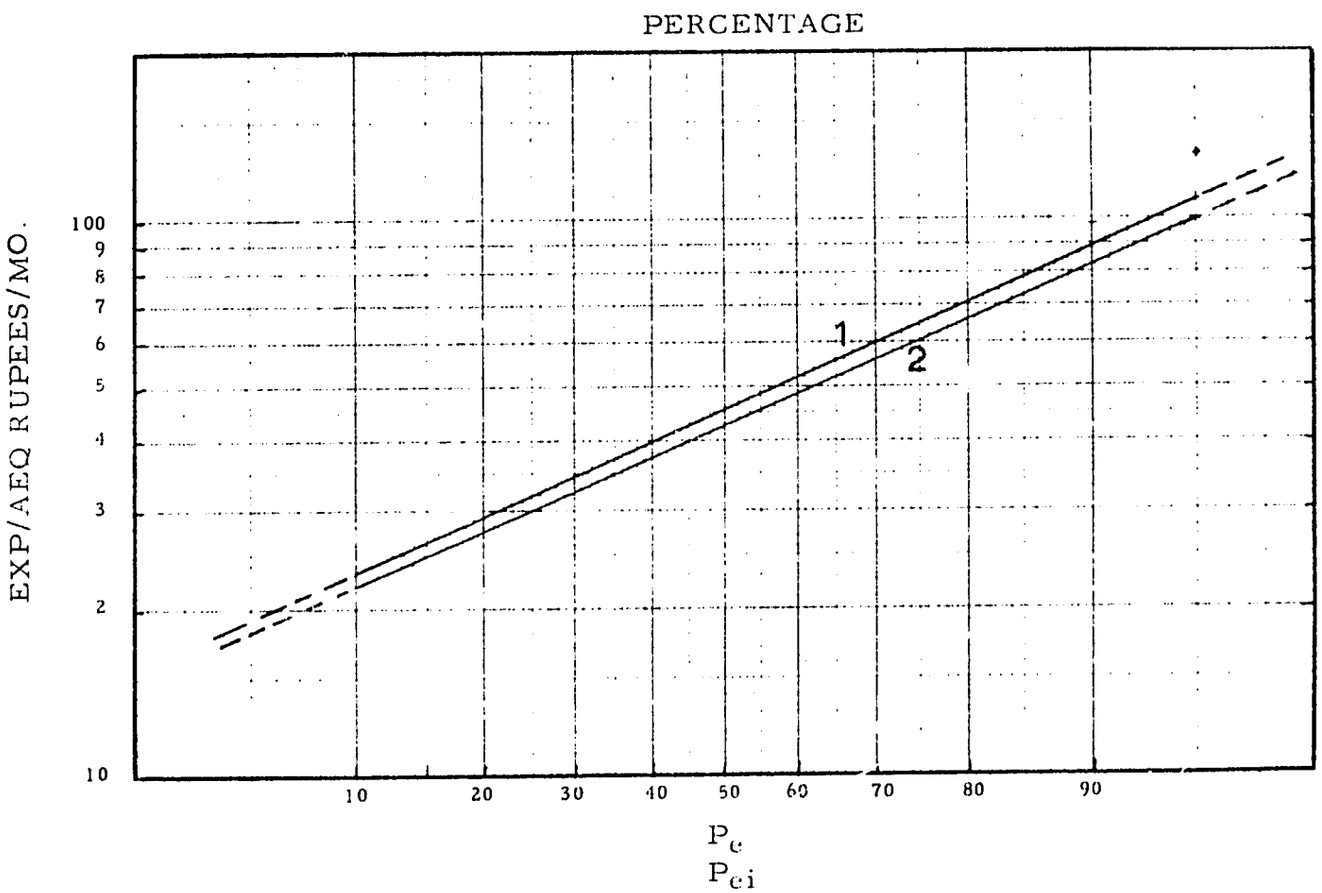
Rank[1] - EXP/AEQ Ranked by family to family partile boundaries.

Rank[2] - EXP/AEQ Ranked by cumulative individuals to individual partile boundaries.

FIGURE 10

MONTHLY EXPENDITURE PER FAMILY CALORIE AEQ (EXP/AEQ)
PROBABILITY FUNCTION (P_e)

EXP/AEQ vs P_e FAMILY BASIS (1)
EXP/AEQ vs P_{ei} INDIVIDUAL BASIS (2)



Ranked by family (curve 1)

$$P_e = m_e \text{ Log } E - m_e \text{ Log } E_{50}$$

$$P_e = 1.89 (\text{Log } E - 1.653)$$

where E = Expenditure in Rs per month per calorie AEQ/family

= EXP/AEQ

Ranked by individual (curve 2)

$$P_{ei} = 1.91 (\text{Log } E - 1.628)$$

where E_i = EXP/AEQ ranked by individual.

5. The Socioeconomic Index (SI) - This index is an example of a closed function which, as defined, has an absolute upper and lower limit. Rectifying such a function into an infinite field is most easily done by re-defining it hyperbolically. The function SI is so defined that it has an upper limit of 20 and a lower limit of $\sqrt{5}$. The rectified form of the function is:

$$SI = f(SE) = \frac{SE - L}{U - SE}$$

where L = lower limit of SE = $\sqrt{5}$
U = upper limit of SE = 20
SE = socioeconomic rank
SI = socioeconomic index

This new function, using data from Table 7, is plotted in Figure 11 and

$$P_{se} = m_{si} \text{ Log } (SI) - m_{si} \text{Log } (SI)_{50}$$

$$P_{se} = [1.09 \text{ Log } (SI) + 0.432]$$

where P_{se} = Socioeconomic Probability Function

SI = Socioeconomic Index.

TABLE 7

RANKING BY FAMILY
SOCIOECONOMIC INDEX (SI)

SE	RANK	PARTILE	SI
3.1	123	1	0.051
3.7	247	1	0.090
4.1	371	1	0.117
4.4	495	1	0.139
4.8	618	2	0.169
5.1	742	2	0.192
5.6	866	2	0.234
6.1	990	2	0.278
6.6	1113	3	0.326
7.0	1237	3	0.366
7.6	1361	3	0.433
8.1	1485	3	0.493
8.8	1608	4	0.586
9.5	1732	4	0.692
10.3	1856	4	0.831
11.2	1980	4	1.019
12.1	2103	5	1.249
12.9	2227	5	1.501
14.3	2351	5	2.116

"SE" is the composite socioeconomic rank for a family

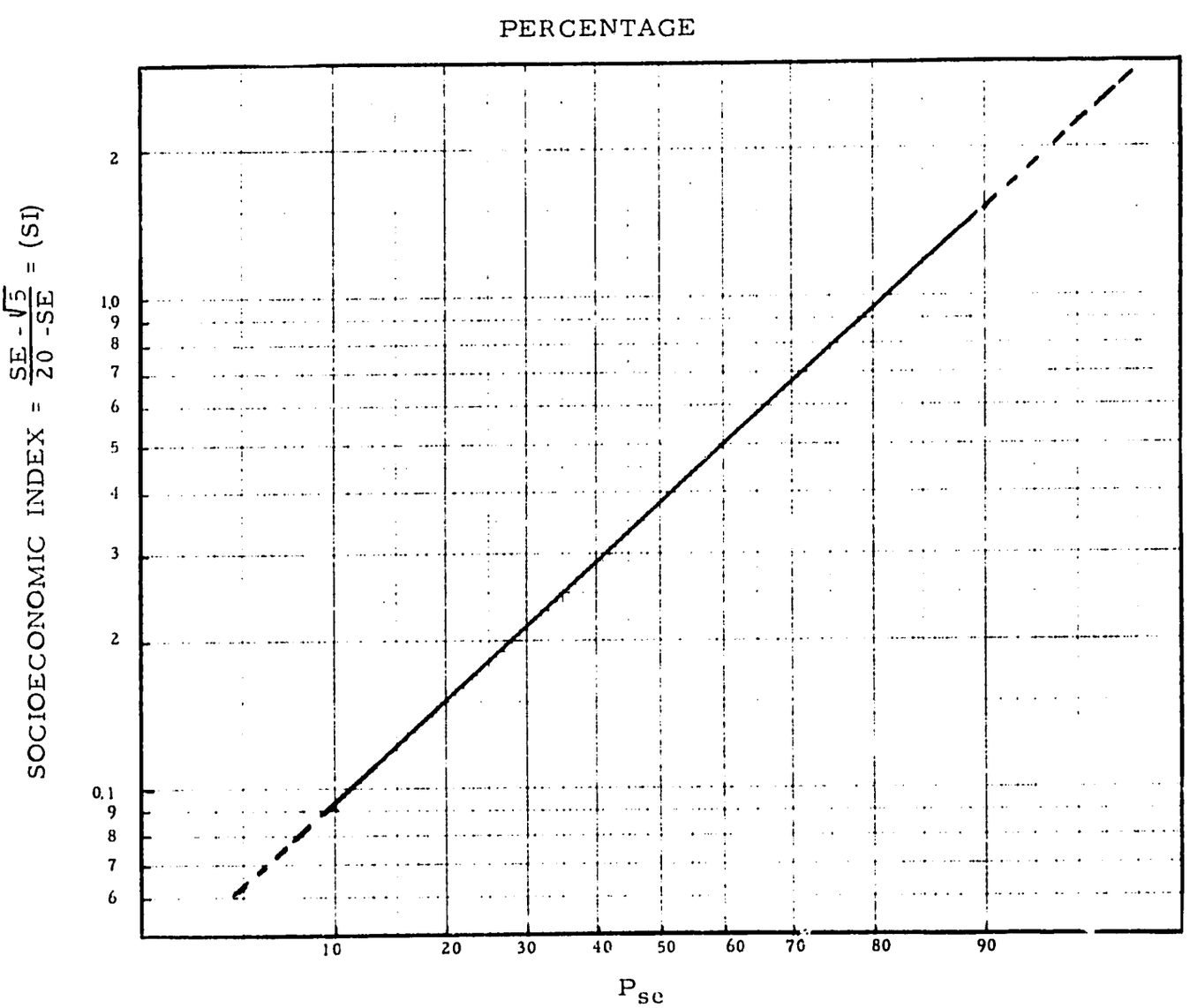
$$SE = \sqrt[4]{JAT * JOB * EXP * EDU}$$

Where:
JAT = Jati rank 5*(1 to 5)
JOB = Job rank 5*(1 to 5)
EXP = Expenditure rank (1 to 20)
EDU = Education rank (1 to 20)

$$SI = \frac{SE - \sqrt{5}}{20 - SE}$$ = Socioeconomic index

FIGURE 11

SOCIOECONOMIC INDEX (SI)
PROBABILITY FUNCTION (P_{se})



6. The District Rank Index (DI) - This function is based upon the nutrition status of individuals. Both calorie and protein levels are considered. An arithmetic average of the $\bar{X}$ values* for District against %C and %P was used for the average nutritional $\bar{X}$ value.** This index, was normalized to partiles by accumulating the sample distribution into an ogive. Data are shown in Table 8, the plot in Figure 12. The limits on $\bar{D}$ suggest the closed system format (see 5. above). A preferred handling would be to estimate pseudo upper and lower limits so that the rectified curve is more linear.

$$P_d = m_d \text{ Log (DI)} - m_d \text{ Log (DI)}_{50}$$

$$P_d = 0.96 \text{ [Log (DI)} + 0.027 \text{]}$$

where P_d = District Probability Function

$$DI = \frac{\bar{D} - 2.0}{4.0 - D}$$

$\bar{D}$ = Average "Calorie-Protein" $\bar{X}$

Summary

A summary of data for the probability functions discussed in this report is presented in Table 9. The data include variable (A), function as plotted (B), intercepts (10%, 50%, 90%) (C), slope of probability function (D), Log of "B" (E), and dimensions of "B" (F).

