Digit Preference in African Survey Data and Their Impact on Parametric Estimates

By
Asmerom Kidane
Department of Economics
University of Dar es Salaam

For presentation at the African Econometric Society Conference
July 11-13 2009
Abuja, Nigeria

Digit preference in African Survey data and their impact on parametric estimates

By
Asmerom Kidane
University of Dar es Salaam

Abstract

Most microeconomic and demographic variables in African countries are collected from sample surveys. Some are comprehensive and cover the whole country while others are region or area specific. In many large scale surveys an appropriate scientific method of sampling (usually multistage stratified cluster sampling) is adopted. However the responses may not be accurate. There are many reasons for wrong reporting such as memory lapse or deliberate attempt to underestimate (such as income) or overestimate (such as expenditure). The most common source of error is the tendency to provide numerical responses that end with certain digits, especially those that end with integer “zero”, followed by the integer “five”. A typical case is “age heaping” where individuals give their ages with umbers ending with these digits. The types of digit preference are not only limited to age reporting. When farmers are asked about acreage planted, amount harvested, number of cattle owned, output consumed and sold , distance from home to the nearest market etc., they appear to give numerical values that end with “zero” or “five”.

In this paper seven variables-four discrete and three continuous- were selected from the 2002-2003 survey of Tanzania. A moderate to extreme form of digit preference was observed. The magnitude of digit preference were also found to be a function of the educational level of respondents. Compared to discrete, continuous variable were more prone to digit preference. A linear and a Cobb Douglas type production were fitted using unsmoothed (variables with digit preference) and smoothed (variables adjusted for digit preference). The results showed better and predictable estimates under smoothed variables. Compared to linear model, a much better result was obtained under the Cobb Douglas production function.

Conducting surveys with few questions along with physical measurement by enumerators on fewer households will go a long way towards reducing the prevalence of digit preference

Table of contents

Section	Title	Page no.
1	Introduction	4
2	Expected and observed age distribution	5
3	Magnitude of age heaping	8
4	Digit preference in other variables	11
4.1	Digit preference for discrete variables	12
4.1.1	Number of trees owned	12
4.1.2	Number of cattle owned	12
4.1.3	Number of goats owned	13
4.2	Digit preference for continuous variables	14
4.2.1	Time taken to fetch water	14
4.2.2	Amount harvested	15
4.2.3	Acreage planted	16
4.3	Selected numerical indicators of digit preference	17
5	Estimates of an agricultural production function	18
6	Conclusion	21
7	Reference	23

Digit preference in African Survey data and their impact on parametric estimates

By
Asmerom Kidane
University of Dar es Salaam

1. Introduction

Most microeconomic and demographic variables in African countries are collected from sample surveys. Some are comprehensive and cover the whole country while others are region or area specific. In many large scale surveys an appropriate scientific method of sampling (usually multistage stratified cluster sampling) is adopted. All requisite procedures such as the preparation of a questionnaire, pilot survey, training of enumerators and supervisors are adapted. This will naturally reduce sampling errors and provide one with representative respondents.

However the responses may not be accurate. There are many reasons for wrong reporting such as memory lapse or deliberate attempt to underestimate (such as income) or overestimate (such as expenditure). The most common source of error is the tendency to provide numerical responses that end with certain digits, especially those that end with integer “zero”, followed by the integer “five”. A typical case is “age heaping” where individuals give their ages with umbers ending with these digits. An individual is more likely to state that his age is 30; this is usually at the expense of the adjacent ages such as 28, 29 or 31, 32. Other variables such as years of schooling and experience are also given in number of years. These types of digit preference are not only limited to age reporting. When farmers are asked about acreage planted, amount harvested, number of cattle owned, output consumed and sold, distance from home to the nearest market etc., they appear to give numerical values that end with “zero” or “five”.

Several techniques have been developed (mostly by demographers and epidemiologists) to measure the magnitude of errors emanating from digit preference as well as techniques of correcting or adjusting the information. Some of the techniques are robust while others are not.

The aim of this study will be to investigate the pattern of digit preference using the results from the 2002-2003 comprehensive agricultural survey conducted in Tanzania. This large scale survey covers small holder farms in rural Tanzania. The sample size is

more than 51,000 households. A closer look at the reported data reveals substantial tendency for digit preference especially those ending with zero and five. An attempt will thus be made to measure the magnitude of age heaping and digit preference of other variables as reported by the heads of households.

