

INDIVIDUAL NUTRIENT ALLOCATIONS FROM  
HOUSEHOLD AGGREGATES:  
VARIATIONS IN THE NUTRIENT COMPOSITION OF DIET  
IN INDONESIA

Andrew Chesher<sup>1</sup>  
Department of Economics,  
University College London

October 1st 2000

ABSTRACT

This paper disaggregates data on the food consumption of Indonesian households to obtain estimates of average rates of energy intake, and rates of energy intake from the three major sources of energy: fat, protein and carbohydrate, for males and females at each year of completed age in rural and urban areas. The disaggregation procedure, which involves the use of roughness penalty methods to obtain smooth estimates of age profiles of nutrient intakes, is outlined. The results indicate that there are marked differences in rates of nutrient intakes across males and females, across ages and across urban and rural dwellers. Two of the three sources of food energy, fat and protein, show much greater sensitivity to variation in household income than does the third energy source, carbohydrate. Females seem to obtain higher proportions of energy from fat and protein during child bearing years. Male nutrient intakes seem to be slightly more sensitive to variation in household income than female intakes, but the proportion of energy obtained from the three energy sources varies with income in a similar fashion for males and females.

---

<sup>1</sup>Department of Economics, University College London, Gower Street, London WC1E 6BT, UK, email [andrew.chesher@ucl.ac.uk](mailto:andrew.chesher@ucl.ac.uk). This paper was prepared for presentation at the annual conference of the Latin American and Caribbean Economic Association, Rio de Janeiro, October 14th 2000.

## 1. INTRODUCTION

It is important to have good information about the nutrient content of the diets of individuals of different sexes, ages and circumstances. For example there may be concerns that diets are inadequate among the poorest people in a country, perhaps lacking sufficient protein and fat or even energy content to support a healthy life. In many rich countries there are concerns that the diets of many people are too rich in fat, leading to obesity and consequent diseases such as diabetes, and to cardiovascular disorders<sup>1</sup>.

Unfortunately information about the variation in the nutrient composition of diet across people of different types and circumstances is scarce. Typically such information is collected using recorded intake surveys of individual consumption. These sorts of surveys are very expensive to conduct and rarely found although there are a few notable examples available, for example the UK National Diet and Nutrition Surveys<sup>2</sup> (NDNS). Even these surveys produce information which requires careful interpretation. The intrusive nature of recorded intake dietary surveys may actually disturb the process that is being observed<sup>3</sup>. For example people with what they regard their eating habit as bad may modify their diet while in the survey, or under-record. In one of the reports of the UK NDNS<sup>4</sup> under recording is flagged as a potentially serious problem and it is suggested that among certain groups of people average recorded energy intakes are insufficient to sustain the expected rates of energy expenditure of these groups of people.

Most of the available information about diet and its composition comes from household (rather than individual) surveys in which data on either household food consumption or amounts of foods entering the household during some recording period are collected. This paper outlines one method for extracting from such data information about the diets of *individuals* of different ages, sexes and circumstances. The method is applied to data from a survey of almost 60,000 Indonesian households covering around 1/4 million people, conducted in 1993.

The method outlined and employed here, essentially a “statistical disaggregation” of total household nutrient acquisitions into amounts provided to individuals of different ages and sex, was developed and applied in Chesher (1997). The method produces flexible, in a sense non-parametric, estimates of the age profiles of nutrient intakes for males and females and estimates of the way in which these vary across types of household, for example rich and poor, urban and rural. Roughness penalty techniques are employed to bring smoothness to the estimated age profiles. These are well suited to this problem in which the object about which we have a “smoothness prior” (the age profile of average rates of nutrient intake) is to some extent hidden from the eye by the observation process, here by aggregation across household members.

In Chesher (1997) British National Food Survey data was employed to produce estimates of average rates of intake of energy, saturated and unsaturated fats, calcium and vitamin C for males and females at each integer year of completed age from 0 to 99. This analysis was extended in Chesher (1998) to provide estimates of age profiles of British nutrient intakes for males and females for each of 20 years, allowing the construction of birth cohort specific time series of estimated nutrient intakes. These suggested that the proportion of energy derived from fat, a key dietary target in the UK and many other countries, had fallen much faster for women than for men over this period, though from a higher starting value for women. Parkin *et al* (1999) employ elements of the procedure to

---

<sup>1</sup>A survey of relevant research and a list of recommendations concerning the fat content of diets is given in World Health Organisation (1990).

<sup>2</sup>See e.g., Gregory *et al*, (1990), Gregory *et al* (1994).

<sup>3</sup>Concerns about the accuracy of recorded intake surveys are expressed in, for example, Bingham *et al* (1995), Black *et al* (1991), and Livingstone *et al* (1990).

<sup>4</sup>Gregory *et al* (1990).

estimate age profiles of morbidity and health service utilisation using British data covering the period 1984 - 96. Deaton and Paxson (2000) apply the method to study individual age - savings rate profiles in Taiwan and Thailand. Vasedkis and Trichopolou (2000) apply the method to study individual food availability. Miquel and Laisney (2000) provide an application to data from Czechoslovakia on household nutrient intakes covering the period 1989-92.

This paper uses data from the 1993 Socioeconomic survey of Indonesia (SUSENAS) recording amounts of around 200 foods consumed in households during one week recording periods that spanned a calendar year. These amounts are converted into amounts of nutrients and the resulting data are employed to estimate age profiles of average rates of intake of energy and of the three main energy producing nutrients, fat, carbohydrate and protein for males and females<sup>5</sup>. These are processed to provide age profiles for the proportion of energy obtained from fat, protein and carbohydrate which are informative about variations in the “quality of diet” across ages, males and females, and urban and rural dwellers. The households studied here live in very different environments, urban and rural, and in a diverse range of islands in the Indonesian archipelago. All the analyses reported here are conducted separately for urban and rural households.

Energy, fat, carbohydrate and protein intakes of individuals are found to vary in a complex but systematic way across people of different ages. The variation is different among urban and rural dwellers, and different for males and females. For example females and males obtain about the same proportion of energy from fat and protein up to age 15 and from age 40 into old age. But between these ages, during which many of the women studied will be bearing children, females obtain significantly higher proportions of energy from fat and protein than do males.

It is possible that larger households are more efficient in converting food entering the household into food consumed by household members. This “economy of scale” effect is investigated and found to be present but rather small.

Around 15% of the households studied have incomes placing them below a poverty line which is defined as about 1.2 times an amount allowing the purchase of 2100 Kilocalories of energy per person per day, the basket of goods considered reflecting the typical diet of the poor. Some of the other households covered in the survey have much higher incomes. Income at the 9th decile of per capita household income is 5 times higher than income at the 1st decile.

We study the association between nutrient intakes for males and females of different ages and household income. There is a small but statistically significant positive association between total energy intakes and household income per head. There is a stronger association between fat and protein intakes, and income implying that higher incomes are associated with a higher quality diet, an effect whose magnitude we measure. We investigate the differential in the sensitivity of males’ and females’ nutrient intakes to variations in household income and find that females’ intakes show slightly less sensitivity. However this is found to be a “scale effect”. The sensitivity of the *proportions of energy* coming from the three energy sources varies with household income in a very similar fashion for males and females.

To interpret these results, and to consider their policy implications, we have to cast them in the context of a model of dietary choice, a task not attempted here, but a natural next step. The main purpose of this paper is to show how, under sufficiently restrictive assumptions, information on the variation in diet across individuals can be wrestled from household aggregate data.

There are quite complex issues involved in interpreting our results and in constructing and estimating a model of dietary choice. For example suppose that people who are

---

<sup>5</sup>Alcohol is not studied here. The SUSENAS records almost no acquisitions of alcohol in this mainly Muslim country.

engaged in labour market activities expend more energy those who do not. Then we might expect to see the sort of association between energy intakes and household income that is apparent in these data. Of course there would be variations across types of occupation that should be taken into account. But this would not explain the much greater sensitivity of fat and protein intakes to income than we see for energy in total, and for carbohydrate. This is likely the consequence of the higher “prices” of fat and protein, and to understand these variations with income requires consideration of choice of diet, and perhaps labour force participation, under different configurations of food prices and wages.

The method used here makes extensive use of the ages of people recorded in the SUSENAS. It is clear that the recorded ages are contaminated by measurement error taking the form of age heaping, in which some ages are recorded to the nearest 5 or 10 years. This is a very common feature of household survey data. In an appendix to the paper the impact on the results of this shortcoming in the data is investigated.

The remainder of the paper is laid out as follows. Section 2 outlines the estimation method, Section 3 describes the data employed, Section 4 gives the empirical results and Section 5 concludes.

## 2. METHOD

**2.1. Disaggregation.** The data employed record amounts of energy and of nutrients (fat, protein and carbohydrate) consumed in each surveyed household during a one week period. For a typical nutrient the recorded amount is modelled as a realisation of a random variable  $Y$ .

The object of interest is the long run average rate of consumption of the nutrient by a person,  $p$ , with characteristics  $x_p$ , for example age and sex, living in a household with characteristics  $v$ . Let this be  $h(x_p, v)$ . Let  $c(x, v)$  be the average long run rate of consumption of the nutrient by a household which contains  $P$  members<sup>6</sup> with characteristics  $x = [x_1, \dots, x_P]$ . Nutrients are private goods, so the average rate of nutrient consumption for a household is the sum of the rates of consumption by household members, giving  $c(x, v) = \sum_{p=1}^P h(x_p, v)$ . In principle  $v$  could contain measures of household composition.

Now consider the relationship between the amount of a nutrient consumed in a household during the recording period,  $Y$ , and the average rate of consumption of that nutrient,  $c(x, v)$  by the household. Providing that a long enough period<sup>7</sup> is studied so that periods of general increase in stocks (e.g. before festivals) are compensated by periods during which stocks are consumed, the expected value of  $Y$  will be equal to the average rate of household consumption and since this will be true for each type of household,  $E[Y|x, v] = c(x, v)$ . With a specification for the function  $h(x_p, v)$  we can conduct GMM estimation based on the following conditional moment restrictions.

$$E\left[Y - \sum_{p=1}^P h(x_p, v) \mid x, v\right] = 0. \quad (1)$$

---

<sup>6</sup>Visitors are not included in the list of household members. At the highest level of aggregation the presence of visitors can be ignored because households acquiring food for visitors will be counterbalanced by households acquiring less food because they have members visiting other households. If the presence of visitors is correlated with the age and sex composition of households then ignoring visitors will cause estimates of nutrient - age profiles to be biased. Calculations using British data reported in Chesher (1997) suggest that this effect is small.

<sup>7</sup>All the analyses here employ whole calendar years of data.

**2.2. Implementation.** So far as household characteristics are concerned, for most of the paper we employ the simplifying assumption  $h(x_p, v) = q(v) \times h(x_p, v_0)$  for some  $v_0$  common to all households. This assumption implies that ratios of intakes of a nutrient across household members are invariant with respect to household characteristics. When studying the impact of income on nutrient intakes we relax this assumption, allowing male and female intakes to vary differently with household income. This is of interest when investigating gender bias in provision of nutrients to household members. A single index model is used for the influence of  $v$ ,  $q(v) = \exp(\delta'v)$ .

The individual characteristics considered here are age ( $a$ ) and sex ( $m_p = 1$  if person  $p$  is male, 0 otherwise, and  $f_p = 1 - m_p$ ). Since nutrient intakes of males and females may be different, separate nutrient intake - age functions are specified. The assumptions made so far result in the following modification of (1).

$$E\left[Y - \left\{ \sum_{p=1}^P (m_p h^M(a_p) + f_p h^F(a_p)) \right\} \exp(\delta'v) | x, v\right] = 0 \quad (2)$$

Here  $x$  now lists the ages and sex of each household member and  $h^M(\cdot)$  and  $h^F(\cdot)$  are sex specific nutrient intake age profiles, independent of household characteristics.

When investigating the possibility of economies of scale in converting food prepared for consumption into nutrients consumed by household members we employ an extension of this moment restriction

$$E\left[Y - \left\{ \sum_{p=1}^P (m_p h^M(a_p) + f_p h^F(a_p)) \right\}^\gamma \exp(\delta'v) | x, v\right] = 0 \quad (3)$$

in which  $\gamma$  measures the scale effect and we might expect to find  $\gamma < 1$ .