---

* The procedure for defining $\bar{X}$ is given under Data Subsets. $\bar{X}$ is the average partile value of a variable for a given data subset. The degree to which it differs from a central value (3.0 for pentiles) indicates the influence of the subset on the variable.

** The function defined in this way is a measure of the quantity of utilizable protein calories. The protein necessary for growth is emphasized. Most children in India appear to receive a diet of sufficient quality but not enough of it; for example, a (%P/%C) ratio of 1.2 or better but only 75% or less of the needed calories. The net result is consumption of 90% or less of the required protein. The degree that the average pentile varies from 3.70 is thus an indication of the degree of stress.

TABLE 8

DISTRICT RANK INDEX CALCULATIONS

District	$\bar{x}$ Calorie	$\bar{x}$ Protein	$\bar{D}$ / $\bar{x}$ Ave.	DI $\frac{\bar{D} - 2.0}{4.0 - \bar{D}}$	Pop. Run. Total	Top of Partile %
Thanj.	376	378	377	7.7	12912	100.0
Tiruc.	379	363	371	5.9	11836	91.6
S. Arc.	341	350	346	2.7	10954	84.9
Coimb.	348	331	340	2.3	9354	72.5
Raman.	306	312	309	1.2	8648	67.0
Dharm.	310	301	305	1.1	7942	61.5
Tirun.	295	305	300	1.0	7403	57.3
Nilgi.	301	290	295	0.9	6410	49.6
Madur.	287	284	285	0.74	5994	46.4
Kanya.	273	278	275	0.6	4940	38.3
Ching.	272	270	271	0.55	4424	34.3
Salem	253	247	250	0.33	3316	25.7
N. Arc.	250	235	242	0.27	2528	19.6
Madras	228	247	237	0.23	1450	11.2

$\bar{x}$ Calories = average %C partile value

$\bar{x}$ Protein = average %P partile value

$\bar{D}$ = district partile average = $\frac{\bar{x}\ \text{Cal} + \bar{x}\ \text{Pro}}{2}$

DI = district rank index = $\frac{\bar{D} - 2.0}{4.0 - \bar{D}}$

$$DI = \frac{\bar{D} - L}{U - \bar{D}}$$

L = 2.0 is a pseudo low value

U = 4.0 is a pseudo high value

FIGURE 12

PROBABILITY FUNCTION (P_d)
DISTRICT RANK INDEX (DI)

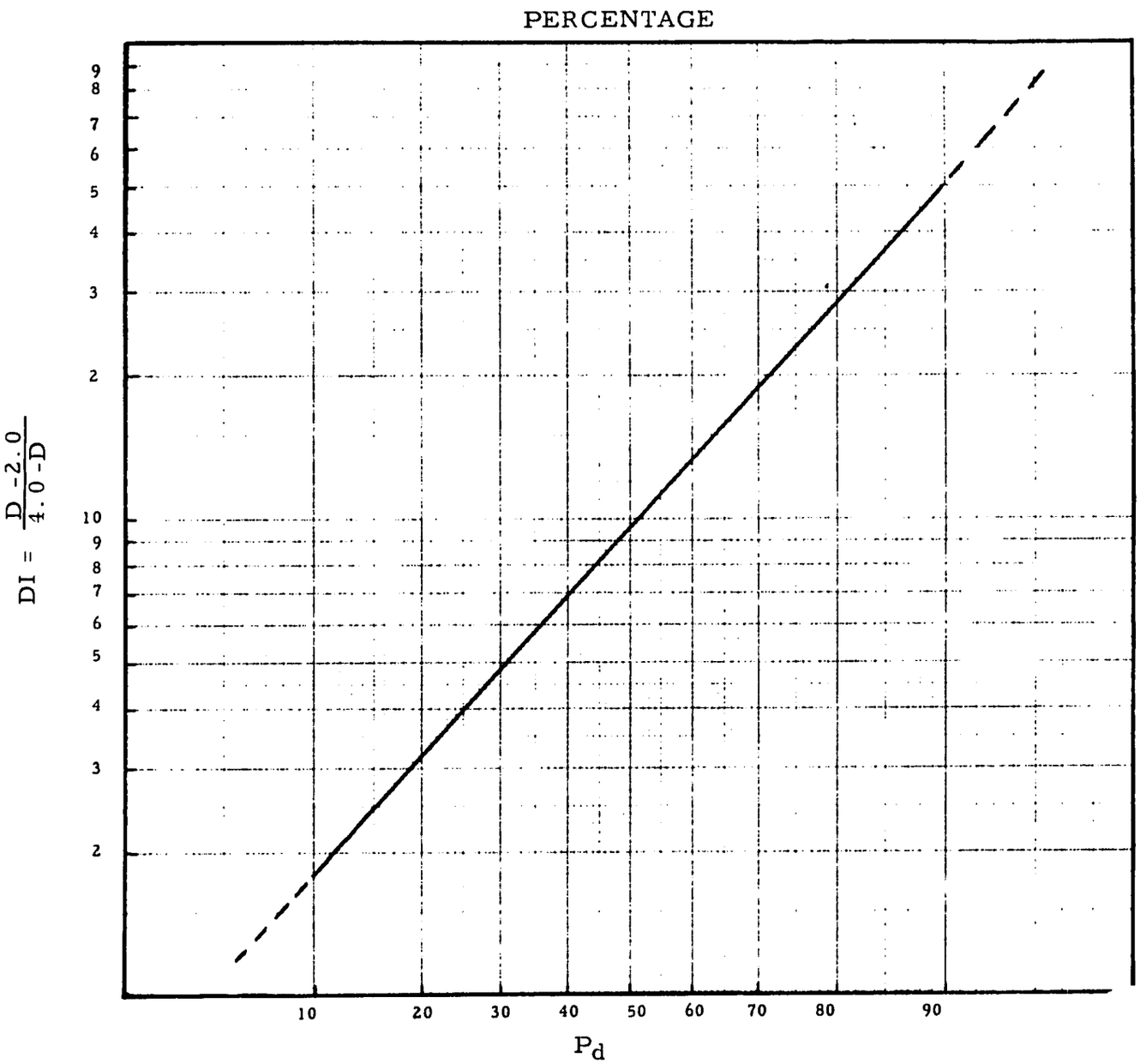


TABLE 9

SUMMARY OF PROBABILITY FUNCTION DATA

Probability* Function	Variable	Function Plotted	Log Normal Intercepts 10%	50%	90%	Slope of Probability Function m_x	Log (B)	Dimension of (B)
	A	Log (B)		C		D	E	F
P_c	% Calorie Need Met IND.	%C	47	79	131	3.06	1.898	Percentage
P_p	% Protein Need Met IND.	%P	56	94	161	2.96	1.973	Percentage
P_a	AEQ/FAM. by FAM.	AEQ/FAM	235	4.0	6.7	2.85	0.602	Number
P_{ai}	AEQ/FAM. IND.	AEQ/FAM	2.8	4.7	7.7	2.87	0.708	Number
P_e	EXP/AEQ by FAM.	EXP/AEQ	23	45	90	1.89	1.653	Rs/Mo/AEQ
P_{ei}	EXP/AEQ by IND.	EXP/AEQ	22	43	82	1.91	1.628	Rs/Mo/AEQ
P_{se}	Socioeconomic Index	$\frac{SE - \sqrt{5}}{20 - SE}$	.088	.37	1.5	1.09	-0.432	Index Number
P_d	District Rank Index	$\frac{\overline{D} - 2.0}{4.0 - \overline{D}}$	.18	.96	5.0	0.96	-0.027	D = Pentile
P_{jt}	Jati	$\frac{JT - 3\sqrt{5}}{20 - JT}$	.093	.41	1.75	1.05	-0.398	Index Number
P_{jo}	Job	$\frac{JB - 3\sqrt{5}}{20 - JB}$	.053	.29	1.60	0.927	-0.538	Index Number
P_{cf}	% Calorie Need Met FAM.	$\%C_F$	52	79	121	3.75	1.880	Percentage
P_{cfi}	P_{cf} Ordered by IND.	$\%C_{FI}$	52	76	111	3.43	1.898	Percentage
P_{pf}	% Protein Need Met FAM.	$\%P_F$	60		141	3.31	1.968	Percentage

*Lower case letter equivalents are used in Probability Function expressions.

Data Subsets

So far the discussion has been about relationships defined in terms of one or two variables. The large Tamil Nadu data sets, comprising several thousand families and over 13,000 persons, provide many opportunities to define data in multiple subsets.

As each definable subset is compared to the original total system the data are usually skewed, an indication of interactions. As the data are compared in pentile arrays it is possible to redefine the centroidal axes for each member of a data subset.

Bar X. Thus, the X and Y positions may be calculated as weighted mean partile positions. For pentile distribution this becomes:

$$X = (X_1 + 2X_2 + 3X_3 + 4X_4 + 5X_5)/ \Sigma X$$

$$\text{or generally } X = \sum_{1}^{n} i X_i \quad / \quad \Sigma X$$

For each level of a subset broken down by column or partile, the standard error of the estimate may be obtained about (X, Y) and a correlation ratio derived from it:

$$\eta = \sqrt{1 - \frac{S_{ay}^2}{\sigma_y^2}}$$

An estimate of S_{ay}^2 $\quad S^2 = \Sigma (Y_a - X_a)^2 M_a$

S^2 = standard error of the estimate assuming partile distribution.