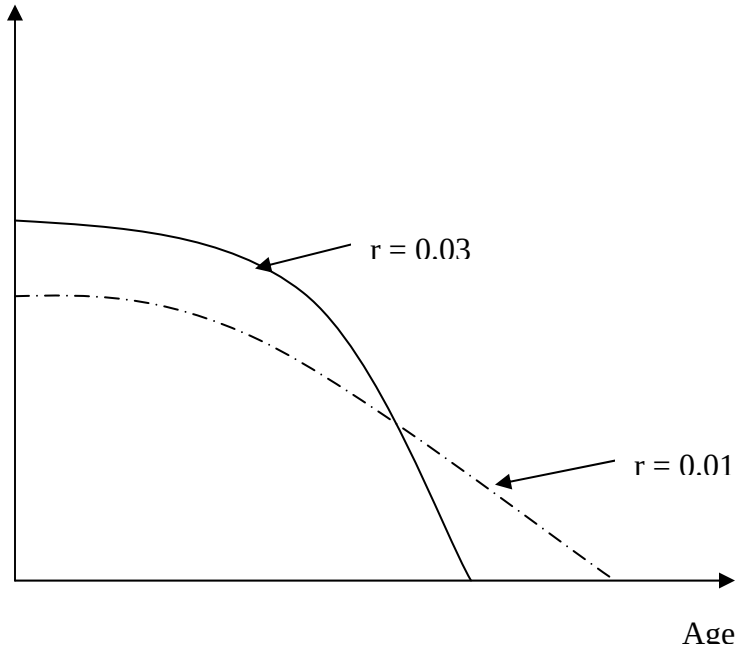
This paper has six parts. Part two will compare reported age distribution with the one based on Model Life Tables (ideal or error free age distributions) and compare the magnitude of the difference. Part three will measure the magnitude of age heaping by applying the most common technique-the Whipple Index. The magnitude of age heaping will also be estimated by variation in the level of education. Part four will consider the magnitude of digit preference in other variables including the number of cattle owned, acreage planted, amount harvested and the time it takes to fetch for water. Part five will estimate and discuss a standard Production Function in the presence of digit presence and after undertaking a smoothing process.. The aim will be to estimate and compare the coefficients and the marginals. Some concluding remarks will be forwarded in Part 6.

2. Expected and observed age distribution

In the absence of age misreporting or age heaping, in the absence of international emigration and immigration, in the absence of war, major epidemic or famine induced mortality, the age distribution of a country's population is a non increasing function of that age. This is true irrespective of the country's level of development. However the decrease in the age distribution among developing countries is higher than in the developed countries. Two hypothetical age distributions are given in Figure 1. In this figure the fast declining graph refers to a fast growing population (r>=0.03) while the slow declining graph refers to a slow growing population (r<=0.01). The above figure is theoretical. For the country under study (Tanzania) one needs to compare the observed age distribution with the expected and discuss the magnitude of the difference. The expected age distribution is