A simplifying assumption made here is that average rates of nutrient consumption by household members are independent of household composition<sup>8</sup>. This doubtless involves a degree of specification error. For example we might expect people in households that contain young children to have different rates of nutrient intake than people in other households. The presence of children may be associated with women who would otherwise be engaged in labour market activities in which they would perhaps have higher rates of nutrient intake. But caring for children expends energy and the direction of the effect is unclear. There is mixed evidence on this from recorded intake surveys of individuals. Mills and Tyler (1992) report average daily energy and fat intake about 4% higher for infants with siblings than for lone infants but Department of Health (1989) reports no significant relationship between dietary patterns of school age children and household composition. Gregory *et al* (1990) in a survey of adults find no significant relationship between energy intake and household composition for males, but among females 8% lower energy intakes for lone than for other mothers.

One way to examine the degree of specification error involved is to estimate the model separately for different household composition types. Chesher (1997) reports estimates of a model like the one used here applied to single cross sections of households categorized into four household composition types: single male households, single female households and multi-person households with and without children. Small differences in estimated rates of consumption were found across these groups with highest rates of consumption in single person households. However where comparisons could be made the relationship with age was very similar across the four groups. Similar results are found for the Indonesian data studied here.

<sup>8</sup>That is, the list of household characteristics,  $\mathbf{v}$ , does not include measures of household composition.

**2.3. Roughness penalties.** Nutrient intakes have a highly nonlinear dependence on age, rising among young people and falling among the old and being elevated during periods of growth and raised physical activity. Because of the complexity of this relationship it is difficult to find a good parametric specification of the functions  $h^M(\cdot)$  and  $h^F(\cdot)$ . Instead we take a flexible approach, leaving the forms of the age - intake functions unspecified, but requiring that the estimated functions exhibit an acceptable degree of smoothness. Conventional kernel methods are awkward to employ in this problem in which (a) the functions for which smooth estimates are required appear in a non-linear model and (b) we have records of aggregate rather than individual data, the number of individuals varying across households. Instead we use roughness penalty methods (see Green and Silverman(1994)).

As noted in Chesher (1997) this problem is reminiscent of one studied in Engle, Granger, Rice and Weiss (1986) in which roughness penalty methods are applied to household electricity consumption data over *one month* billing periods in order to estimate the relationship between *daily* electricity demand and *daily* temperature. The ages and sexes of our household members correspond to the daily temperatures in the work of Engel *et al* and their monthly electricity totals correspond to the household nutrient acquisitions here.

The age data in the SUSENAS are recorded to the nearest whole year. Define binary variables identifying whole years of age of  $P$  household members as follows

$$z_{pa} = \begin{cases} 1 & , \quad a_p = a \\ 0 & , \quad otherwise \end{cases} \quad , \quad p = 1, \dots, P$$

and write the sex specific age profile functions as follows.

$$h^S(a_p) = \sum_{a=0}^{99} z_{pa} \beta_a^S, \quad S \in \{M, F\}. \quad (4)$$

The parameters  $\beta_a^M$  and  $\beta_a^F$  are the average rate of nutrient consumption by respectively males and females aged  $a$  years in a household with characteristics  $v_0$ .

Employing these specifications in (3) yields

$$E\left[Y - \left\{ \sum_{p=1}^P \left( n_p^M \beta_p^M + n_p^F \beta_p^F \right) \right\}^\gamma \exp(\delta'v) | x, v\right] = 0 \quad (5)$$

where  $n_p^S$  is the number of persons of age  $p$  and sex  $S$  in the household.

Even with the large amount of data provided by the SUSENAS the estimates of sex-age specific nutrient intakes we obtain using (4) in conjunction with (2) or (3) show excessive variation across ages. Accordingly we impose a roughness penalty during estimation, adding to the GMM minimand<sup>9</sup> a term

$$R(\lambda, \beta^M, \beta^F) = \lambda^2 \left( \beta^{M'} A' A \beta^M + \beta^{F'} A' A \beta^F \right) \quad (6)$$

where  $\beta^S = [\beta_0^S, \dots, \beta_{99}^S]'$ ,  $S \in \{M, F\}$ ,  $A$  is the  $98 \times 100$  matrix

$$A = \begin{bmatrix} 1 & -2 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1 \end{bmatrix}$$

<sup>9</sup>Linear (when  $\gamma$  and/or  $\delta$  are not being estimated) and non-linear (otherwise) least squares criteria are used throughout.

and  $\lambda \geq 0$  controls the severity of the roughness penalty.

Each of the terms  $\beta^{S'} A' A \beta^S$  is the sum of squared second differences of the sex-age specific nutrient intakes. As  $\lambda$  is increased, rough estimated profiles are increasingly penalised. This procedure is essentially a discrete analogue of the roughness penalty method employed in making smoothed estimates of continuous functions in which the roughness penalty is proportional to the integrated squared second derivative of the estimated function leading to the well known cubic spline estimator (Whaba (1990)).

When constant returns to scale are imposed ( $\gamma = 1$ ) estimation is very simple using standard least squares estimation software. One appends to the survey data  $2 \times 98$  additional pseudo-observations with values 0 for  $Y$  and for the variables in  $v$ , and with values given by  $\lambda$  times the rows of  $A$  for the counts of household members, giving a data structure as follows.

$$\mathcal{D} = \begin{bmatrix} y & N^M & N^F & v \\ 0 & \lambda A & 0 & 0 \\ 0 & 0 & \lambda A & 0 \end{bmatrix}$$

Here, with  $T$  denoting the number of household records,  $y$  is  $T \times 1$  and contains the observed household nutrient flows,  $N^S$ ,  $S \in \{M, F\}$  are  $T \times 100$  matrices containing counts of household members by age and sex, and  $v$  is  $T \times k$ , containing values of household characteristics. The pseudo-observations contribute precisely the amount (6) to the criterion<sup>10</sup>.

Imposition of a roughness penalty causes a reduction in the variances of estimates but necessarily imposes a cost, namely bias. Considering the model without household characteristics, and writing the data structure employed as

$$\mathcal{D} = \begin{bmatrix} y & X \\ 0 & \lambda W \end{bmatrix}$$

where e.g.

$$X = [ N^M \quad N^F ]$$

we have for the bias of the roughness penalised ordinary least squares estimates of the coefficients on  $X$

$$E[\hat{\beta}|X] - \beta = \left( \left( I + \frac{\lambda^2}{T} (T^{-1} X' X)^{-1} W' W \right)^{-1} - I \right) \beta$$

and for their variance matrix

$$V(\hat{\beta}|X) = \sigma^2 \left( I + \frac{\lambda^2}{T} (T^{-1} X' X)^{-1} W' W \right)^{-1} (X' X)^{-1} \left( I + \frac{\lambda^2}{T} W' W (T^{-1} X' X)^{-1} \right)^{-1}$$

where  $\sigma^2$  is the conditional variance of  $Y$  given  $x$  here assumed independent of  $x$ . Further discussion and results for the non-linear model in which household characteristics appear can be found in Chesher (1997).

The choice of  $\lambda$ , which controls the severity of the roughness penalty, is made subjectively, a sufficiently large value being chosen to produce acceptably smooth estimated age profiles. Chesher (1997) reports investigation of ‘‘optimal’’ choices of  $\lambda$  using cross-validation methods. The conclusion drawn there<sup>11</sup> is that, with the large samples available in this sort of problem, cross-validation methods resolve the trade-off between bias and variance in a way which leads to estimated age profiles that are far too rough given what we would expect the actual across age variation in nutrient intake rates to be.

<sup>10</sup>When  $\gamma$  is estimated it is necessary to stop the penalty  $R(\lambda, \beta^M, \beta^F)$  appearing in the criterion as  $R(\lambda, \beta^M, \beta^F)^\gamma$  but this is easily done in, for example PROC MODEL of SAS (1993).

<sup>11</sup>See also Silverman’s contribution to the discussion of Chesher (1997).

### 3. DATA

The data employed here consist of records obtained from 58,725 of the households who responded in the 1993 Survei Sosial Ekonomi Nasional (SUSENAS) of Indonesia<sup>12</sup>. The history and structure of the survey is described in Surbakti (1995). These records do not include any from households with 10 or more members or with missing values of any of the variables employed in the analysis<sup>13</sup>.

The households contributing records to the data set contained in total 257,886 members, an average of 4.4 per household, with ages ranging from 0 to 99. Frequency distributions of completed years of age for males and females in urban and rural areas are shown in Figures 1 and 2. There is some age heaping with unnaturally high relative frequencies at certain ages that are multiples of 5. The effect is not noticeable among urban dwellers at ages less than 20, but among rural dwellers it is noticeable from age 5 upwards.

With the help of enumerators, each household recorded amounts of around 200 foods consumed in the household during the week prior to the interview. As in all surveys that gather information by recall there is a possibility of under recording. This suggests that the results we obtain later on the composition of diet may be more accurate than the results on amounts of nutrients consumed.

The records of amounts of foods were processed using nutrient conversion factors developed for Indonesia to produce amounts of fat, carbohydrate, protein and energy entering the household during the recording period. All these amounts were converted to Kilocalories (Kcal) per day<sup>14</sup>. Considering all households we find that on average fat, protein and carbohydrate are sources of respectively 18%, 10% and 72% of energy<sup>15</sup>. For the households in the lowest decile of income per head these proportions are 13%, 9% and 77% and for households in the top decile they are 24%, 12% and 65%. This is as one might expect because carbohydrate is a relatively cheap source of energy. However some of this variation may reflect differences in household composition if the relative contributions of energy sources vary with sex and age. This is a matter on which the analysis to follow may shed some light.

Indonesia contained 27 provinces at this time<sup>16</sup> and in some of the analyses we include binary variables,  $P_i$ , identifying these. All analyses are conducted separately for urban and rural households<sup>17</sup>. Around 40% of households lived in urban areas.

Households recorded expenditure on food and other items entering the household and expenditure on services. The recall period for non-food items was one or twelve months depending on the item. At various points in the analysis we make use of the recorded values of total food expenditure, total expenditure and total household income, all expressed *per capita* in 1993 Rupiah per month. Food expenditure averages around 56% of total expenditure taking all households into account. For households in the lowest decile of per capita income, food accounts for around 68% of total expenditure. For households in the top decile of per capita income it accounts for 41%.

<sup>12</sup>The history and structure of the survey is described in Surbakti (1995).

<sup>13</sup>The SUSENAS contains a few households (1,917 out of a total of 59,642 - 1.3%) with between 10 and 18 members. Very large households were excluded mainly in order to reduce the scale of the estimation task. Around 50 households were removed because they had missing values for household income.

<sup>14</sup>Grams of fat, carbohydrate and protein were converted to Kcal of energy using the conversion factors respectively: 9, 3.75, and 4, and energy consumed by the household was calculated as the sum of the resulting amounts.

<sup>15</sup>By way of contrast, the proportion of energy from fat in Great Britain is currently a little under 40%.

<sup>16</sup>East Timor was covered in the 1993 SUSENAS.

<sup>17</sup>Note that Jakarta is treated as a province and has no rural area.