M_a = weight of row a.

S_{ay} = standard deviation of y in row a

where Y_a mean position of row a

X_a mean position of centroid along row a.

The net effect of this calculation will be to align arbitrarily all rows and columns to the diagonal. This is automatically true in perfect correlation.

In arrays of two variables distributed by partiles $Y_a - X_a \longrightarrow 0$ but varies symetrically. This relationship is useful when arbitrary boundaries are used to define subsets, e.g., geographic boundaries (district, rural/urban split), demographic boundaries (age, sex), and when the variations are nonsymetrical η is the preferred correlation index if the relationship between variables is truly linear $\eta = \rho = r$. But if the relationship is nonlinear, then $\eta > \rho > r$ and as the difference becomes large η is probably the better coefficient and the relationship is probably nonlinear.

This method was used to develop the district ranking indexes. The average value between X for %C and X for %P for each district had the effect of ranking the districts along the diagonal in three dimensional space were district, %C and %P were the variables. Such a tactic makes it possible to do correlation analysis on variables not usually quantified.

The N-Dimensional Display

The probability grid is uniquely qualified for the presentation of data beyond the normal two or three dimensional displays. Color codes charts, pie charts, hostograms, are all used but present only a limited picture of interaction without becoming complicated. The present system, while still complex, permits the user to evaluate visually, and in an orderly way, the interactions of a modestly complex system. It is possible to identify objective or target areas and see how disperse or concentrated they are. This provides the planner with a better feeling for the required magnitude of a proposed intervention scheme - a feeling for the entropy* of the system.

Correlation coefficients obtainable at this stage will then have more meaning. The latter impart to the planner the visual sense of magnitude of the problem. The visual presentation provides accuracy, although not any more precision. For the planner, accuracy in specifying the target can greatly reduce cost in any program for reaching it.

---

* Entropy is a term from the physical sciences which is a direct measure of degree of randomness. It is being heard more frequently in the social sciences. One useful rule about entropy is that the larger the entropy (more random the system) at the initial state the more work will be required to develop order.

The experience in developing the system has demonstrated that the system may be optimized for the presentation of five variables on the same grid. Preferably, the variables are presented as a set of nested grids, a two variable (X, Y) outer matrix with a three-dimensional (R, S, Z) inner display. The development of such a display follows.

The Data Presentation

The process used to develop 5 variables in 5 one dimensional arrays into a full 5 dimensional display should provide material of particular use to planners. For this reason, the illustrations are centered about one of the data sets selected by a member of the Tamil Nadu Nutrition Planning Staff.* The variables tabulated are:

1. (X axis) %C = % fulfillment of calorie need by individual. Data in Table 3 and Figure 7.

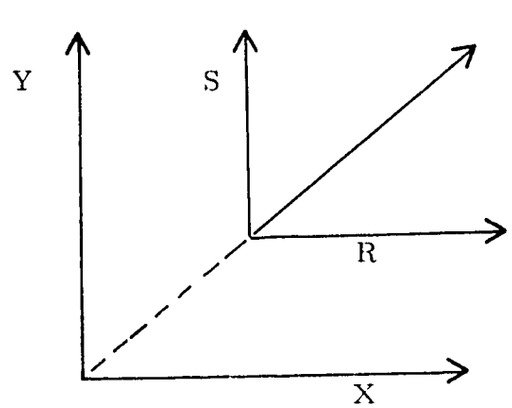
2. (Y axis) $\%C_{fi}$ = % fulfillment of calorie need by family, rank ordered by individual. Data in Table 10 and Figure 13 (Curve 2).

3. (R axis) AEQ/FAM = number of adult equivalents per family, rank ordered by individual. Data in Table 5 and Figure 9 (Curve 2).

4. (S axis) EAP/AEQ = expenditure (Rs/month) per adult equivalent per family, rank ordered by individual. Data in Table 6 and Figure 10 (Curve 2).

5. (Z axis) Age groups arranged to provide a maximum of information about the young infant.

The layout of the axes is schematically demonstrated as follows:



* S. Rajagopalan, Senior Statistician

TABLE 10

PERCENTAGE OF CALORIE NEED MET BY FAMILY

RANKED BY FAMILY ($\%C_F$)

RANKED BY INDIVIDUAL ($\%C_{FI}$)

$\%C_F$	Rank[1]	Partile	Rank[2] Family No.	By Individual
41	124	1		
51	248	1		
56	372	1	484	20% of individuals
61	496	1		
64	621	2		
68	745	2		
71	869	2	943	40% of individuals
74	993	2		
77	1117	3		
80	1242	3		
84	1366	3	1414	60% of individuals
87	1490	3		
90	1614	4		
94	1738	4		
98	1863	4	1910	80% of individuals
104	1987	4		
111	2111	5		
119	2235	5		
132	2359	5		

$\%C_F$ - % fulfillment of calorie need by family based upon 24 hour consumption recall (in %)

Rank[1] - $\%C_F$ is rank ordered by family partile position.

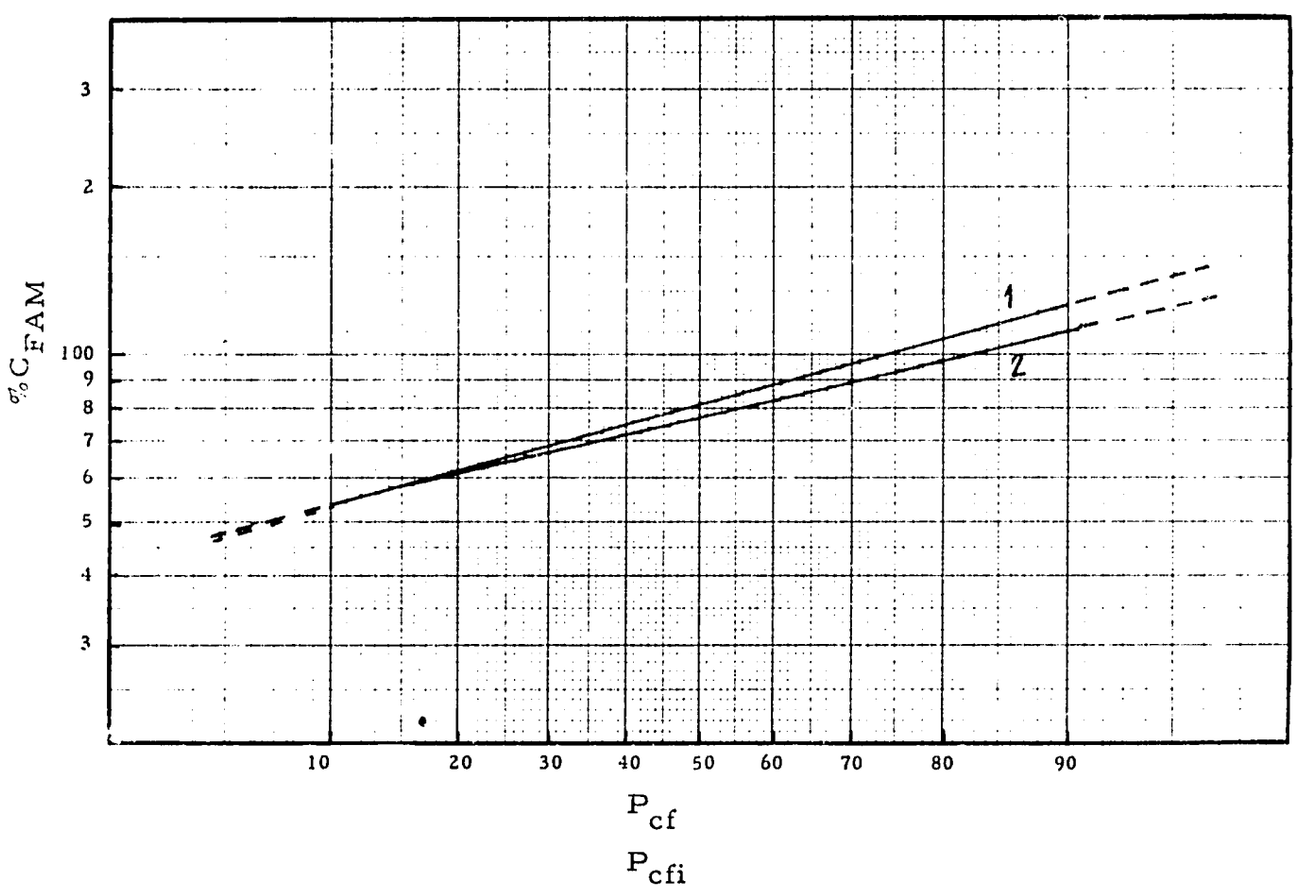
Rank[2] - $\%C_{FI}$ is rank ordered by summing the individuals to the net partile boundaries.

FIGURE 13

PERCENTAGE OF CALORIE NEED MET BY FAMILY ($\%C_{FAM}$)
PROBABILITY FUNCTION (P_{cf})

$\%C_{FI}$ vs P_{cf} Ranked by Individual 1

$\%C_{FAM}$ vs P_{cfi} Ranked by Family 2



The age groups ordered on the Z axis (item 5 above) are:

1. 0 - 6 mos.
2. 7 - 18 mos.
3. 19 - 30 mos.
4. 31 - 5 yrs.
5. 6 - 10 yrs.
6. 11 - 15 yrs.
7. 16 - 20 yrs.
8. 21 - 25 yrs.
9. 26 - 30 yrs.
10. 31 - 40 yrs.
11. 41 - 50 yrs.
12. 51 - 65 yrs.
13. 65 yrs.

The groups are selected to provide more access to the weaning years when the child is known to be most stressed. Even more age categories were desirable but not possible with the data available. It is felt that in future work great care should be taken in interviewing, to ascertain a child's age to the nearest quarter year until the age of 3 years. A closer estimate would permit a 6 or 7 age group separation for the preschool child.

The Array Size - The array including all total and subtotal columns will have $(5 + 1)^4 (13 + 1) = 18144$ boxes, of which $5^4 \times 13$ or 8125 are data boxes, the remainder are various totals and subtotals.

Assignment of Axes - With 18144 boxes in a grid, it is important to assign the variables to axes in such a way that dispersion of data is minimized. This is accomplished through preliminary analysis. The two most highly correlated variables are preferred as the outer array variables (Y and X). In this case, $\%C_{fi}$ versus $\%C_i$ are the preferred pair, Table 11. The pentile boxes are arranged conventionally so that the poorest off position is in the lower left box and the best off is in the upper right. Most of the data are seen to cluster (highest density) about the diagonal connecting these corners.