Figure 1

Hypothetical age distributions

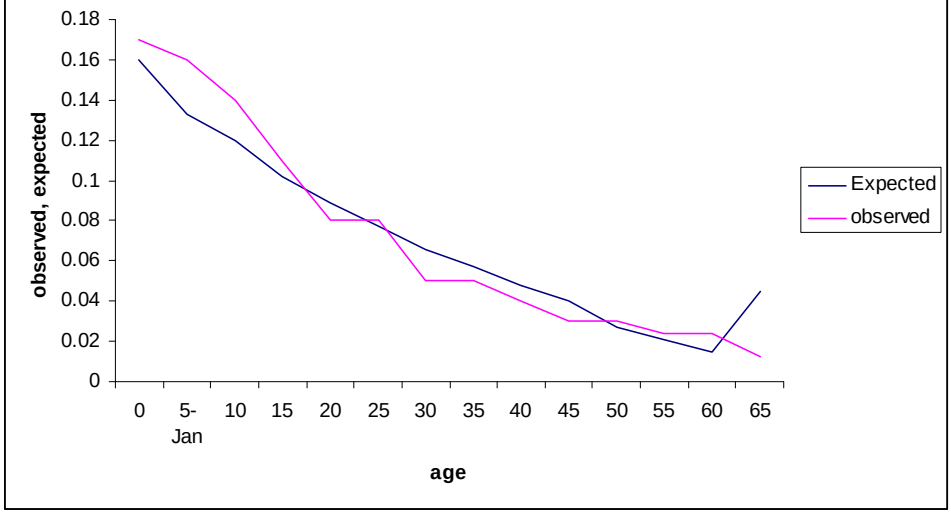


estimated indirectly from limited information and from Model Life Tables. There are different Model Life Tables; the most common one is the Coale Demeney Model Life Tables (Coale, Demeney**,** 1976)**.** From the latest Statistical Bulletin of Tanzania as well as the US Bureau of Census of the United States Tanzania's population growth rate is given as being about 2.3 percent. On the other hand from the Demographic and Health Surveys (DHS) of Tanzania (2005) one can get an estimate of birth rate which is 42.4 per thousand. With this information at hand one can identify the appropriate model that is consistent or that fits the age distribution of present day Tanzania. Coale and Demeney identify four different models commonly known as East, West, North and South Models. These models vary not by geographical areas but by historical patterns of fertility and mortality. For a developing country like Tanzania, the West Model is appropriate. Thus with the stated growth and birth rates of Tanzania the right Model is Level 14 West. The corresponding expected age distribution along with the observed is given in table 1 and figure 2. The observed estimate is based on the 2002-2003 agricultural survey.

Table 1 Expected and observed age distribution of Tanzanian male population 2004-2008			
Age interval	Expected*	Observed**	Difference(O-E)
0-4	0.160	0.170	0.010
5-9	0.133	0.160	0.027
10-14	0.120	0.140	0.020
15-19	0.102	0.110	0.008
20-24	0.089	0.080	-0.009
25-29	0.077	0.080	0.003
30-34	0.066	0.050	-0.016
35-39	0.057	0.050	-0.007
40-44	0.048	0.040	-0.008
45-49	0.040	0.030	-0.010
50-54	0.027	0.030	0.003
55-59	0.021	0.024	0.001
60-64	0.015	0.024	0.009
65-	0.045	0.012	-0.033

*derived from West Model Life Table (GRR=2.5), r=2.39
**based on agricultural survey of 2002-2003

Figure 2
Observed and expected age distribution-Tanzania 2005



There are several reasons for presenting age distribution in terms of five year interval. One of the reasons is to "eliminate" the pattern of age heaping or digit preference. Given that the expected age distribution is the right one, the result above shows that there is an overestimation of younger members at the expense those who are middle aged.. This will

inflate the dependency ratio along with other macroeconomic consequences of the high dependency.

The best way to measure the magnitude of age heaping is to present the age distribution in terms of single years. One also needs a numerical measure or an index of digit preference. The most common index of age heaping is the Whipple Index which is given as follows;

$$WI = \frac{\sum (P25 + P30 + P35 + P40 + P45 + P50 + P55 + P60)}{\frac{1}{5}\sum (P23 + P24 + ... + P62)} * 100$$