Figure 1: Relative frequency distributions of ages of urban males (52,013) and females (53627)

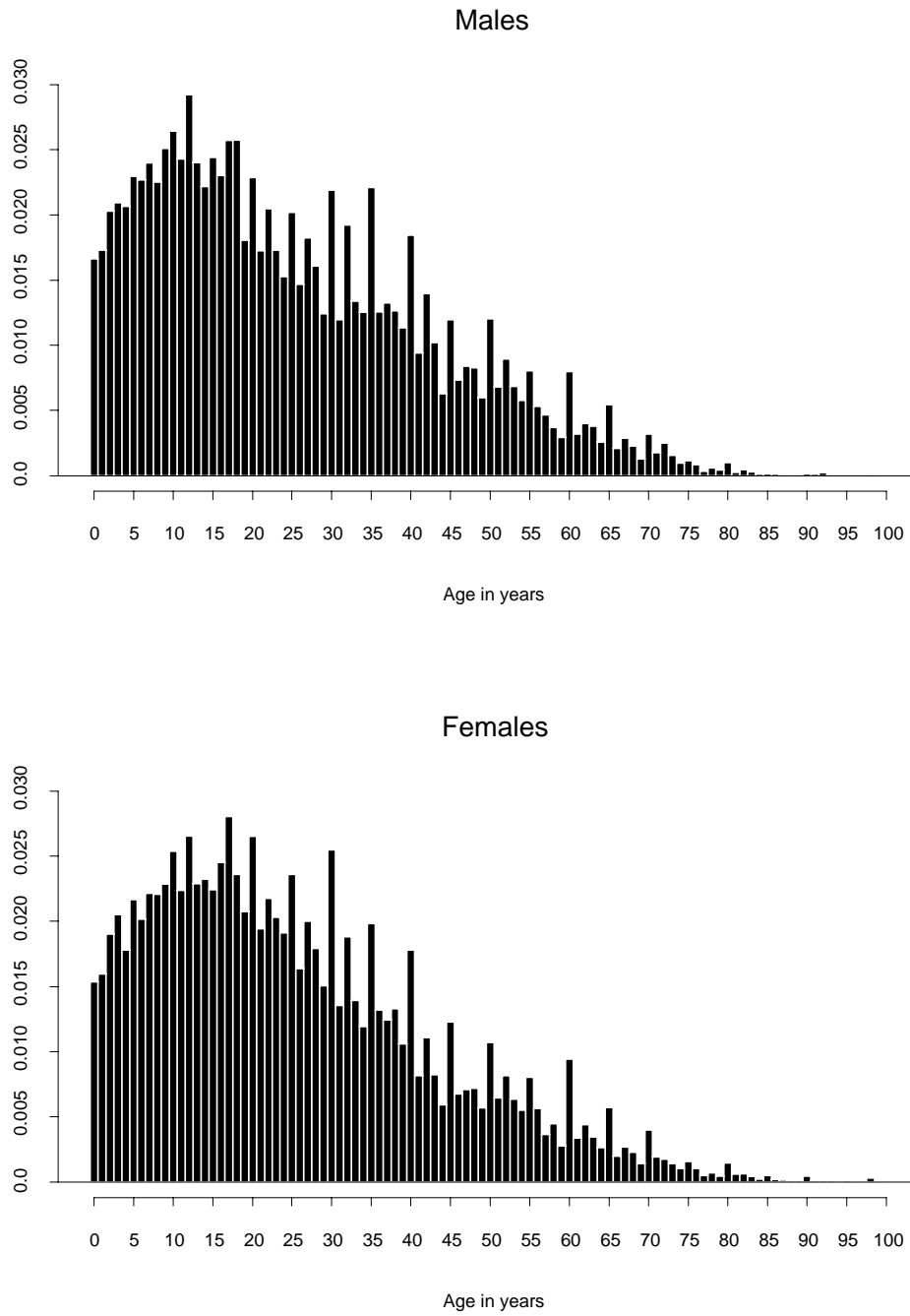
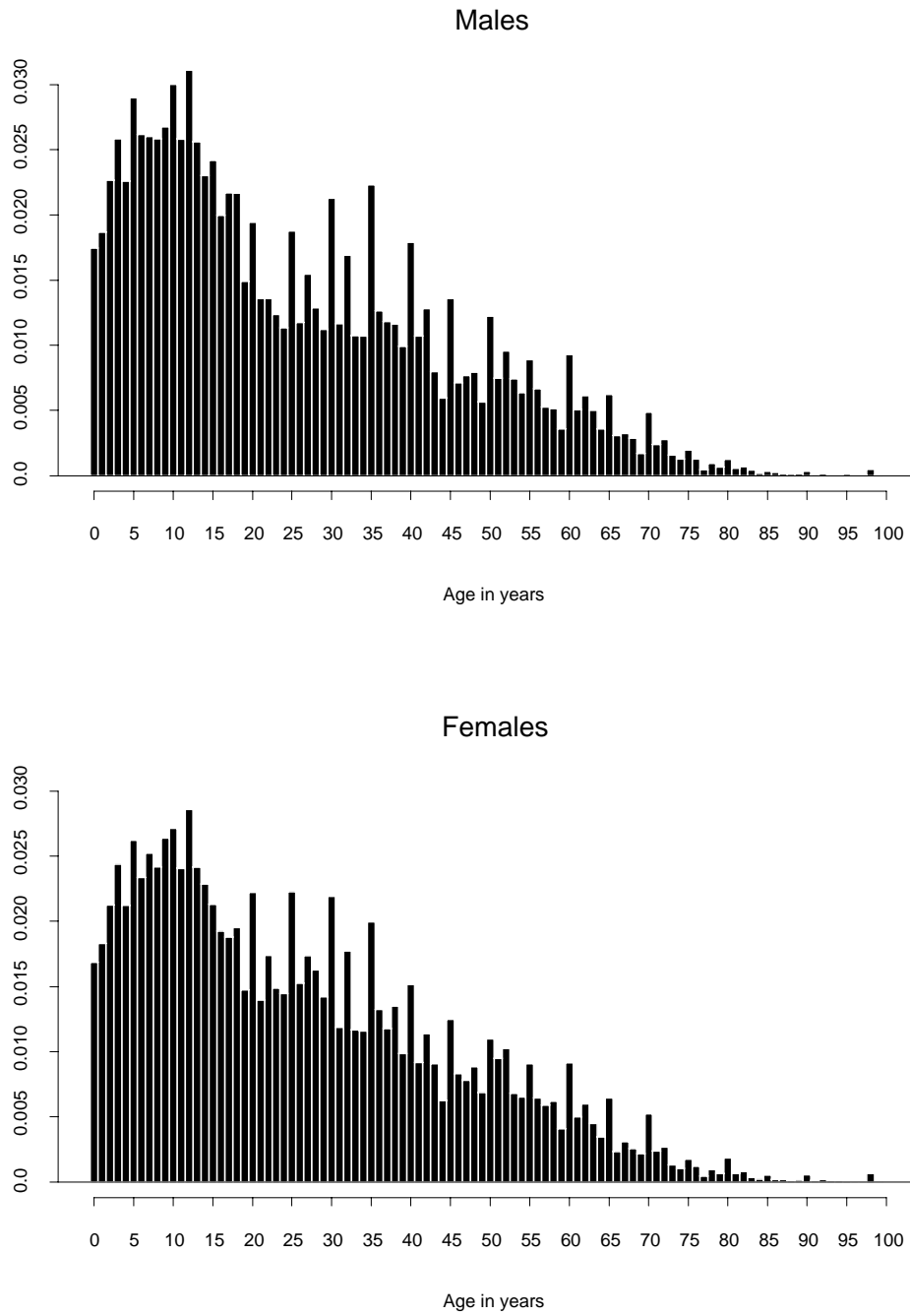


Figure 2: Relative frequency distributions of ages of rural males (76,015) and females (76,213)



## 4. RESULTS

In this Section results obtained using the method described in Section 2 and the SUSENAS data are presented.

**4.1. Energy intakes: Smoothing and accuracy.** The first results in Figure 3 show the impact of smoothing. Data on energy consumption are analysed separately for urban and rural dwellers with no control for any household characteristics. The dashed, erratic lines in Figure 3 show age profiles estimated with minimal smoothing<sup>18</sup> ( $\lambda = 2$ ); the solid lines show age profiles estimated with  $\lambda = 200$ . The general shape of the age profiles are evident even in the non-smoothed estimates but the smoothing clearly brings benefits. As one would expect the non-smoothed estimates are much more erratic at higher ages<sup>19</sup> where the data are less informative (see Figures 2 and 1).

In both urban and rural areas estimated energy consumption is higher for males than for females at all almost all ages though the difference is very small for people less than 20 years of age and is smaller in rural than in urban areas. Among rural dwellers the profiles are relatively flat from age 20 to 40 but for urban males at least, there is a steady increase in energy consumption over this age range. All the age profiles decline as age rises above about 50 years.

Standard errors for the non-smoothed estimates increase steadily with age from around 60 Kcal/person/day at age 5 to around 250 Kcal/person/day at age 80. Smoothing increases accuracy substantially (of course at the cost of some bias), standard errors for smoothed estimates increasing from around 20 Kcal/person/day at age 5 to around 90 Kcal/person/day at age 80. Figure 4 shows smoothed estimated age profiles for males and females with pointwise 90% confidence intervals superimposed.

Comparing males and females there is a clear overlap in the intervals up to around age 20 but at higher ages the intervals are distinct suggesting that the estimated differences between the male and female energy intake age profiles are revealing a genuine gender difference<sup>20</sup>. Figure 5 replots these curves putting results for males in the upper pane and results for females in the lower pane. It is clear that, up to around age 35, the estimated energy intake age profiles are significantly higher for males and for females in rural areas than in urban areas. Among older people the difference is somewhat marginal but seems to persist among females.

**4.2. Sources of energy: carbohydrate, fat and protein.** The results in Section 4.1 pertained to energy intakes. This Section provides estimates of intakes of the three *sources* of energy: fat, protein and carbohydrate, all expressed in Kcal/person/day. We continue to present separate results for urban and rural dwellers. As before there is no other control for household characteristics.

Figures 6, 7 and 8 show estimated age profiles for respectively fat, protein and carbohydrate, results for urban dwellers in the upper panes, results for rural dwellers in the lower panes, each estimated age profile accompanied by pointwise 90% confidence intervals.

There are clear differences both across males and females and across urban and rural dwellers. Figure 6 suggests that rural dwellers have slightly higher carbohydrate intakes than urban dwellers at most ages, and that the amount by which males' intakes are higher than females is larger among urban dwellers, except among older people. The age profile

<sup>18</sup>With the small value, but non-zero value  $\lambda = 2$  essentially the same results are produced as with no smoothing at all, except that the very few ages (in the 90's) at which no household members are recorded are "bridged", allowing a complete age profile to be produced.

<sup>19</sup>Profiles are not plotted for ages exceeding 80 years because the limited amount of data causes even the smoothed estimates to become unreliable.

<sup>20</sup>When we calculate profiles of the difference between males and females estimated intakes with 90% pointwise intervals we find that the intervals do not contain zero at higher ages.

Figure 3: Estimated smoothed and non-smoothed age profiles of energy intakes, male and female urban and rural dwellers

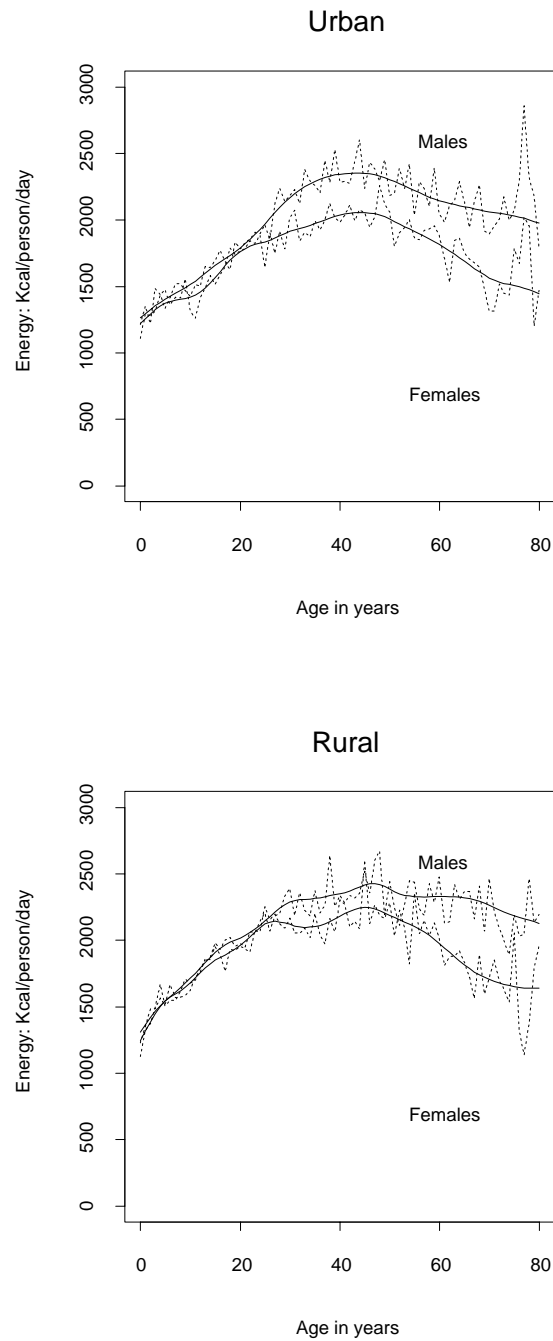


Figure 4: Smoothed estimates of energy intake age profiles, males and females in urban and rural areas with pointwise 90% confidence intervals