TABLE 11

INDIVIDUALS ARRAYED IN A PENTILE GRID
OF

X = PERCENTAGE FULFILLMENT OF INDIVIDUAL CALORIE NEED ($\%C_I$)
VERSUS
Y = PERCENTAGE FULFILLMENT OF FAMILY CALORIE NEED
(PARTILED BY INDIVIDUAL) ($\%C_{FI}$)

$\%C_{FI}$ Pentile	$\%C_I$ Pentile 1	2	3	4	5	Total	Function* Value c / (b)
5	49	61	170	620	1776	2676	121
							(104)
4	111	271	652	1009	565	2608	94
							(86)
3	247	631	851	622	158	2509	80
							(74)
2	580	1003	675	246	57	2561	67
							(60)
1	1598	624	241	93	33	2589	52
Total	2585	2590	2589	2590	2589	12943	
Function Value c	47	64	79	97	131		
(b)	(5.7)	(7.1)	(8.7)	(10.9)			

$\%C$

*Function Value· c = at center of partile
(b) = at partile boundary

The next step is to select the more correlated pair from the remaining three. This pair (S and R) is Expenditure per AEQ versus family size (reciprocal) (1 / (AEQ/Family)), Table 12. Since the normal correlation for this pair of variables is negative, the family size, (AEQ/FAM) array is reversed. The largest family in the low position, the smallest family in the high position. This change has the effect of accentuating the concentration of data along the lower left to upper right diagonal.

The next step is the preparation of the four dimensional array, Table 13. In this array, note that the totals in each square are summarized in Table 11, also that the grand total block is Table 12.

The net effect is then the expansion of information about each single value of Table 11 into a full 2-dimensional pentile array, resulting in four dimensions overall. As an analogy for this development, consider that the 12943 individuals in the grand total box represent a plant which develops branches as pictured in Figure 14. Each branch then develops buds of expenditure versus family size, as in Figure 15 for three dimensions and Figure 16 in four dimensions.

The final step is the expansion of each grid value into the Z axis of age distribution which brings about the full flowering of the plant. The final chart is a computer printout 120 by 36 inches. Only the lower right hand section is shown here in Figure 17 in which the grand total block is in effect the expansion of Table 12 into a third dimension of age distribution. The layout diagram indicating the order in which the arrays expand is shown in Figure 18.

0 indicates the grand total box 12943 individuals
1 indicates the pentile totals
2 indicates Table 11 and Table 12 grids and other two-dimensioned combinations.

In general, notice the direction of increase of array number from lower right to upper left as the dimensionality of the system increases.

TABLE 12

INDIVIDUALS ARRAYED IN A PENTILE GRID
OF

R - FAMILY SIZE - ADULT EQUIVALENTS/FAMILY (AEQ/FAM)
S - EXPENDITURE PER ADULT EQUIVALENT (EXP/AEQ)

PENTILES	5	4	3	2	1	Total	Function* Value c (b)
5	356	420	518	451	860	2605	81
							(65)
4	531	414	481	546	579	2551	57
							(48)
3	561	448	495	597	479	2580	43
							(38)
2	594	644	546	483	389	2656	33
							(28)
1	619	583	557	514	280	2551	22
	2659	2509	2597	2591	2587	12943	
Function Value c	8.2	6.0	4.8	3.8	2.8		
(b)	(6.8)	(5.3)	(4.3)	(3.4)			

*Funtion Value: c = at center of partile
(b)= at partile boundary

FOUR DIMENSIONAL ARRAY

%C Partile By Individual

3	1	6	1	3	14	3	2	2	6	4	17
1	3	3	4	2	13	3	3	1	2	2	11
2	2	2	2	0	8	1	2	2	5	2	12
1	2	2	2	1	8	0	3	4	3	0	10
1	0	1	2	2	6	2	1	2	4	2	11
8	8	14	11	8	49	9	11	11	20	10	61
2	2	3	5	9	21	6	8	11	9	18	52
6	4	2	8	4	24	23	15	16	16	16	86
7	1	6	9	10	33	18	6	12	20	8	64
0	4	5	4	1	14	6	10	9	9	2	36
3	7	2	3	4	19	6	8	8	7	4	33
18	18	18	29	28	111	59	47	56	61	48	271
4	4	8	5	15	36	12	26	15	14	27	94
8	8	5	11	12	44	19	15	24	27	37	122
4	16	18	8	6	52	20	26	28	19	21	114
11	22	15	16	9	73	26	53	31	31	14	155
10	6	8	13	5	42	40	26	32	23	25	146
37	56	54	53	47	247	117	146	130	114	124	631
25	18	21	15	16	95	63	39	28	15	40	185
19	16	17	16	17	85	24	38	36	39	34	171
35	27	24	23	22	131	56	38	44	35	45	218
33	24	21	30	17	125	48	26	64	39	29	206
27	43	35	31	8	144	45	60	55	34	29	223
139	128	118	115	80	580	236	201	227	162	177	1003
25	51	52	17	98	243	18	32	34	5	33	122
83	24	26	39	53	225	31	12	15	20	16	94
56	65	60	60	38	279	33	26	22	19	21	121
141	73	55	52	61	382	50	22	16	18	20	126
109	115	108	90	47	469	35	45	25	41	15	161
414	328	301	258	297	1598	167	137	112	103	105	624
59	76	90	43	141	409	102	107	90	49	122	470
117	55	53	78	88	391	100	83	92	104	105	484
104	111	110	102	76	503	128	98	108	98	97	529
186	125	98	104	89	602	130	114	124	100	65	533
150	171	154	139	66	680	128	140	122	109	75	574
616	538	505	466	460	2585	588	542	536	460	464	2590

TABLE 13

FAMILY SIZE BY EXPENDITURE IMBEDDED IN FAMILY CALORIE NEED MET BY INDIVIDUAL
% C INDIVIDUAL PARTILE BY INDIVIDUAL

2	1	13	15	13	44	14	14	39	43	57	167	45	60
3	10	10	15	5	43	7	25	32	30	38	132	29	73
4	1	6	5	5	21	17	14	16	32	28	107	48	35
5	9	5	5	4	28	20	43	25	20	20.	128	43	79
3	3	16	8	4	34	14	15	31	21	5	86	56	37
17	24	50	48	31	170	72	111	143	146	148	620	221	284
15	20	19	23	57	134	26	48	25	47	83	229	12	20
51	24	22	24	37	158	76	39	29	38	53	235	49	15
49	19	24	41	26	159	55	42	32	71	45	245	40	8
18	31	25	16	15	105	32	45	41	23	31	172	26	29
28	23	14	17	14	96	32	38	20	33	5	128	13	16
161	117	104	121	149	652	221	212	147	212	217	1009	140	88
14	18	30	25	52	139	8	18	19	17	29	91	5	4
18	24	30	39	43	154	19	13	30	23	21	106	6	10
27	26	44	36	41	174	24	27	34	23	19	127	11	8
26	66	46	49	29	216	19	46	36	36	19	156	4	13
61	23	28	30	26	168	49	23	28	24	18	142	16	7
146	157	178	179	191	851	119	127	147	123	106	622	42	42
32	16	22	16	20	106	12	4	11	7	9	43	2	1
23	21	20	24	20	108	9	10	12	8	14	53	4	3
29	30	33	32	21	145	9	14	8	10	8	49	3	4
41	23	33	38	18	153	14	10	5	7	5	41	3	3
29	39	39	35	21	163	12	15	16	14	3	60	3	9
154	129	147	145	100	675	56	53	52	46	39	246	15	20
5	11	14	5	13	48	1	1	8	5	17	32	2	1
17	6	6	8	4	41	2	2	3	4	2	13	1	1
11	8	14	10	10	53	2	2	3	6	0	13	0	1
15	7	11	5	3	41	11	1	4	2	1	19	1	0
20	17	3	14	4	58	3	3	3	4	3	16	0	4
68	49	48	42	34	241	19	9	21	21	23	93	4	7
68	66	98	84	155	471	61	85	102	119	195	562	66	86
112	85	88	110	109	504	113	89	106	103	128	539	89	102
120	84	121	124	103	552	107	99	93	142	100	541	102	56
105	136	120	113	69	543	96	145	111	88	76	516	77	124
141	105	100	104	69	519	110	94	98	96	34	432	88	73
546	476	527	535	505	2589	487	512	510	548	533	2590	422	441

CALORIE NEED MET				TOTAL					
115	121	194	535	67	78	175	186	271	777
93	112	105	412	43	114	139	163	152	611
37	84	74	278	72	54	63	128	109	426
62	56	78	318	69	136	98	86	103	492
64	50	26	233	76	56	114	85	39	370
371	423	477	1776	327	438	589	648	674	2676
19	32	38	121	61	98	77	116	205	557
32	31	36	163	205	97	101	117	146	666
21	31	22	122	169	76	95	172	111	623
18	17	8	98	82	119	98	69	57	425
17	8	7	61	82	92	61	68	34	337
107	119	111	565	599	482	432	542	553	2608
2	2	9	22	43	70	74	63	132	382
12	5	6	39	70	70	101	105	119	465
2	7	3	31	86	103	126	93	90	498
11	5	2	35	86	200	139	137	73	635
1	6	1	31	176	85	97	96	75	529
28	25	21	158	461	528	537	494	489	2509
1	0	1	5	134	78	83	53	86	434
5	3	0	15	79	88	90	90	85	432
0	5	2	14	132	113	109	105	98	557
1	0	2	9	139	86	124	114	71	534
1	1	0	14	116	166	146	115	61	604
8	9	5	57	600	531	552	477	401	2561
1	1	5	10	51	96	109	33	166	455
0	0	2	4	134	45	50	71	77	377
3	4	2	10	102	102	102	99	71	476
1	0	0	2	218	103	87	77	85	570
0	1	2	7	167	184	139	150	71	711
5	6	11	33	672	530	487	430	470	2589
138	156	247	693						
142	151	149	633						
63	131	103	455						
93	78	90	462						
83	66	36	346						
519	582	625	2589						