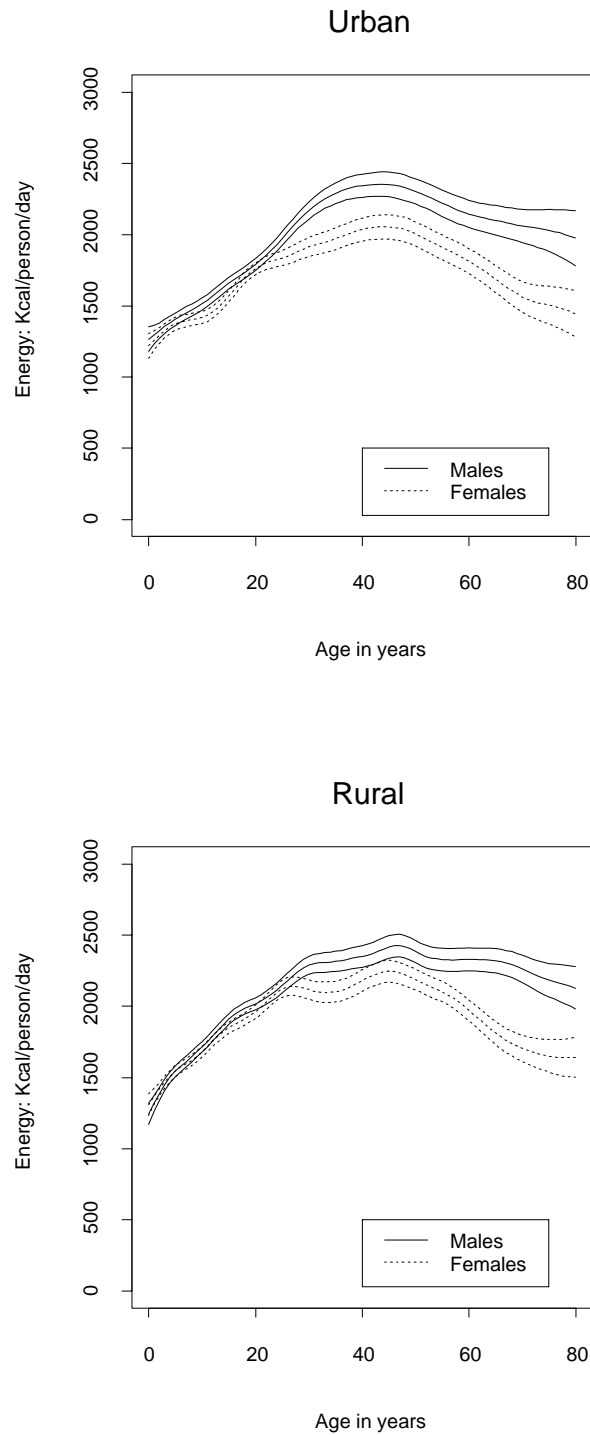


Figure 5: Smoothed estimates of energy intake age profiles, males and females in urban and rural areas with pointwise 90% confidence intervals

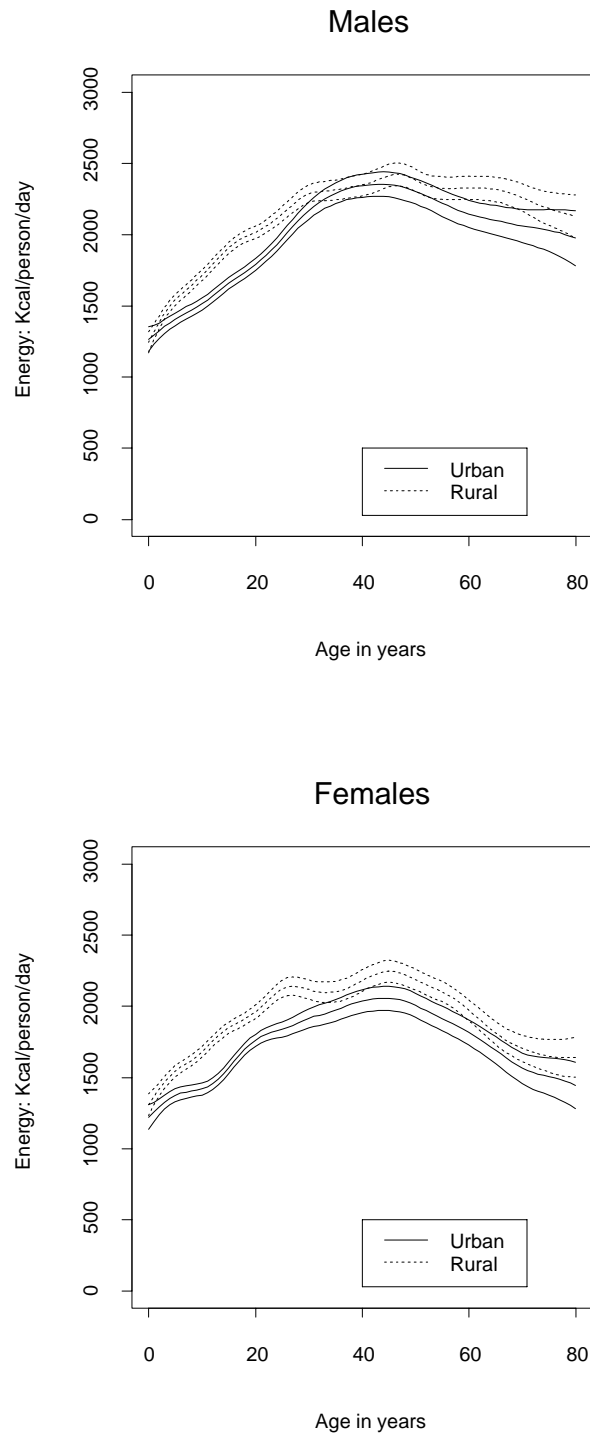


Figure 6: Estimated intakes of carbohydrate (Kcal/person/day) with pointwise 90% intervals

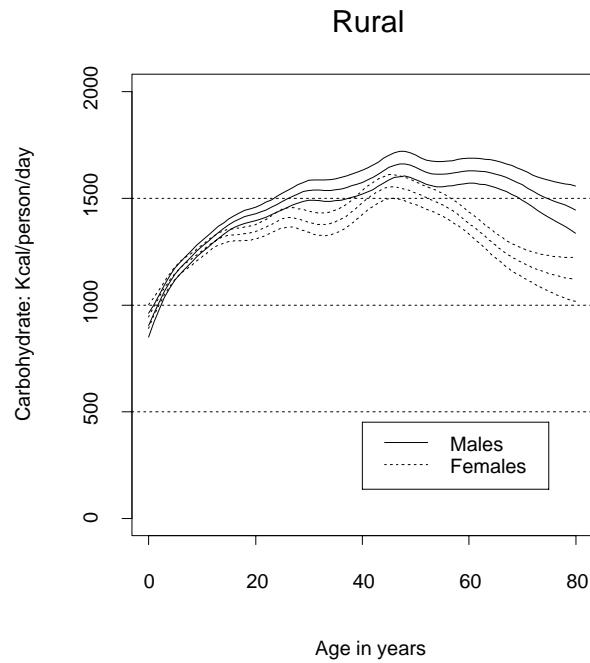
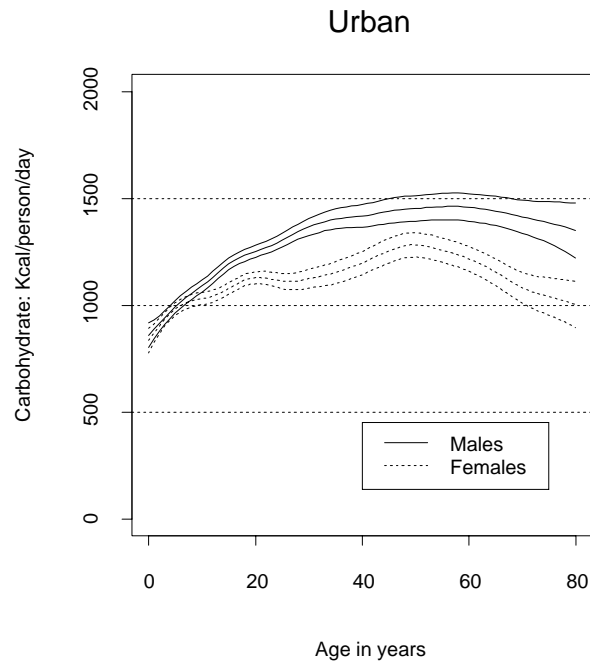


Figure 7: Estimated intakes of fat (Kcal/person/day) with pointwise 90% intervals

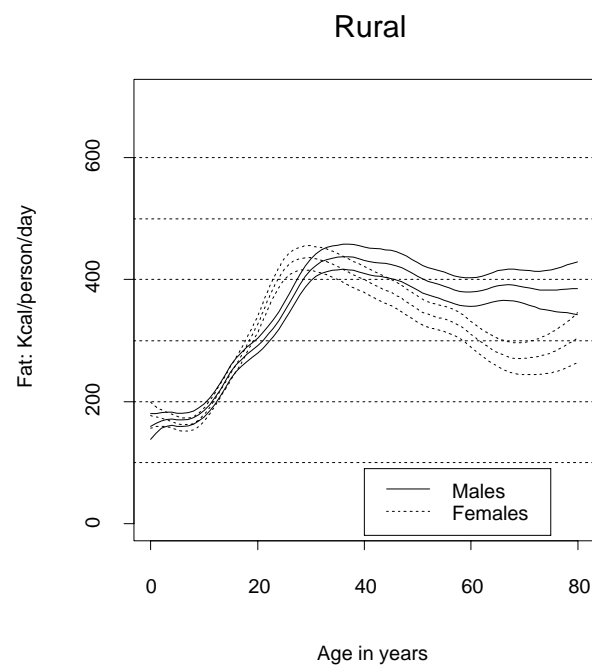
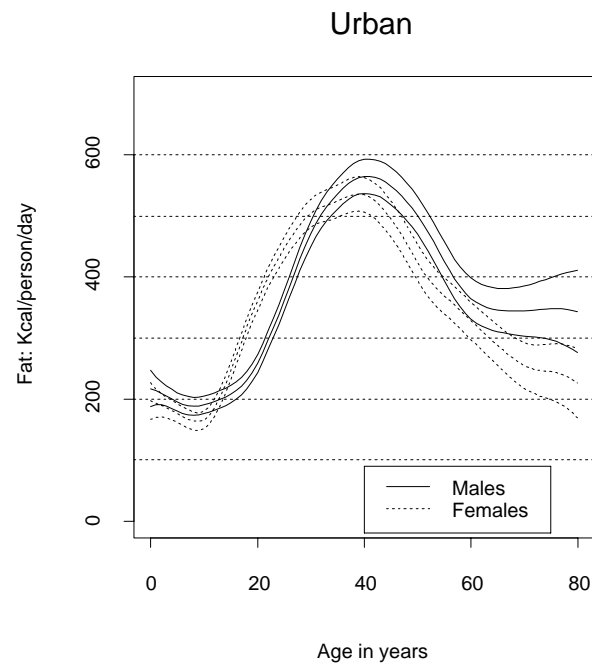
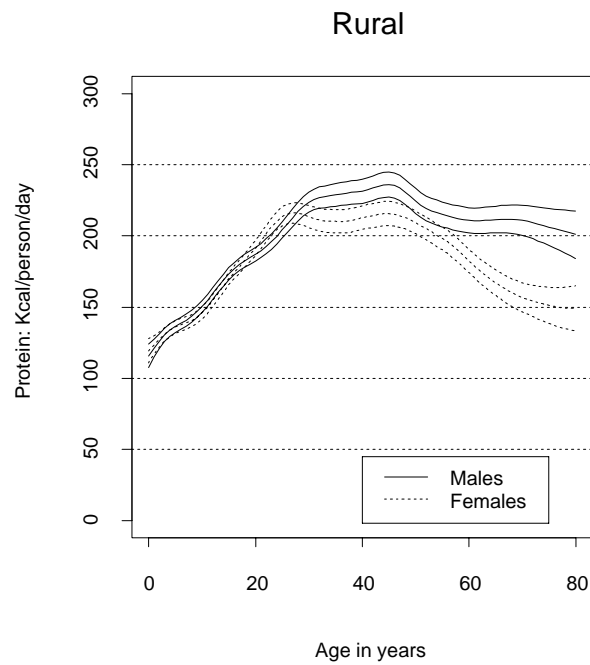
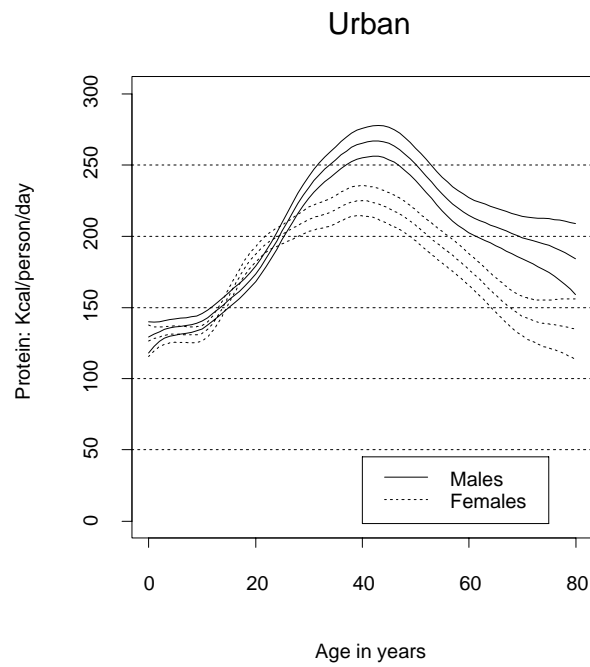