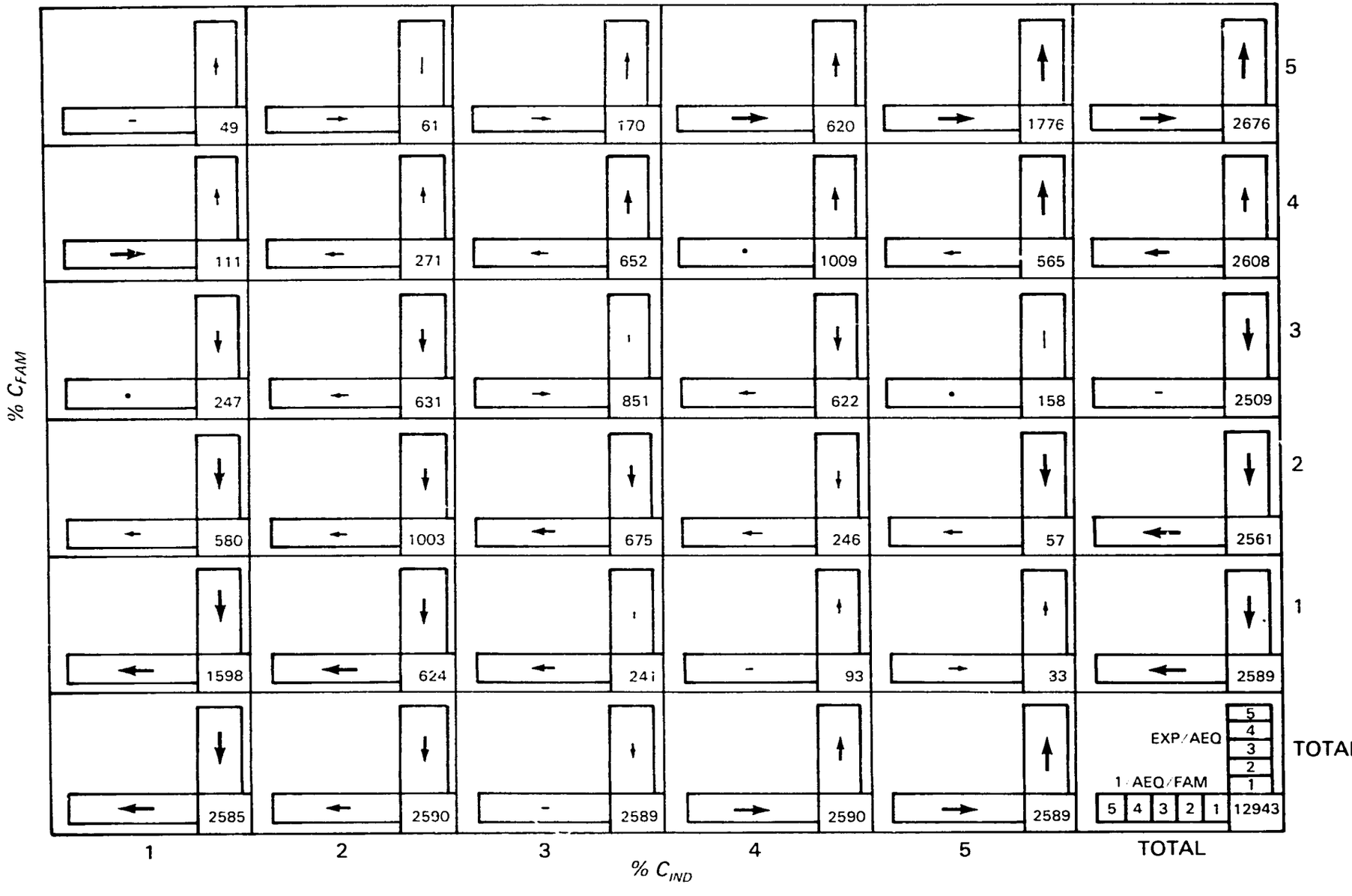
TABLE 12

5	4	3	2	1	Total	
356	420	518	451	860	2605	(5)
531	414	481	546	579	2551	(4)
561	448	495	597	479	2580	(3)
594	644	546	483	389	2656	(2)
617	583	557	514	280	2551	(1)
2659	2509	2597	2591	2587	12943	Total

FIGURE 14 Exploded View of % C_{FAM} vs % C_{IND} (Table 13)

% C_{FAM} \ % C_{IND}	1	2	3	4	5	TOTAL
5	49	61	170	620	1776	2676
4	111	271	652	1009	565	2608
3	247	631	851	622	158	2509
2	580	1003	675	246	57	2561
1	1598	624	241	93	33	2589
TOTAL	2585	2590	2589	2590	2589	12943

FIGURE 1 5 Three Dimensional Expansions of Data About the % C_{FAM} vs % C_{IND} Core (Table 13)

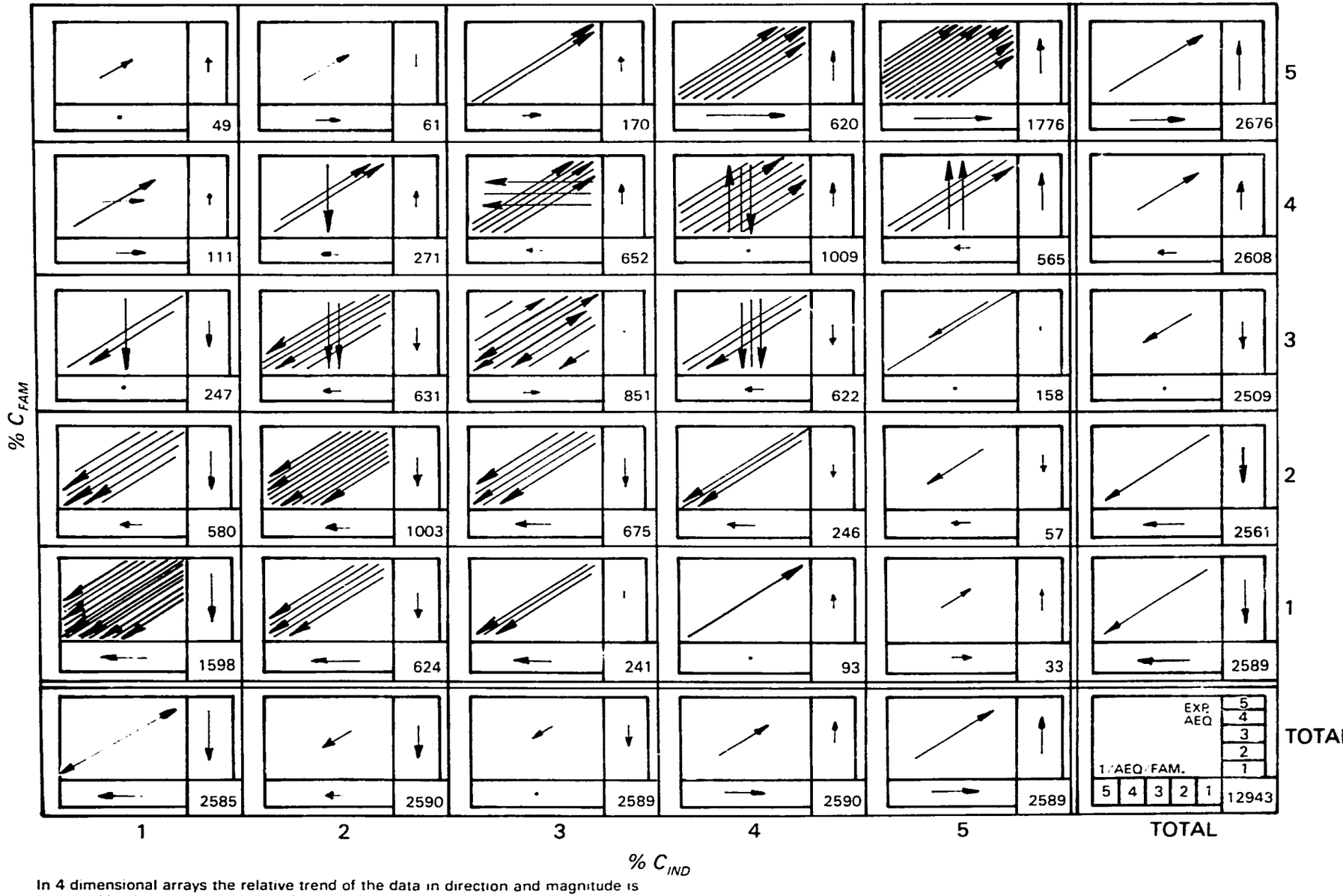


The relative trend of the data in direction and magnitude is indicated by the arrows.

FIGURE 1 6 Four Dimensional Extension of data about $\%C_{FAM}$ vs $\%C_{IND}$

(family size (AEQ/FAM) vs expenditure/AEQ) are the imbedded variables

(see data table 13)



In 4 dimensional arrays the relative trend of the data in direction and magnitude is indicated by the number and size of arrows. In lower order arrays, only one arrow is used

FIGURE 17

LOWER RIGHT HAND PORTION OF FIVE DIMENSIONAL MATRIX EXPANDED
SHOWING AGE DISTRIBUTION OF INDIVIDUALS ORIENTED ABOUT TABLE 12

4	8	4	10	18	44
2	6	6	4	6	24
9	6	2	3	1	21
3	8	3	2	2	18
8	2	5	2	1	18
26	30	20	21	28	125

10	15	24	20	49	118
17	17	25	21	28	108
18	6	10	19	15	68
9	25	14	7	12	67
10	9	6	6	9	40
64	72	79	73	113	401

8	9	25	23	51	116
13	16	16	22	19	86
17	9	12	23	16	77
16	15	12	7	18	68
10	17	13	10	3	53
64	66	78	85	107	400

8	11	23	42	44	128
9	13	24	25	35	106
7	4	10	25	22	68
5	19	14	17	17	72
10	5	15	11	14	55
39	52	86	120	132	429

9	2	8	10	18	47
5	5	9	8	18	45
9	3	3	6	11	32
4	6	3	6	9	28
4	3	3	7	2	19
31	19	26	37	58	171

8	20	15	15	36	94
11	10	11	11	15	58
18	11	9	8	8	54
16	20	9	7	7	59
18	9	9	10	4	50
71	70	53	51	70	315

31	38	51	35	103	258
49	36	41	35	60	221
46	34	49	42	41	212
46	57	42	30	39	214
51	37	29	31	31	179
223	202	212	173	274	1084

36	36	53	43	94	262
41	38	40	47	51	217
50	41	39	54	46	230
53	48	49	37	42	229
49	53	44	38	15	199
229	216	225	219	248	1137

28	44	71	81	139	363
50	56	75	81	101	363
43	59	73	90	82	347
52	79	79	74	55	339
58	75	94	86	50	363
231	313	392	412	427	1775

41	30	41	41	110	263
50	26	33	62	88	259
41	19	42	69	67	238
35	31	38	49	50	203
30	25	36	52	29	172
197	131	190	273	344	1135

4	4	9	5	7	29
9	4	5	5	4	27
8	0	0	4	7	19
3	3	2	3	4	15
3	0	0	1	1	5
27	11	16	18	23	95

3	3	4	5	2	17
3	3	3	1	3	13
5	4	1	0	3	13
9	2	0	0	2	13
5	3	0	0	0	8
25	15	8	6	10	64

4	5	10	5	8	32
5	11	10	5	2	33
7	8	1	8	0	24
7	10	5	3	3	28
13	3	7	4	1	28
36	37	33	25	14	145

5	14	17	18	15	69
11	17	22	15	11	96
11	13	14	15	6	59
13	20	20	17	10	80
13	14	19	9	1	56
53	78	92	94	43	360

5	4	7	14	16	46
8	3	12	15	12	50
5	1	5	14	13	38
2	7	8	7	3	27
6	9	7	10	2	34
26	24	39	60	46	195

1	5	1	3	7	17
4	2	3	4	1	14
1	1	0	5	5	12
1	3	3	5	3	15
1	1	2	2	1	7
8	12	9	19	17	65

41	28	41	38	98	246
57	24	24	27	71	203
68	26	26	38	64	222
50	48	19	35	53	205
61	22	16	30	36	165
277	148	126	168	322	1041

48	35	38	28	40	189
65	39	30	19	26	179
71	46	28	33	24	202
78	49	26	20	23	196
79	59	30	20	21	209
341	228	152	120	134	975

38	55	51	27	29	200
67	62	61	34	18	242
74	76	53	47	19	269
88	99	61	33	18	299
99	90	81	51	13	334
366	382	307	192	97	1344

30	68	75	63	57	293
67	59	86	107	41	360
76	66	86	98	34	360
92	109	115	85	32	433
78	98	104	81	28	389
343	400	466	434	192	1835

24	24	37	35	57	177
25	26	33	51	43	178
28	31	37	51	30	177
33	38	43	46	20	180
37	45	47	49	17	195
147	164	197	232	167	907

11	9	11	13	24	68
18	13	9	18	18	76
14	10	14	19	21	78
18	22	18	23	12	93
19	14	17	17	10	77
80	68	69	90	85	392

1	3	2	0	5	11
1	2	3	3	4	13
3	1	2	1	1	8
4	1	3	2	3	13
1	3	4	0	0	8
10	10	14	6	13	53

4	3	4	1	7	19
2	3	4	3	6	18
2	0	3	8	3	16
1	5	6	2	4	18
4	4	2	4	1	15
13	15	19	18	21	86

66	85	138	156	247	693
89	102	142	151	149	633
102	56	63	131	103	455
77	124	93	78	90	462
88	73	83	66	36	346
422	441	519	582	625	2589

10	18	20	20	39	107
20	15	28	33	23	119
23	23	27	30	27	130
22	30	25	27	28	132
21	41	33	24	15	134
96	127	133	134	132	622

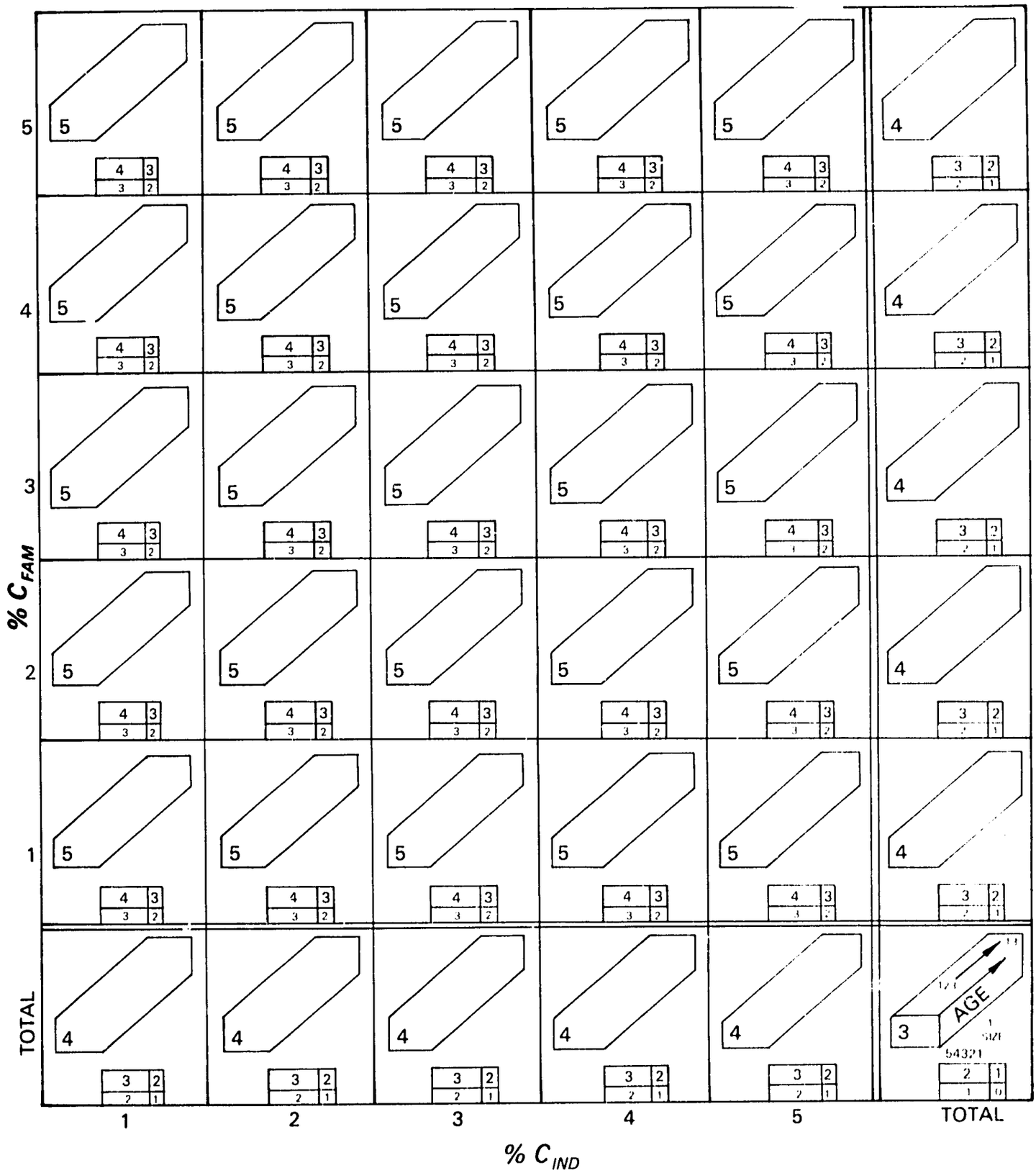
10	15	14	12	34	85
11	10	10	21	24	76
9	6	12	18	16	61
11	14	22	17	10	74
17	15	17	25	11	85
58	60	75	93	95	381

Table 12

356	420	518	451	860	2605
531	414	481	546	579	2551
561	448	495	597	479	2580
594	644	546	483	339	2656
617	583	557	514	280	2551
2659	2509	2597	2591	2587	12943

FIGURE 18 Five Dimensional Array

Outer array is % C_{FAM} vs % C_{IND} with EXP/AEQ vs. $(AEQ/FAM)^{-1}$ imbedded in it and age group imbedded within the latter.



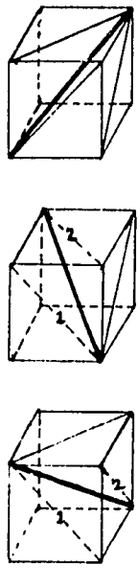
CHAPTER VI

THE VECTOR MODEL

Introduction

The material presented in this chapter describes a means with examples for determining the significance of combined correlation coefficients as the number of variables in a system is increased. For example, by the use of this methodology, the relative significance of cultural as distinguished from economic variables may be determined. Indeed, the modifying effect of adding or subtracting variables one by one allows the analyst, hence the planner, to isolate the variable or ultimately the combination of variables which, by their effect on the correlation coefficient, prove to be most important for providing the maximum resolution of a dynamic relationship among these variables. Conversely, the procedure, by warning in terms of overwhelming correlation, may reveal a static relationship which tends to obscure the dynamics of the system.

Vector Reversal



An example in three space will demonstrate the general correlation principles of the partile grid. If a major diagonal is drawn in a simple cube, the projections on the faces are then the correlation axes in two space. Reversing one axis in two space automatically reverses one more pair in three space. Thus, in three space the product of the signs in the correlation grid is always positive. In all higher space the rule holds. Therefore, when using a correlation matrix to generate a coefficient in (2 + N) space, the product of the signs of the two-dimensional correlation coefficients should be positive.

Confounding Variables

Isolated negative values will identify variables which have confounded components and which, therefore, may not be considered in isolation. In Table 1, are listed the cross correlation coefficient matrix for 10 variables arranged in rank order assuming $\%C_F$ as the dominant variable. Column 1A and 1B are variations on percent calorie need met which will be discussed in the context of the basic table. In Table 1, note that

EXP/AEQ by district and by rural/urban split are negative. These two negative values indicate that the data will not permit a simple discussion of expenditure by district without incorporating the rural urban relationship. It is quite possible that the family size relationship in "EXP/AEQ" is the confounding factor, or a significant part of it.

Other than the confounding pairs, the cause for which must be determined, the rules for analysis are:

(a) The minimum amount of work necessary to define the system is the definition of all correlation coefficients between each pair of variables in the system (isolated negative pairs of coefficients indicate a need to find a new variable).

(b) When an all positive grid of paired correlation coefficients is defined, the maximum amount of work necessary to define the system has been accomplished.

Steps of the Analysis

The next step is to determine the combination of variables most descriptive of the system. From the matrix of correlation coefficients in Table 1 a matrix of nondetermination coefficients (k^2) is prepared (Table 2).

$$k^2 = 1 - \rho^2 = \frac{S^2}{\sigma^2}$$

In a partile matrix all variables have an equal standard deviation (σ); the remainder of the analysis follows.

Based on Table 2 or similar data:

(a) On row A sum the columns for Σk^2.

(b) On row B prepare a running sum of row A.