Figure 8: Estimated intakes of protein (Kcal/person/day) with pointwise 90% intervals



of fat intakes (Figure 7) is much more peaked, and peaks higher, among urban dwellers than among rural dwellers. Among urban dwellers, females' fat intakes exceed males' from around age 16 to 35. There is a similar feature in the results for rural dwellers but for a shorter age interval. In both cases the age interval over which these male - female differences arise falls during the major period of child bearing. Figure 8 suggests that protein intakes are similar for males and females up to around age 25, females' intakes falling below males' at older ages.

Some of these variations mirror those found for energy intakes - compare Figures 6, 7 and 8 with Figure 3 which gives the energy intake results. However there are differences in the age profiles and to gain more information on these we now consider the proportions of energy obtained from the three energy sources. These are calculated as simple ratios of the estimated nutrient intakes relative to the estimated energy intakes. Figures 9, 10 and 11 show the resulting age profiles accompanied by pointwise 90% confidence intervals<sup>21</sup>.

From early childhood to around 10 years the proportion of energy from carbohydrate rises, then falling to a low at around age 40, thereafter rising again. The changes across ages in the proportions of energy from fat and protein both mirror these changes. The amplitudes of these variations are larger among urban dwellers than among rural dwellers. The proportion of energy from fat and protein for females exceeds the corresponding proportions for males between ages 15 and 35 in urban and rural areas. The proportion of energy from carbohydrate is higher at most ages among rural than among urban dwellers. This may be associated with different levels of income in urban and rural areas, a point we shall return to shortly.

**4.3. Economies of scale.** The intention of the enumerators in the SUSENAS was to record amounts of foods consumed by households but it is likely that what is in fact recorded is the amount of food employed in making food for consumption. It is possible that where there are larger nutrient demands, and so larger amounts of food processed, there are economies of scale in the production of food for consumption from the raw materials brought into the household. This might occur because there is less waste as a proportion of food processed, the more food is processed.

To obtain some idea of the existence and magnitude of scale economies the model employed earlier was re-estimated in the following form, with no control for household characteristics.

$$E\left[Y - \left\{ \sum_{p=1}^P \left( n_p^M \beta_p + n_p^F \beta_p^F \right) \right\}^\gamma \mid x\right] = 0$$

The estimate of  $\gamma$  obtained for urban and rural dwellers using the records of household energy consumption ( $Y$ ) are respectively (estimated standard errors in parentheses) 0.86 (0.005) and 0.88 (0.004). Both estimates are significantly less than one. One interpretation of this result is that there are indeed economies of scale, each 10% increase in the energy demands of a household requiring an increase of around 8.7% in the amount of energy in food prepared for consumption. However some care is required here. Suppose that households with many members have a lower proportion of members working than households with a small number of members, perhaps because large households tend to have many children amongst their members. If working entails higher expenditure of energy then we would expect to see the sort of effect that these estimates of  $\gamma$  imply.

**4.4. Household characteristics.** The association between poverty and nutrition is of great interest. In this section we consider how energy intakes and intakes from the three sources of energy vary with household income. First we present results obtained using a

<sup>21</sup>The intervals use estimated standard errors calculated using the delta method, taking into account the (generally positive) correlations between estimated intakes of energy and of its components.

Figure 9: Proportion of energy from carbohydrate with pointwise 90% intervals

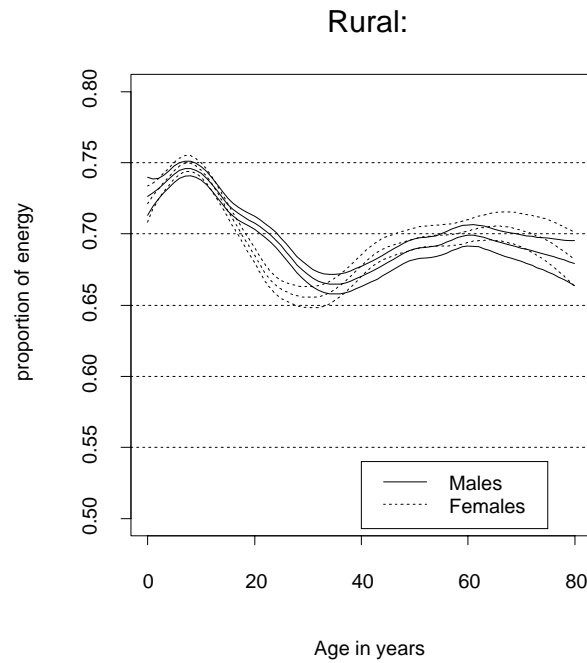
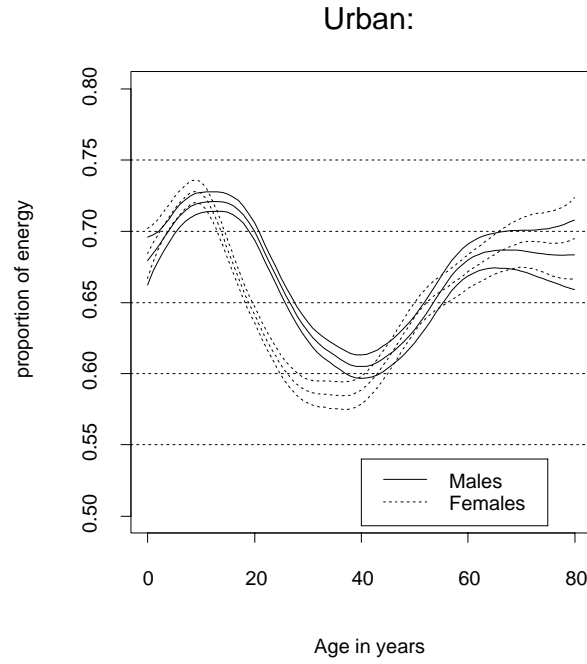


Figure 10: Proportion of energy from fat with pointwise 90% intervals

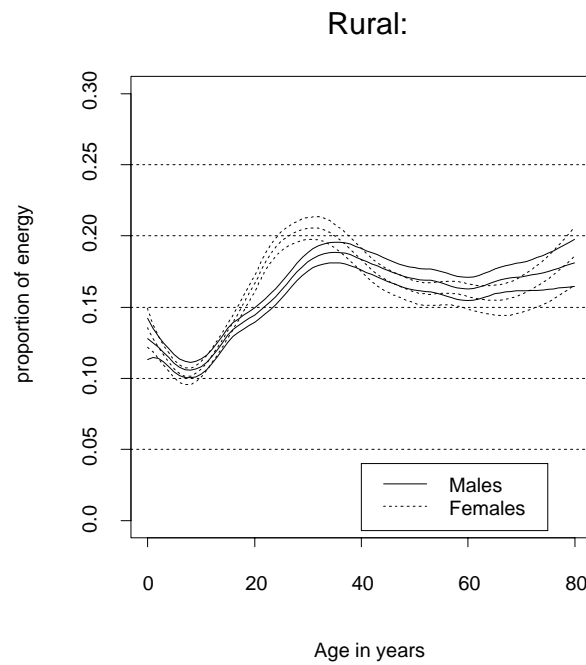
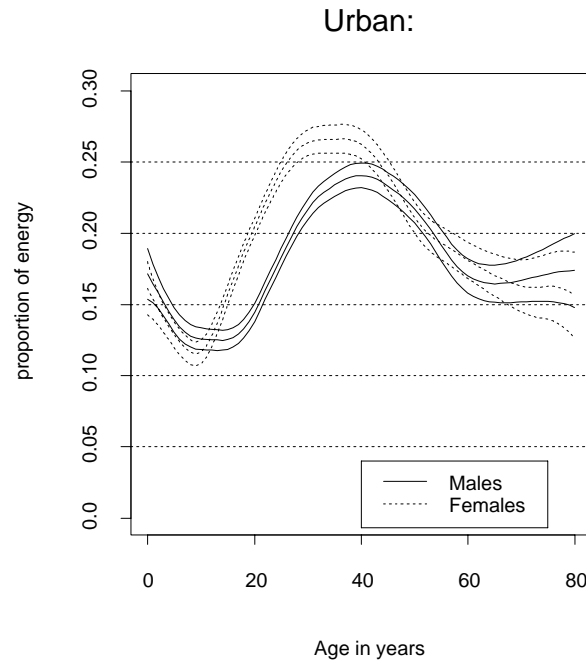


Figure 11: Proportion of energy from protein with pointwise 90% intervals

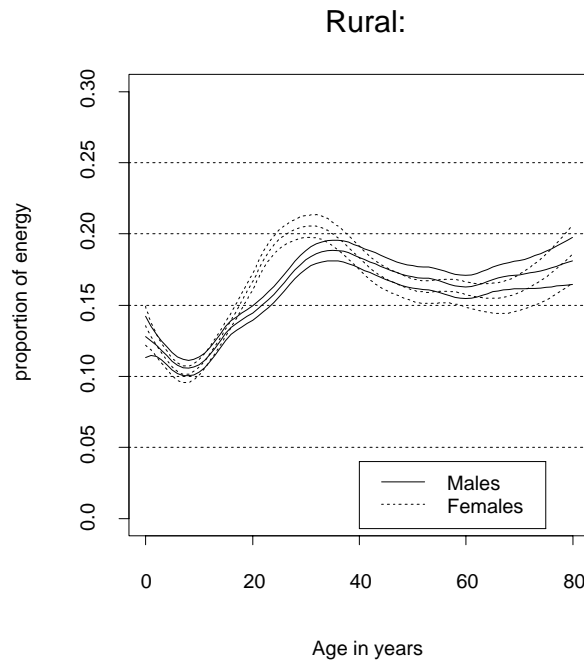
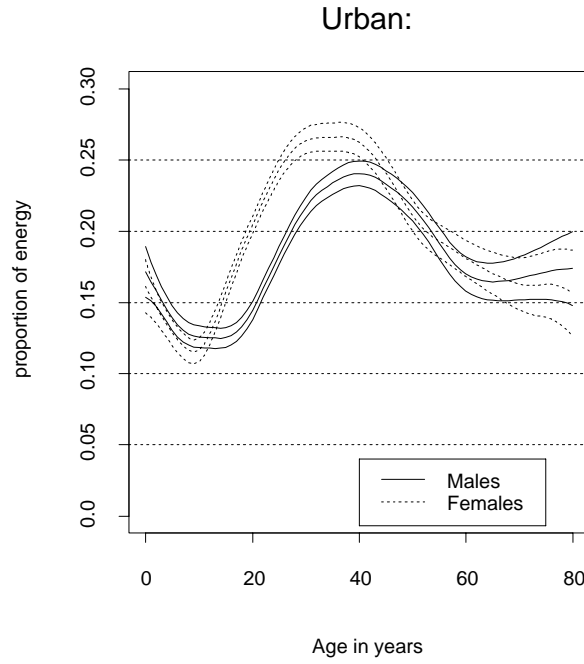


Table 1: Estimated coefficients on log income per head (estimated standard errors)

Nutrient	Urban	Rural
Energy	0.133 (.0025)	0.145 (.0023)
Carbohydrate	0.068 (.0027)	0.092 (.0026)
Fat	0.302 (.0036)	0.324 (.0035)
Protein	0.210 (.0029)	0.224 (.0025)

Table 2: Estimated coefficients on log income per head and its square (estimated standard errors)

Nutrient	Urban		Rural	
	log income	(log income) <sup>2</sup>	log income	(log income) <sup>2</sup>
Energy	0.200 (.0085)	-0.014 (.0018)	0.248 (.0069)	-.030 (.0021)
Carbohydrate	0.094 (.0089)	-0.005 (.0020)	0.170 (.0075)	-0.024 (.0023)
Fat	0.683 (.0162)	-0.072 (.0033)	0.650 (.0130)	-.081 (.0035)
Protein	0.363 (.0105)	-0.031 (.0022)	0.393 (.0080)	-0.046 (.0023)

simple representation of the income effect, including log household income per head as a variable in  $v$  in the model (2), including additionally indicator variables identifying the 27 provinces<sup>22</sup> of Indonesia. The results are shown in Table 1. In producing these results the scale parameter  $\gamma$  was set to one.

When the relationship between nutrient intakes and permanent income is of main interest it is common to find total expenditure used as a measure of permanent income rather than income. When log expenditure per head is used in place of log income per head we find somewhat greater sensitivity. This may be because for many households food expenditures make up a large fraction of total expenditure. In this situation transitory shocks to food intakes, due e.g. to weddings, visitors and temporary absence of household members, will result in transitory shocks to total expenditure. One possible attack on this problem is to employ total expenditure rather than income and to instrument using for example, measures of housing quality and durable ownership which are available in the SUSENAS. This is one of the items on the research agenda.

The estimated income coefficients are all significantly greater than zero and they are much larger<sup>23</sup> for intakes of fat and protein than for carbohydrate. The income coefficient for energy intakes, as one would expect falls in the range of the other coefficients (energy intakes are the sum of carbohydrate, fat and protein intakes) and closer to the coefficients on the dominant energy source, carbohydrate, than to the others.

These results suggest that among urban dwellers a 10% difference in household income per head is associated with differences at all ages for males and females in intakes of energy,

<sup>22</sup>26 for rural dwellers since Jakarta has no rural areas.

<sup>23</sup>The differences are all statistically significantly different from zero using a size 0.001 test.

Table 3: Estimated income elasticities of nutrient intakes at 1st and 9th decile and median income (by urban and rural) for urban and rural dweller

Nutrient	Urban			Rural		
	1st decile	median	9th decile	1st decile	median	9th decile
Energy	.16	.15	.10	.18	.16	.14
Carbohydrate	.08	.08	.06	.14	.11	.07
Fat	.53	.42	.15	.54	.45	.32
Protein	.30	.25	.13	.33	.28	.21

carbohydrate, fat and protein of respectively 1.3%, 0.7%, 2.9%. The corresponding figures for rural dwellers are slightly larger<sup>24</sup>. Thus 10% higher income per head is associated with a difference in the proportion of energy from carbohydrate of  $-0.6\%$  and differences in the proportion of energy obtained from fat and protein of respectively  $+1.6\%$  and  $+0.7\%$ .

Further investigation shows that there is more concavity in the nutrient intake - household income relationship than the constant elasticity form used above allows. Table 2 shows estimates of income coefficients when additionally the square of log income per head is included in the list of household characteristics,  $v$ . In each case the coefficient on squared log income per head is negative and statistically significantly different from zero, suggesting that income elasticities of nutrient intakes are higher among lower income households than among higher income households. Table 3 shows the income elasticities at 1st and 9th decile and median income (quantiles calculated separately for urban and rural dwellers) for each of the nutrients. There are clearly substantial variations with income.

So far household characteristics have been entered in a single multiplicative term so that different levels of, for example income, result in the same proportionate effect on the average nutrient intakes of all household members. It is interesting to see whether there is evidence to suggest that different types of household members (e.g. young and old, male and female) have unequal differences in average nutrient intakes when low and high income households are compared. In order to do so, and focussing here on differences across males and females, the model is extended, estimation being based on the following moment condition

$$E[Y - \sum_{p=1}^P (m_p h^M(a_p) \exp(\delta'_M v) - f_p h^F(a_p) \exp(\delta'_F v)) | x, v] = 0$$

where, as before, the age effects are represented by step functions with changes at each integer year of age and smoothness in the estimated age profiles is induced by means of a roughness penalty. The resulting estimates of the male and female specific coefficients on log income per head are shown in Table 4. The estimated income elasticities for females are generally slightly smaller than those estimated for males indicating smaller differences in nutrient intakes for females than for males when we compare high and low income households. However the differences are generally small. The differences are in most cases significantly different from zero, but this in part reflects the very large samples being employed here (35,539 rural and 23,538 urban households).

As explained in the introduction, the positive income coefficient obtained for energy intakes may be a consequence of variation across households in labour force activities.

<sup>24</sup>The coefficients for urban and rural dwellers are statistically significantly different from one another, but this mainly reflects the accuracy of estimation in this large sample of what are in practical terms rather small differences.

Table 4: Estimated male and female specific coefficients on log income per head (estimated standard errors) and t statistics for the hypothesis of equal coefficients

Nutrient	Urban			Rural		
	Male	Female	t statistic for $H_0$ : $\delta^M = \delta^F$	Male	Female	t statistic for $H_0$ : $\delta^M = \delta^F$
Energy	0.144 (.0073)	0.133 (.0075)	2.17	0.162 (.0071)	0.142 (.0075)	4.28
Carbo- hydrate	0.085 (.0077)	0.062 (.0082)	4.14	0.111 (.0078)	0.082 (.0083)	5.54
Fat	0.314 (.0117)	0.307 (.0108)	0.93	0.346 (.0120)	0.327 (.0116)	2.39
Protein	0.213 (.0083)	0.218 (.0086)	-0.87	0.238 (.0078)	0.224 (.0080)	2.78

Perhaps the small male - female differentials in sensitivity to household income have a similar cause.

The difference between an income coefficient for a nutrient (e.g. fat) and the income coefficient for energy measures the sensitivity to income variation of the proportion of energy obtained from the nutrient. Interestingly these differences are in all cases very similar for males and females. Even though the total amount of dietary energy seems to be less sensitive to income for females than for males, males' and females' dietary composition shows very similar variation with income.

## 5. CONCLUDING REMARKS

This paper has applied a technique of statistical disaggregation to data on Indonesian households' food consumption to obtain estimates of average rates of nutrient intake for males and females at each year of completed age in rural and urban areas. There are marked differences across males and females, across ages and across urban and rural dwellers. Two of the three sources of food energy, fat and protein, show much greater sensitivity to variation in household income than does the third energy source, carbohydrate. Females seem to obtain higher proportions of energy from fat and protein during child bearing years. Male nutrient intakes seem to be slightly more sensitive to variation in household income than female intakes, but the proportion of energy obtained from the three energy sources varies with income in a similar fashion for males and females.

Provided that the assumptions underlying our procedures are valid, we should regard the methods used here as having uncovered some "facts" about individuals' rates of nutrient intakes from records of household rates of intake. These "facts" are of course capable of a variety of interpretations. For example, the positive association between energy intakes and household income may be being driven by labour market factors, energy expenditure and income being higher among labour market participants than among others. The differential sensitivity of fat, protein and carbohydrate intakes to income perhaps reveals features of the demand for energy sources which are available at different prices. To understand these issues and to draw any policy implications from the "facts" presented here requires development of a model for *individual* dietary choice and labour

force participation and development of methods for estimating such a model using the *household* level, aggregate, data which is all that is normally available. This is the next task in this research agenda.

#### APPENDIX: AGE HEAPING

The method used in this paper makes extensive use of the age data recorded in the SUSENAS but there is clearly some mis-recording of ages in this survey. The age relative frequency distributions in Figures 1 and 2 show clear spikes at many ages which are multiples of 5. In this Appendix we consider the impact of this heaping on the estimated age profiles, propose a model for age heaping and show how it can be used to examine the sensitivity of estimated profiles to different assumptions about age heaping.

**5.1. General considerations.** We consider the impact of age heaping in the context of a problem, like that set out in the main part of this paper, in which age is a discrete variate, recorded in completed years of age. To avoid overly complex notation, no distinction between males and females is made for the moment. Let  $X = [X_0, \dots, X_{99}]$  and  $Z = [Z_0, \dots, Z_{99}]$  be vectors of binary random variables with<sup>25</sup>  $\iota'X = \iota'Z = 1$ , such that for a randomly sampled person of actual age  $i$ , and recorded age  $j$ ,  $X_i = Z_j = 1$ . Let  $a' = [0, 1, \dots, 99]$ . Then actual and recorded ages are respectively  $a_X = X'a$ ,  $a_Z = Z'a$ .

Age heaping is a form of measurement error in which an age close to a value in a set of heaping points  $\mathcal{K}$ , (e.g.,  $\mathcal{K} = \{5, 10, 15, \dots, 95\}$ ) may be recorded at the heaping point. Ages recorded at non-heaping points are accurately recorded but ages recorded at heaping points may not be. Formally for  $k \notin \mathcal{K}$ ,  $P[X_k = 1|Z_k = 1] = 1$  and for  $k \in \mathcal{K}$ ,  $P[X_k = 1|Z_k = 1] \leq 1$ . In a simple age heaping process, ages which are actually at heaping points are accurately recorded, that is, for  $k \in \mathcal{K}$ ,  $P[Z_k = 1|X_k = 1] = 1$ . In a more complex process one might have ages close to a multiple of 5 sometimes recorded as the multiple of 5 and other times recorded accurately, but some ages that are multiples of 5 but not of 10 sometimes being recorded as the nearest multiple of 10.

First we consider the impact of age heaping on the regression of some variate  $Y$  on  $X$ . That is we consider how the regression of  $Y$  on  $Z$ , which is what we would estimate using conventional estimation methods with data on  $Y$  and  $Z$ , relates to the regression of  $Y$  on  $X$ , which is what we would like to learn about. Later we will consider the impact of age heaping on the regression of  $Y$  on  $X$  and another set of covariates,  $V$ , the household characteristics in the model for nutrient intakes.

**The regression of  $Y$  on  $X$  alone.** Our model for the average rate of consumption of a person with  $X = x$  is

$$E[Y_p|X = x] = x'\beta$$

from which it follows directly that expected consumption for a person with recorded age  $Z = z$  is<sup>26</sup>

$$E[Y_p|Z = z] = E[X|Z = z]'\beta.$$

Let  $g_{ij} = P[X_i = 1|Z_j = 1]$  and define the matrix  $G = [g_{ij}]$ . Then, noting that  $X$  and  $Z$  are vectors of binary variables with  $\iota'X = \iota'Z = 1$ ,

$$E[X|Z = z] = Gz$$

<sup>25</sup> $\iota$  is a 100 element vector of ones.