(c) On row C total the number of k^2 values in the running sum -

$$\sum_{1}^{n-1} i = \text{Total} = \frac{(n-1)(n)}{2}$$

where n = number of variables

(d) Compute the overall correlation coefficient for variables

1 to c (c = column c)

$$\rho\, 1 - c = \sqrt{1 - B/C} = \sqrt{1 - \frac{\sum_{1}^{n} k^2}{n}}$$

(e) Calculate the relative degree to which the system is explained. This is equivalent to determining the length of the diagonal in the total system explained by the data. As long as this value steadily increases the new variable provides a substantive input.

$$Q = n \times \rho^2$$

If the correlation is perfect than $\rho = 1$ and n is the square of the diagonal of the n dimensional grid.

Interpretations

Note that the very high correlation between %P and %C overwhelms the multiple correlation. This indicates that anything said about %P alone or about %C alone will generally be applicable to the other. Consequently, these two variables should be ungrouped.

Table 3 shows the new data with %P deleted. Even here, the preliminary drop in correlation following "District" is a consequence of the ranking method (based on %P and %C). Since recovery is immediate (line E rising) it was retained because it also interacts with the other variables.

The probability function equation is then developed.

$$P_c = \rho \ (P_d + P_{ag} + P_e + P_a + P_{jo} + P_{ru} + P_{ed})$$

The conditions for linear regression being met:

(a) all factors homonormal

(b) all factors homoscedastic

(c) all factors orthogonal

The general equation follows:

$$m_e F(\%C) = \rho \left[m_d F(DI) + m_{ag} F(ag) + m_e F(EXP/AEQ) + m_a F(AEQ/FAM) \right]$$
$$+ \rho \, m_{jo} F(job) + \rho \, m_{jt} F(jati) + \rho \, m_{ru} F(Rur/Urb) + \rho \, m_{ed} F(Edu)$$

The slopes m_x and intercepts $F(x)_{50}$ data are found in Table 9.

Note the almost stable overall ρ as the last four variables are added, also that each individually correlated poorly with %C. This is a strong indication that the four factors should be discussed together as all interact with nutrition.

For example, it is observed that many higher ranked caste groups are better educated, have preferred jobs, earn and spend more money, eat only quality foods, but are more poorly off nutritionally than many lower ranked caste groups.

Even though the individual correlation of each factor with %C is poor, it may be concluded as a collective result that:

1. A family of higher ranking caste is likely to be more poorly nourished although this may be a matter of choice. [veg. versus non-veg. is a significant example.]

2. Families of lower ranked jobs fare more poorly than others.

3. The better educated generally fare better.

4. Urban families usually eat less well than rural families.

These relationships as well as others are expanded in various sections of the Project report.

By setting aside these four and the controversial "District" (which may not be used without rural-urban data when discussing Expenditure) and as previously noted, %P there are left four variables of which %C may be considered by family or by individual. To expand the concept of %C we may take variables 1A and 1B from Table 1 and develop the (k^2) non-determination Table 4 and find complete deterioration of correlation as new variables are added (Group 1, rows A to E).

Developing a new trial by eliminating either %C Individual or %C Family Group 2 and Group 3, rows A to E result. The data Group 3 would indicate that there was a good correlation matrix using %C Family, but that on an individual basis %C Individual (Group 2) Age was a grossly confounding variable. It was this developing picture which led to the "silhouette hypothesis" discussed in Volume I. In summary, "Age" showed no simple correlation with individual nutritional status.

These data point out the danger of projecting nutritional status on the basis of national food balance sheets or on the basis of family data alone. Only if the intrafamily distribution of food can be assessed is there a real chance of understanding the nutritional problems of Tamil Nadu and probably any other location.

It is perhaps fortuitous that based on the quantity of calories received, the infant is still identified as a target group. The limit to which he is permitted to consume could not be specified from the data.

TABLE 1

CORRELATION COEFFICIENT MATRIX

n		1A	1B	1	2	3	4	5	6	7	8	9	10
		% C_I	% C_{FI}	% C_F	% P	Dist.	Age	$\frac{EXP}{AEQ}$	$\left(\frac{AEQ}{FAM}\right)^{-1}$*	Job	$(Caste)^{-1}$	$(R/U)^{-1}$	Educ.
1A	% C_I	1.00	0.85	0.84	0.91	0.58	0.25	0.38	0.30			0.19	
1B	% C_{FI}		1.00	0.90	0.92	0.66	0.48	0.41	0.35	0.30	0.21	0.22	0.03
1	% C_F			1.00	0.93	0.60	0.48	0.90	0.33	0.30	0.21	0.14	0.03
2	% P_I				1.00	0.59	0.50	0.42	0.40	0.33	0.22	0.16	0.03
3	Dist					1.00	0.08	-0.29**	0.31	0.03	0.20	0.48	0.16
4	Age						1.00	0.26	0.10	0.05	0.05	0.11	0.05
5	$\frac{EXP}{AEQ}$							1.00	0.47	0.54	0.39	-0.43**	0.54
6	$\left(\frac{AEQ}{FAM}\right)^{-1}$								1.00	0.11	0.50	0.33	0.57
7	Job									1.00	0.36	0.08	0.28
8	$(Caste)^{-1}$										1.00	0.08	0.44
9	$(R/U)^{-1}$											1.00	0.44
10	Educ.												1.00

* The use of the reciprocal relationship reverses the sign of the correlation coefficient as well as directing the data toward the major correlation diagonal.

** The two negative values indicate that a fourth variable is confounding the relationship (perhaps family size?).

TABLE 2

TABLE OF COEFFICIENTS OF NONDETERMINATION

$1 - \rho^2 = k^2$

n		1	2	3	4	5	6	7	8	9	10
		% C	% P	Dist.	Age	$\frac{EXP}{AEQ}$	$\left(\frac{AEQ}{FAM}\right)^{-1}$	Job	$(Caste)^{-1}$	$(R/U)^{-1}$	Educ.
	% C	0	0.135	0.640	0.770	0.84	0.891	0.919	0.927	0.968	0.999
	% P		0	0.652	0.750	0.827	0.840	0.891	0.952	0.979	0.999
	Dist			0	0.994	0.916	0.904	0.999	0.960	0.770	0.974
	Age				0	0.932	0.990	0.998	0.998	0.988	0.998
	$\frac{EXP}{AEQ}$					0	0.779	0.708	0.848	0.815	0.708
	$\left(\frac{AEQ}{FAM}\right)^{-1}$						0	0.988	0.750	0.891	0.675
	Job							0	0.870	0.994	0.922
	$(Caste)^{-1}$								0	0.994	0.806
	$(R/U)^{-1}$									0	0.806
	Educ.										0
A	Σk^2	0	0.135	1.292	2.516	3.515	4.404	5.494	6.305	7.394	7.887
B	$\sum_1^n A$	0	0.135	1.427	3.943	7.458	11.862	17.356	23.661	31.055	38.942
C	$\sum_1^{n-1} c$		1	3	6	10	15	21	28	36	45
D	$\rho_{1 \to n} = \sqrt{1 - B/C}$		0.93	0.72	0.59	0.50	0.46	0.42	0.39	0.37	0.37
E	$n \times D^2$		1.73	1.56	1.39	1.25	1.27	1.23	1.22	1.23	1.37

TABLE 3

TABLE OF COEFFICIENTS OF NONDETERMINATION

$1 - \rho^2 = k^2$

n	1	2	3	4	5	6	7	8	9
	% C	Dist.	Age	$\frac{EXP}{AEQ}$	$\left(\frac{AEQ}{FAM}\right)^{-1}$	Job	$(Caste)^{-1}$	$(R/U)^{-1}$	Educ.
% C	0	0.64	0.770	0.840	0.891	0.910	0.927	0.968	0.999
Dist		0	0.994	0.916	0.904	0.999	0.960	0.770	0.974
Age			0	0.932	0.990	0.998	0.998	0.988	0.998
$\frac{EXP}{AEQ}$				0	0.779	0.708	0.848	0.815	0.708
$\left(\frac{AEQ}{FAM}\right)^{-1}$					0	0.988	0.750	0.891	0.675
Job						0	0.870	0.994	0.922
$(Caste)^{-1}$							0	0.994	0.806
$(R/U)^{-1}$								0	0.806
Educ.									0
A $\sum k^2$		0.64	1.764	2.688	3.564	4.603	5.353	6.420	6.886
B $\sum^{n} A$	0	0.64	2.404	5.092	8.656	13.259	18.612	25.032	31.918
C $\sum_{1}^{n-1} c$		1	3	6	10	15	21	28	36
D $\rho_{1 \to n} = \sqrt{1 - B/C}$		0.60	0.45	0.39	0.37	0.34	0.34	0.33	0.34
E $n \times D^2$		0.72	0.60	0.61	0.67	0.70	0.80	0.85	1.02

TABLE 4

COEFFICIENT OF NONDETERMINATION (k^2) TABLE
FOR A SELECT FIVE VARIABLES

			$\%C_I$	$\%C_{FI}$	AEQ^{-1}/FAM	EXP/AEQ	AGE
	1A	$\%C_I$	0.000	0.278	0.910	0.856	0.938
	1B	$\%C_{FI}$		0.000	0.878	0.832	0.770
	6	$(AEQ/FAM)^{-1}$			0.000	0.832	0.990
	5	EXP/AEQ				0.000	0.932
	4	AGE					0.000
	n1-5						
Group 1	A	Σk^2		0.278	1.788	2.520	3.630
	B	run. Σk^2		0.278	2.066	4.586	8.216
	C	$\sum_1^{n-1}$		1	3	6	10
	D	$\rho = \sqrt{1-B/C}$		0.85	0.560	0.49	0.42
	E	$n \times \rho^2$		1.444	0.934	0.943	0.892
	n1 + 3-5						
Group 2	A	Σk^2			0.910	1.688	2.860
	B	run. Σk^2			0.910	2.598	5.658
	C	$\sum_1^{n-1}$			1	3	6
	D	$\rho = \sqrt{1-B/C}$			0.30	0.37	0.24
	E	$n \times \rho^2$			0.18	0.40	0.23
	n2 - 5						
Group 3	A	Σk^2			0.878	1.646	2.692
	B	run. Σk^2			0.878	2.542	5.234
	C	$\sum_1^{n-1}$			1	3	6
	D	$\rho = \sqrt{1-B/C}$			0.35	0.39	0.36
	E	$n \times \rho^2$			0.24	0.46	0.51

CHAPTER VII

A CONVENTIONAL APPROACH TO SYSTEMS MODELLING FOR NUTRITION PLANNING

CONTENTS

	Page No.
Introduction	139
Data Needs and Availability	141
General Approach	142
(a) Food Balance Sheets	
(b) Optimization of the Cost Diet Ratio	
(c) Demand Analysis	
(d) Supply Analysis	
(e) Food Processing Subsystem Analysis	
Summary	146

A CONVENTIONAL APPROACH TO SYSTEMS MODELLING FOR NUTRITION PLANNING

Introduction. A model which helps in understanding the relationships between variables (inhibiting or accelerating) in a system, is a useful tool for nutrition planning.