<sup>26</sup>We assume that the age heaping is a pure measurement error process in the sense that  $E[Y|X = x, Z = z] = E[Y|X = x]$ .

and so

$$E[Y|Z = z] = z'G'\beta = z'\gamma$$

where  $\gamma = G'\beta$ . If we had data on individual consumption and estimated its regression on the mis-recorded age indicators,  $z$ , then we would estimate an intake at age  $j$ ,  $\gamma_j$ , which is a weighted average of the intakes at age  $j$  and at other ages, that is:

$$\gamma_j = \sum_i P[X_i = 1|Z_j = 1]\beta_i.$$

When there is age heaping, all recorded ages that are observed at ages other than heaping points (e.g. multiples of 5) are error free, but some of the ages recorded at heaping points are incorrect. It follows that, when there is age heaping the coefficients in the regression of  $Y$  on  $Z$  at ages  $k \notin \mathcal{K}$  are equal to the corresponding coefficients in the regression of  $Y$  on  $X$ , that is

$$\gamma_j = \beta_j \quad , j \notin \mathcal{K}.$$

At ages in the set of heaping points  $\gamma_k$  is a weighted average of the  $\beta_i$ 's where  $i$  runs through the set of ages whose values can be mis-recorded as  $i$ . We might expect these to be limited to ages close to  $k$ , suggesting that age heaping induces a degree of smoothness in the  $\gamma_i$ 's relative to the  $\beta_i$ 's.

**Aggregation.** Letting  $Y_h$  be the average rate of consumption by a household with  $m$  members with recorded ages indicated by  $z_1, \dots, z_m$  and let  $n = \sum_{p=1}^m z_p$  which is a vector of counts of household members at each integer year of recorded completed age.. Then the expected rate of consumption of the household, conditional on this recorded age composition is

$$E[Y|z_1, \dots, z_m] = \sum_{p=1}^m E[Y_p|Z = z_p] = \sum_{p=1}^m z'_p G' \beta = n' \gamma.$$

So, when there is age heaping, least squares estimation of the regression of  $Y$  on  $n$  will produce an estimate of  $\gamma = G'\beta$  which will only be equal to  $\beta$  when  $G = I$  in which case there is no age heaping. Applying the roughness penalty method will produce smoothed estimates of  $\gamma$  rather than of  $\beta$ .

If we knew the matrix  $G$ , and if it were non-singular then we could retrieve estimates of  $\beta$  using  $\hat{\beta} = G'^{-1}\hat{\gamma}$ . Of course  $G$  may be singular in which case all element of  $\beta$  cannot be identified from age heaped data but there will be linear combinations of the elements that are identifiable. Alternatively one could obtain estimates of identifiable elements of  $\gamma$  directly by estimating the regression of  $Y_h$  on  $Gn$  rather than on  $n$ . Indeed when applying the roughness penalty method this would a better course of action, for then one would be smoothing estimates of  $\beta$  rather than  $\gamma$ .

Now suppose there are different age profiles for males and females with average rates of age specific consumptions given by  $\beta^F$  and  $\beta^M$  and suppose, as may be reasonable, that the age mis-recording process is identical for males and females. Assume also that sex is not mis-recorded. Then, arguing as above, the regression of  $Y$  on  $n_M$  and  $n_F$  has coefficients  $\gamma^M = G'\beta^M$  and  $\gamma^F = G'\beta^F$  and it is these that will be estimated when age heaped data are employed.

**The regression of  $Y$  on  $X$  and  $V$ .** In our model for nutrient intakes including household characteristics, the average rate of consumption by a person with actual age indicated by  $X = x$  and household characteristics  $V = v$  is

$$E[Y|X = x, V = v] = (x'\beta) \exp(\delta'v)$$

from which it follows directly that expected consumption for a person with recorded age  $Z = z$  and  $V = v$  is<sup>27</sup>

$$E[Y|Z = z, V = v] = (E[X|Z = z, V = v]'\beta) \exp(\delta'v)$$

Let  $g_{ij}(v) = P[X_i = 1|Z_j = 1, V = v]$  and define the matrix  $G(v) = [g_{ij}(v)]$ . Then, as before,

$$E[X|Z = z, V = v] = G(v)z$$

and it follows that

$$E[Y|Z = z, V = v] = (z'G(v)'\beta) \exp(\delta'v).$$

The difficulty that arises here is that  $G(v)$  will typically depend upon  $v$ , even if we suppose that the age heaping process is independent of  $V$  in the sense that

$$P[Z_j = 1|X_i = 1, V = v] = P[Z_j = 1|X_i = 1].$$

We will return to this in the context of the specific model of age heaping introduced now.

**5.2. A model for age heaping.** The simple model for age heaping studied here has ages that are multiples of 5 being recorded correctly and ages within 2 years of a multiple of 5 being recorded as the multiple of 5 with some probabilities between zero and one. In particular suppose that an age within one (two) year(s) of an age  $k$  which is a multiple of 5, is recorded correctly with probability  $\alpha_k^1$  ( $\alpha_k^2$ ) and is recorded as age  $k$  with probability  $1 - \alpha_k^1$ , ( $1 - \alpha_k^2$ ). So, for  $k \in \mathcal{K} = \{5, 10, 15, \dots, 95\}$ , we have the following, for  $j = 1, 2$ .

$$\begin{aligned} P[Z_{k-j} = 1|X_{k-j} = 1] &= \alpha_k^j \\ P[Z_k = 1|X_{k-j} = 1] &= 1 - \alpha_k^j \\ P[Z_k = 1|X_k = 1] &= 1 \\ P[Z_k = 1|X_{k+j} = 1] &= 1 - \alpha_k^j \\ P[Z_{k+j} = 1|X_{k+j} = 1] &= \alpha_k^j \end{aligned}$$

When  $\alpha_k^1 = \alpha_k^2 = 0$  for all  $k \in \mathcal{K}$ , then there is complete age heaping, equivalent to grouping ages into 5 year intervals<sup>28</sup> and recording the interval midpoints instead of the actual ages. When  $\alpha_k^1 = \alpha_k^2 = 1$  for all  $k$  then ages are recorded accurately. Figures 1 and 2 suggest that the SUSENAS presents a situation somewhere between these two extremes.

Let  $C$  be the matrix with elements<sup>29</sup>  $c_{ij} = P[Z_j = 1|X_i = 1]$ . Then  $C$  will be block diagonal taking the form<sup>30</sup>

$$C = \begin{bmatrix} I_3 & 0 & 0 & \dots & 0 & 0 \\ 0 & C_5 & 0 & & 0 & 0 \\ 0 & 0 & C_{10} & & 0 & 0 \\ & 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & & C_{95} & 0 \\ 0 & 0 & 0 & \dots & 0 & I_2 \end{bmatrix} \quad (7)$$

<sup>27</sup>Again we assume that the age heaping is a pure measurement error process in the sense that

$$E[Y|X = x, Z = z, V = v] = E[Y|X = x, V = v].$$

<sup>28</sup>For example [8, 12], [13, 17] with mid points 10, 15.

<sup>29</sup>Note that  $i$  indexes rows and  $j$  indexes columns.

<sup>30</sup>We assume for simplicity that ages 0, 1, 2, 98, and 99 are recorded correctly with probability 1.

where  $I_n$  is a  $n \times n$  identity matrix and

$$C_k = \begin{bmatrix} \alpha_k^2 & 0 & 1 - \alpha_k^2 & 0 & 0 \\ 0 & \alpha_k^1 & 1 - \alpha_k^1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 - \alpha_k^1 & \alpha_k^1 & 0 \\ 0 & 0 & 1 - \alpha_k^2 & 0 & \alpha_k^2 \end{bmatrix}.$$

Let  $P$  be the matrix of joint probabilities,  $P = [p_{ij}] = P[X_i = 1 \cap Z_j = 1]$  and let  $d_i = P[X_i = 1]$ . Then  $P$  has the same block structure as  $C$ ,

$$P = \begin{bmatrix} P_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & P_5 & 0 & & 0 & 0 \\ 0 & 0 & P_{10} & & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & & P_{95} & 0 \\ 0 & 0 & 0 & \dots & 0 & P_{98} \end{bmatrix} \quad (8)$$

where

$$P_0 = \begin{bmatrix} d_0 & 0 & 0 \\ 0 & d_1 & 0 \\ 0 & 0 & d_2 \end{bmatrix}, \quad P_{98} = \begin{bmatrix} d_{98} & 0 \\ 0 & d_{99} \end{bmatrix},$$

and for  $k \in \mathcal{K}$ ,

$$P_k = \begin{bmatrix} \alpha_k^2 d_{k-2} & 0 & (1 - \alpha_k^2) d_{k-2} & 0 & 0 \\ 0 & \alpha_k^1 d_{k-1} & (1 - \alpha_k^1) d_{k-1} & 0 & 0 \\ 0 & 0 & d_k & 0 & 0 \\ 0 & 0 & (1 - \alpha_k^1) d_{k+1} & \alpha_k^1 d_{k+1} & 0 \\ 0 & 0 & (1 - \alpha_k^2) d_{k+2} & 0 & \alpha_k^2 d_{k+2} \end{bmatrix}.$$

The matrix of conditional probabilities  $G$  is got by dividing each element of  $P$  by the sum of the elements in the column that it occupies<sup>31</sup>, leading to

$$G = \begin{bmatrix} I_3 & 0 & 0 & \dots & 0 & 0 \\ 0 & G_5 & 0 & & 0 & 0 \\ 0 & 0 & G_{10} & & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & & G_{95} & 0 \\ 0 & 0 & 0 & \dots & 0 & I_2 \end{bmatrix} \quad (9)$$

where

$$G_k = \begin{bmatrix} 1 & 0 & (1 - \alpha_k^2) d_{k-2}/s_k & 0 & 0 \\ 0 & 1 & (1 - \alpha_k^1) d_{k-1}/s_k & 0 & 0 \\ 0 & 0 & d_k/s_k & 0 & 0 \\ 0 & 0 & (1 - \alpha_k^1) d_{k+1}/s_k & 1 & 0 \\ 0 & 0 & (1 - \alpha_k^2) d_{k+2}/s_k & 0 & 1 \end{bmatrix} \quad (10)$$

and  $s_k$  is the sum of the entries in column  $k$ , namely

$$s_k = \sum_{j=-2}^2 d_{k+j} - \alpha_k^1 (d_{k-1} + d_{k+1}) - \alpha_k^2 (d_{k-2} + d_{k+2}) \geq d_k.$$

<sup>31</sup>We must have  $d_i > 0$  for all  $i$  and  $\alpha_k^j > 0$  for all  $j$  and  $k$  to be able to take this step.

Recalling that regression of  $Y_h$  on  $n$  has coefficients  $\gamma = G'\beta$ , it is clear that in this model of heaping  $\gamma_i = \beta_i$  when  $i$  is not a multiple of 5 even though counts of household members at these ages are mis-recorded. At ages  $k$  which are multiples of 5,  $\gamma_k \neq \beta_k$  if there is age heaping, and

$$\begin{aligned} \gamma_k &= s_k^{-1} \left( (1 - \alpha_k^2) \beta_{k-2} + (1 - \alpha_k^1) d_{k-1} \beta_{k-1} + d_k \beta_k \right. \\ &\quad \left. + (1 - \alpha_k^1) d_{k+1} \beta_{k+1} + (1 - \alpha_k^2) \beta_{k+2} \right) \end{aligned}$$

which is a weighted average of the  $\beta$  coefficient at age  $k$  and the four adjacent  $\beta$  coefficients.

As heaping vanishes, i.e.,  $\alpha_k^1 \rightarrow 1$  and  $\alpha_k^2 \rightarrow 1$ ,  $\gamma_k \rightarrow \beta_k$  and as heaping becomes complete, i.e.,  $\alpha_k^1 \rightarrow 0$  and  $\alpha_k^2 \rightarrow 0$ ,  $\gamma_k \rightarrow \sum_{j=-2}^2 d_{k+j} \beta_{k+j} / \sum_{j=-2}^2 d_{k+j}$  and coefficients  $\gamma_i$ ,  $i \notin \mathcal{K}$  become undefined. The effect of this form of age heaping is to cause age profile variations around multiples of 5 to tend to appear smoother than they in fact are, although only the values on the profile at the multiple of 5 are affected.

**5.3. Corrected estimates.** Given an estimate of  $\gamma$  and knowledge of the  $\alpha_k^j$ 's it may be possible to retrieve an estimate of  $\beta$ . To do this  $G$  must be non-singular but it is easy to see that, for all  $k$ ,  $\det(G_k) = d_k/s_k \in [0, 1]$ . So retrieval will be possible whenever  $d_k > 0$  at all multiples of 5. Obviously  $d_i > 0$  is required if  $\beta_i$  is to be identifiable. So identifiability of  $\beta$  (which also requires  $\alpha_k^j > 0$  for all  $j, k$ ) ensures that we can obtain  $\beta$  as  $\beta = G'^{-1}\gamma$ .

The matrix  $G'^{-1}$  has the same block diagonal structure as  $G$  with each  $G_k$  replaced by  $G_k'^{-1}$  where

$$G_k'^{-1} = \frac{1}{d_k} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -(1 - \alpha_k^2) d_{k-2} & -(1 - \alpha_k^1) d_{k-1} & s_k & -(1 - \alpha_k^1) d_{k+1} & -(1 - \alpha_k^2) d_{k+2} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

It follows that  $\beta_i = \gamma_i$  except at ages  $k$  which are multiples of 5 where

$$\begin{aligned} \beta_k &= d_k^{-1} \left( -(1 - \alpha_k^2) d_{k-2} \gamma_{k-2} - (1 - \alpha_k^1) d_{k-1} \gamma_{k-1} + s_k \gamma_k \right. \\ &\quad \left. - (1 - \alpha_k^1) d_{k+1} \gamma_{k+1} - (1 - \alpha_k^2) d_{k+2} \gamma_{k+2} \right). \end{aligned} \quad (11)$$

In order to implement a correction based on (11) one would need information on the probabilities of mis-recording,  $\alpha_k^j$ , and on the marginal probabilities  $d_i = P[X_i = 1]$  the latter requiring access to error free data on ages. In the absence of information on the  $d_i$ 's one can exploit the information in the marginal distribution of recorded (heaped) ages. Let  $f_j = P[Z_j = 1]$ . The column sums of the matrices  $P_k$  are equal to  $[f_{k-2} \ f_{k-1} \ f_k \ f_{k+1} \ f_{k+2}]$ . It follows that  $s_k = f_k$  and the following.

$$\left. \begin{aligned} d_{k-2} &= f_{k-2} / \alpha_k^2 \\ d_{k-1} &= f_{k-1} / \alpha_k^1 \\ d_k &= f_k - (1 - \alpha_k^1) (f_{k-1} + f_{k+1}) / \alpha_k^1 - (1 - \alpha_k^2) (f_{k-2} + f_{k+2}) / \alpha_k^2 \\ d_{k+1} &= f_{k+1} / \alpha_k^1 \\ d_{k+2} &= f_{k+2} / \alpha_k^2 \end{aligned} \right\} \quad (12)$$

The  $f_k$ 's can be estimated, so with knowledge of the  $\alpha_k^j$ 's one could retrieve estimates of  $\beta$  from estimates of  $\gamma$ .

The parameters  $\alpha_k^j$  cannot be identified without information concerning the joint distribution of true and recorded ages, or further assumptions. However one can use the results above to investigate the sensitivity of results to different assumptions about age

heaping, comparing estimated age profiles as the  $\alpha_k^j$  parameters are varied over plausible ranges of values. In doing that while using the roughness penalty method outlined earlier, it is appropriate to apply the roughness penalty to the  $\beta$  coefficients. To do this one rewrites the model for household nutrient acquisitions as

$$Y_h = (Gn_M)' \beta^M + (Gn_F)' \beta^F + \varepsilon$$

using the transformed age counts<sup>32</sup>,  $Gn_M$  and  $Gn_F$ , in the estimation, applying the roughness penalty method as before.

With additional information about the age distribution one could attempt estimation of the  $\alpha_k^j$  parameters. For example suppose one took the view that the true age distribution is in some sense “smooth”. Then one estimation strategy would entail choosing as estimates of the  $\alpha_k^j$ 's the values that minimise the sum of the squared second differences of the  $d_i$ 's

$$\sum_{i=2}^{99} (d_i - 2d_{i-1} + d_{i-2})^2$$

employing the relationships (12) and the observed proportions at each recorded year of age as data<sup>33</sup> on the  $f_i$ 's. The estimation problem is numerically a rather simple one if we reparameterise in terms of  $\rho_k^j = 1/\alpha_k^j$  because then the objective function is a quadratic function<sup>34</sup> of the  $\rho_k^j$ 's. If this strategy were followed one would probably want to put some structure on the  $\alpha_k^j$ 's, for example removing, or at least limiting, their dependence upon  $k$ .

**5.4. The regression of  $Y$  on  $X$  and  $V$ .** Now consider the implications of the specific model of age heaping introduced above when one wishes to estimate the regression of  $Y$  on  $X$  and  $V$ . As noted earlier we have

$$E[Y|Z = z, V = v] = (z'G(v)'\beta) \exp(\delta'v)$$

where  $G(v) = [g_{ij}(v)]$  and  $g_{ij}(v) = P[X_i = 1|Z_j = 1, V = v]$ .

The matrix  $G(v)$  is obtained much as above, except that we must take care to condition on  $V$  at various points in the development. Suppose that the age heaping process is in fact independent of  $V$  in the sense that

$$P[Z_j = 1|X_i = 1, V = v] = P[Z_j = 1|X_i = 1]. \quad (13)$$

Then the  $\alpha_k^j$ 's do not depend upon  $v$  and the matrix  $C$  given in (7) is independent of  $v$ . However the matrix  $P$  given in (8) is not independent of  $v$  unless  $X$  and  $V$  are independent which will generally not be the case. The reason is that the elements of  $P$  now have to give joint probabilities for  $X$  and  $Z$  *conditional* on  $V$ , i.e.  $P$  becomes  $P(v) = [p_{ij}(v)]$  where

$$p_{ij}(v) = P[X_i = 1 \cap Z_j = 1|V = v] = P[Z_j = 1|X_i = 1]P[X_i = 1|V = v]$$

in which we have exploited (13). So in producing the matrix  $P$  we must multiply elements on row  $i$  by the conditional probability that  $X_i = 1$  given  $V = v$ . Let

$$d_i(v) = P[X_i = 1|V = v].$$

<sup>32</sup>Evaluated at the chosen values of  $\alpha_k^j$ 's.

<sup>33</sup>The joint distribution of the  $\hat{f}_i$ 's is multinomial which leads directly to a weighted criterion and more efficient estimation.

<sup>34</sup>The constraints  $\rho_k^j \geq 1$  will have to be imposed.

Then  $P(v)$  has the same structure as in (8) but now, in the definitions of the blocks of  $P$ , each  $d_i$  must be replaced by  $d_i(v)$ .

The column sums of the matrix  $P(v)$  now depend upon  $v$  but on dividing through we arrive at a matrix  $G(v)$  with the structure given in (9) and with blocks  $G_k$  as in (10) but with each  $d_i$  replaced by  $d_i(v)$  and with  $s_k$  replaced by  $s_k(v)$  where

$$s_k(v) = \sum_{j=-2}^2 d_{k+j}(v) - \alpha_k^1(d_{k-1}(v) + d_{k+1}(v)) - \alpha_k^2(d_{k-2}(v) + d_{k+2}(v)) \geq d_k(v).$$

Only the elements of  $G(v)$  in columns which are multiples of 5 years depend upon  $v$  and the matrix  $G(v)$  is, away from these columns, an identity matrix. It follows that when  $z$  does not have a unit element in a column  $k$  where  $k$  is a multiple of 5 years,

$$E[Y|Z = z, V = v] = (z'\beta) \exp(\delta'v).$$

This means that if individual data were available one could estimate  $\delta$  and all the elements of  $\beta$ ,  $\beta_i$ , apart from those for which  $i$  is a multiple of 5 by omitting data on those people with ages that are multiples of 5. And one could proceed in the same way with household aggregate data, omitting all households in which any member has a recorded age which is a multiple of 5, although in practice this might entail a considerable loss of data. Note that the same procedures would work if the age heaping process was not independent of  $V$ .

An alternative way to proceed that does not entail discarding data is to employ information on recorded ages, as in the previous section. One could estimate a model for  $f_i(v) = P[Z_i = 1|V = v]$  and then with knowledge of the  $\alpha_k^j$ 's produce estimates of the  $d_i(v)$ 's using (12). It would not be appropriate to obtain corrected estimates using (11) because the estimates of the  $\gamma$  coefficients will have been obtained from a mis-specified model in which the full dependence of the regression on  $v$  was not captured. Instead one should use the estimates of the  $d_i(v)$ 's to form up an estimate of the matrix  $G(v)$  and then estimate

$$Y_h = \left( \beta_0 + (\widehat{G}(v)n_M)' \beta^M + (\widehat{G}(v)n_F)' \beta^F \right) \exp(\delta'v) + \varepsilon$$

using data on all households.

#### ACKNOWLEDGMENTS

Financial support provided by ESRC grant R 008237386 is gratefully acknowledged.

## REFERENCES

- Bingham, S.A., Cassidy, A., Coles, T.J., Welch, A., Runswick, S.A., Black, A.E., Thurnham, D., Bates, C., Khaw, K.T., Key, T.J.A., and Day, N.E. (1995) Validation of Weighed Records and Other Methods of Dietary Assessment using the 24 h Urine Nitrogen Technique and Other Biological Markers. *British Journal of Nutrition*, **73**, 531-550.
- Black, A.E., Goldberg, G.R., Jebb, S.A., Livingstone, M.B.E., Cole, T.J., and Prentice, A.M. (1991) Critical Evaluation of Energy Intake Data Using Fundamental Principles of Energy Physiology: 2. Evaluating the Results of Published Surveys. *European Journal of Clinical Nutrition*, **45**, 583-599.
- Chesher, A.D. (1996) Individual Demands from Household Aggregates: Time and Age Variation in the Composition of Diet. *Journal of Applied Econometrics*, **13**, 505-524.
- Chesher, A.D. (1997) Diet Revealed? Semiparametric Estimation of Nutrient Intake - Age Relationships (with discussion). *Journal of the Royal Statistical Society, A*, **160**, 389-428.
- Deaton, A.S., and Paxson, C. (2000) Growth and savings among individuals and households, *Review of Economics and Statistics*, **82**, 212-225.
- Department of Health (1989). The Diets of British Schoolchildren. *Report on Health and Social Subjects No 36*. London: HMSO.
- Engle, R.F., Granger, C.W.J., Rice, J., and Weiss, A. (1986) Semiparametric Estimates of the Relation between Weather and Electricity Sales. *Journal of the American Statistical Association*, **81**, 310-320.
- Green, P.J., and Silverman, B.W. (1994) *Nonparametric Regression and Generalised Linear Models: A Roughness Penalty Approach*, London: Chapman and Hall.
- Gregory, J.R., Collins, D.L., Davies, P.S.W., Hughes, J.M., and Clarke, P.C. (1994) *National Diet and Nutrition Survey: Children Aged 1½ to 4½ Years*. London: HMSO.
- Gregory, J.R., Foster, K., Tyler, H., and Wiseman, M. (1990) *The Dietary and Nutritional Survey of British Adults*. London: HMSO.
- Livingstone, M.B.E., Prentice, A.M., Strain, J.J., Coward, W.A., Black, A.E., Barker, M.E., McKenna, P.G., and Whitehead, R.G. (1990) Accuracy of Weighed Dietary Records in Studies of Diet and Health. *British Medical Journal*, **300**, 708-712.
- Miquel, R., and Laisney, F. (2000) Consumption and Nutrition: Age - Intake Profiles for Czechoslovakia 1989 - 92. *Economics of Transition*, forthcoming.
- Mills, A., and Tyler, H. (1992). *Food and Nutrient Intakes of British Infants Aged 6 - 12 Months*. London: HMSO.
- Parkin, D., Rice, N., and Sutton, M. (1999) Non- and Semi-parametric Estimation of Age and Time Heterogeneity in Repeated Cross-sections: An Application to Self-reported Morbidity and General Practitioner Utilization. *Health Economics*, **8**, 429-440.
- SAS Institute Inc., (1993) *SAS/ETS Users's Guide, Version 6, 2nd Edition*, Cary, NC: SAS Institute Inc.
- Vasedkis, V.G.S., and Trichopolou, A. (2000) Nonparametric Estimation of Individual Food Availability along with Bootstrap Confidence Intervals in Household Budget Surveys. *Statistics and Probability Letters*, **46**, 337-345.
- Wahba, G. (1990) *Spline Models for Observational Data*, Philadelphia: SIAM.
- World Health Organization. (1990). *Diet, Nutrition and the Prevention of Chronic Diseases, Report of a WHO Study Group*. WHO Technical Report Series No. 797, Switzerland: WHO.