A nutrition system can be broadly divided into three subsystems:

i. Consumer subsystem
ii. Food Production subsystem
iii. Processing and Distribution subsystem

In planning nutrition programs, it is necessary to have an insight into factors which influence each subsystem and evaluate the degree of interdependence among the subsystems. There are many factors outside the nutrition system which have an impact on it since the nutrition system is an integral part of the overall economic system.

Agricultural production is the backbone of the nutrition system. For example, in a developing country like India nearly 50% of the national income accrues from the agricultural sector alone. Clearly any intervention in the agricultural subsystem may have significant impact on economic development in general, and nutrition in particular.

With nutrition located at the center, the modelling task becomes clearer in the sense that relationships within and between subsystems are the primary objective. Each subsystem mentioned has its own components and thus its own sets of relationships. The model builder, as a first step, would describe mathematically the relationships within each sub system. Secondly, he would describe (mathematically) the interdependence among the subsystems and subsequently, identify those variables which influence the whole chain of the integrated system. Each subsystem independently can be defined by a set of mathematical equations having exogenous and endogenous variables such as:

$$F \quad (Y, X_1 \; X_2 \; X_3 \; . \; . \; . \; . \; . \; . \; . \; . \; . \; X_n)$$

An approach to the mathematical formulation of relationships within the integrated nutrition system is described by the following sets of simultaneous equations:

Consumer Subsystem

$D_t = Y_t \times p_t$ where $Y_t = F(X_1, X_2 \ldots\ldots X_n)$

D_t = Demand for year 't'

Y_t = Per capita consumption in year 't'

P_t = Population in year 't'

$X_1, X_2 \ldots\ldots X_n$ = Variables which influence the per capita consumption like production prices, income, etc.

Production Subsystem

$D_t = K(P - E + I - W)$* where P = Production

E = Export

I = Import

W = Seed, Feed and waste

K = Nutrient proportionality lost or gained during processing

Processing Subsystem

$K = F(t, C_p)$ where t = Technology

C_p = Consumer preference

* Generally in economic theory, a ceteris paribus assumption is made. Demand is = to supply in an equilibrium economy.

With the help of the above or similar sets of equations, estimation of demand, production, supply and consumption of nutrients can be accomplished. A systems model as described above, if successfully carried out, provides a comprehensive understanding of the nutritional implications. It can also help in identifying meaningful programs which will enhance nutrition in a given region within a given frame of time.

Data Needs and Availability. The data base available in Tamil Nadu from secondary sources is quite voluminous. Certain aspects of the available information lend themselves readily to the type of analysis described in the previous section. However, some informational needs for specific variables of the integrated nutrition system are not satisfactory. For example, income information is either not available or unreliable.

Another technique for studying the nutrition system is the input/output analysis. This is described elsewhere in this report.

A modeller, therefore, not only has to make judgements as to the extent of continuous information requirement but also decide which methodology to employ. Frequently, the initial conceptual framework of a model has to be reduced so as to correlate with the realities of data availability.

For example, we have sufficient data on the agricultural subsystem for a period of 10 years. These include:

1. Land use
2. Cropping pattern
3. Production pattern
4. Productivity
5. Price combinations
   etc.*

For the consumer subsystem, the Data Bank includes such information as the 1961 Food Habits Survey in Tamil Nadu, ICMR nutrition surveys and geographically limited surveys carried out by various research and academic institutions. In the case of food processing subsystem, information about the processing industry was made available from the Annual Survey of Industries and the ORG Survey.

* The detailed list of the contents of the Data Bank is contained in Chapter IV.

Having this data in an organized form, the method of analysis might be along the so-called conventional econometric approach. The data might have been sufficient to have made such an analysis useful for planning purposes.

This is not to say that this approach would have yielded the best possible results but the steps described in the following sections are common and therefore prerequisites to any technique in econometric modelling.

General Approach. The first step is to analyze the available data and observe trends through elementary analytical manipulations, and identify initial hypotheses requiring verification by more advanced analytical techniques. An elementary scheme of analysis such as the following would be a useful guide in structuring the basic framework of a systems model:

(a) Food Balance Sheets

Food balance sheets for individual commodities which consist of:

i. Total State food supplies derived as a balance between production, changes in stock, import and export.

ii. State supply for commodities obtained as a balance between total food supplies and seed-feed waste.

iii. Per capita food availability per year and daily nutrient intake in terms of calories and proteins.

Expressing the above in functional terms -

Net availability $Y = P + I - E \pm S - W$

where P = Production

I = Import

E = Export

S = Changes in stock

W = Seed feed, storage loss and non-food utilization

This is a useful tool for estimating in an aggregate way, the nutrition availability of a region. The Food and Agricultural Organization of the U. N. uses the same approach for projecting the supply of agricultural commodities. As the Food Habits Survey results became available, this approach could be varied.

(b) Optimization of the Cost Diet Ratio

One of the important objectives of the study is to identify or recommend development of a low-cost-nutrient-rich food item based on indigenously available resources. Conventional linear programming can help in identifying the least-cost combination of a nutritious diet. For example,

$$\text{Minimize Cost } C = \Sigma P_r f_r$$

P_r = the price

f_r = the food article

subject to getting a given nutrient $N = \Sigma n_r f_r$

where n_r is the nutrient content in a unit of f_r.

This method can also be used for identifying the most advantageous diet for feeding programs.

(c) Demand Analysis

The demand for a given food commodity is a function of its price, prices of substitutes, prices of complement, income levels, tastes and preferences, etc. This can be expressed mathematically as follows:

$$Q_d = F\ (P,\ P_s,\ P_c,\ Y,\ T)$$

where Q_d = Quantity demanded

P = Price of commodity

P_s = Price of substitutes

P_c = Price of complement

Y = Income

T = Tastes, etc.

The household consumption survey provides adequate information to carry out such an analysis.

Conventional econometric models are very useful in establishing functional relationships between the forces at work in a given socio-economic system. The data available satisfy most of the requirements for an austere econometric model except for income statistics. Household income data is not reliable in India. The practice is to use expenditure information as an approximation to disposable income. Since demand analysis depends heavily upon factors of income elasticities, these are estimated through the following procedure:

1. $\text{Log } Y = a + b \text{ Log } X$
2. $Y = a + b \text{ Log } X$
3. $\text{Log } Y = a - b/X$
4. $\text{Log } Y = a - b/X - c \text{ Log } X$

where Y is Consumption, X is Income, a, b and c are Constants.

Prior to selecting the variables to be included in demand equations and avoiding pre-judgement, factor analysis by calculating zero order correlation matrices must be undertaken.

An aggregate demand function for a given year or period of time is presented below:

$$Q^t = Q_0 \prod_{1}^{12} \left(1 + \frac{\gamma_x}{100}\right)$$

where

Q_t = per capita quantity demanded at period of time 't'

Q_o = per capita quantity demanded at the base year

γ_x = percent change in variables (some have to be assumed, others calculated using previous trends)

e_x = partial elasticity of demand with reference to the variables considered.

(d) Supply Analysis

Estimation of demand alone has not brought us close enough to identifying solutions to the problem of nutrition. In fact, having estimated the demand, many more questions have been raised than solved. The most important questions raised: is the supply of available and potential food items adequate to meet the demand? Constructing a supply function is basically similar to that of demand, in terms of methodology. It is the informational inputs which differ. In a supply function, data inputs such as fertilizers, pesticides, land use pattern, price policies, subsidy programs and others; play an important role. In addition, general economic development considerations influence the supply portion of the model as well. Here again, the data available from secondary and primary sources may be utilized in constructing a set of supply functions. These must now be related to the demand that is likely to be generated. For example, the alternatives:

(i) Dynamic Causal Chain Method by the following set of equations:

$$S_t = \alpha_1 P_{t-1} + \beta_1 X_{1t} + \varepsilon_1$$

$$D_t = \alpha_2 P_t + \beta_2 X_{2t} + \varepsilon_2$$

$$P_t = \alpha_3 (D_{t-1} - S_t) + \beta_3 X_{3t} + \varepsilon_3$$

where

S = supply; D = demand; and P = price; these are the endogenous variables of the model.

x_i = Exogenous variables representing outside influences such as rainfall, technology, etc.

$_i$ = disturbance term residuals - they have zero expectation and are not correlated with exogenous and logged endogenous variables.

(ii) Interdependence System which has the following set of equations:

$$Q_t = S_t = \alpha_1 P_t + \beta_1 X_{1t} + \varepsilon_1 \text{ (demand curve equation)}$$

$$Q_t = D_t = \alpha_2 P_t + \beta_2 X_{2t} + \varepsilon_2 \text{ (supply curve equation)}$$

While one set of equations portrays the demand characteristics of the population, the supply set of equations shows the extent to which the demand could be met.

(e) Food Processing Subsystem Analysis

The food processing subsystem plays an important role in linking the chain from production to consumption. Most of the agricultural products go through some sort of processing procedure before reaching the ultimate consumer. Food processing causes either an addition or deletion of nutrients. The magnitude of these food alterations is largely a function of the state of technology in the food industry. Technology in turn depends on general economic conditions, investment incentives, availability of capital and the speculative disposition of businessmen. The price of the processed food item is a function of all these factors in addition to production costs.

The functional relationship in the food processing subsystem can be expressed as:

$$P = F\ (P_o,\ T\ \ C_p,\ I)$$

where

P = final processed product

P_o = Input

T = Technology

C_p = Consumer preference

I = Capital investment

Summary The foregoing description of an approach to nutrition systems analysis might have been undertaken with the available data. Such an approach would have been initiated while realizing that mid-way, we may have had to change our minds. Clear indications may have emerged from such analysis that our proposed approach cannot be supported by the available data. This type of situation confronts every researcher, but the flexibility of the conventional approach might have been sufficient in our purposes.

The model might be used for simulation purposes and thus a "feel" for the reality of it could be obtained. The set of equations explaining the consumer behavior, production pattern, and processing system constitute the integrated nutrition system model. With the help of this model a number of feasible interventions might be identified which later could have been subjected to cost-benefit analysis. This ultimate procedure provides the decision-maker with a sound basis for intervention judgements.