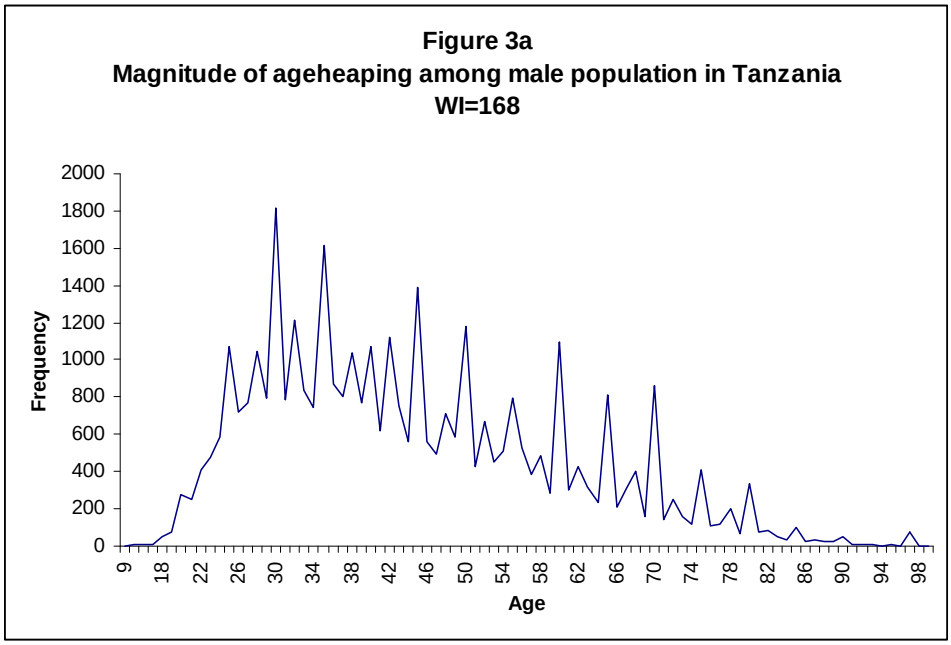
In the above equation the P23 etc. refer to population aged 23 years. One should note that the above equation assumes the existence of digit preference with ages ending with 0 and 5. The values of the above index ranges between 0 and 500. When WI=100 there is no problem of digit preference while WI=500 implies an extreme form of age heaping where all observations end with the digit zero or five. In practice a value of W<= 110 implies that age heaping is not significant while a value of W>=150 implies high prevalence.

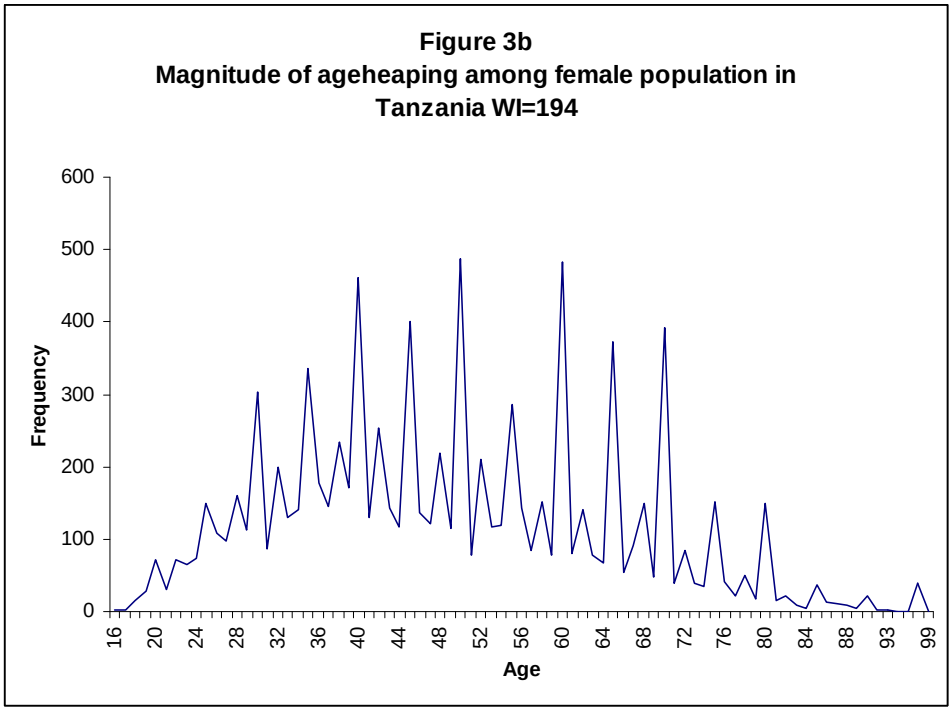
The above index is expected to measure digit preference for population aged 23 to 62 years. It can easily be modified to include wider or narrower age range. The index may also be modified so as to make it applicable to other variables where digit preference is suspected.

3. Magnitude of age heaping

The following paragraphs present the pattern and magnitude of age heaping among the surveyed households in Tanzania. The results are based on the 2002-2003 agricultural survey conducted in Tanzania. The reported ages are given by the respondent who is usually a male. Female headed households were very few and were omitted from analysis. Besides his own, age the head of household also provided the age of his wife and other members of the household. Chances are the age preference among wives is expected to be more pronounced than that of husbands.

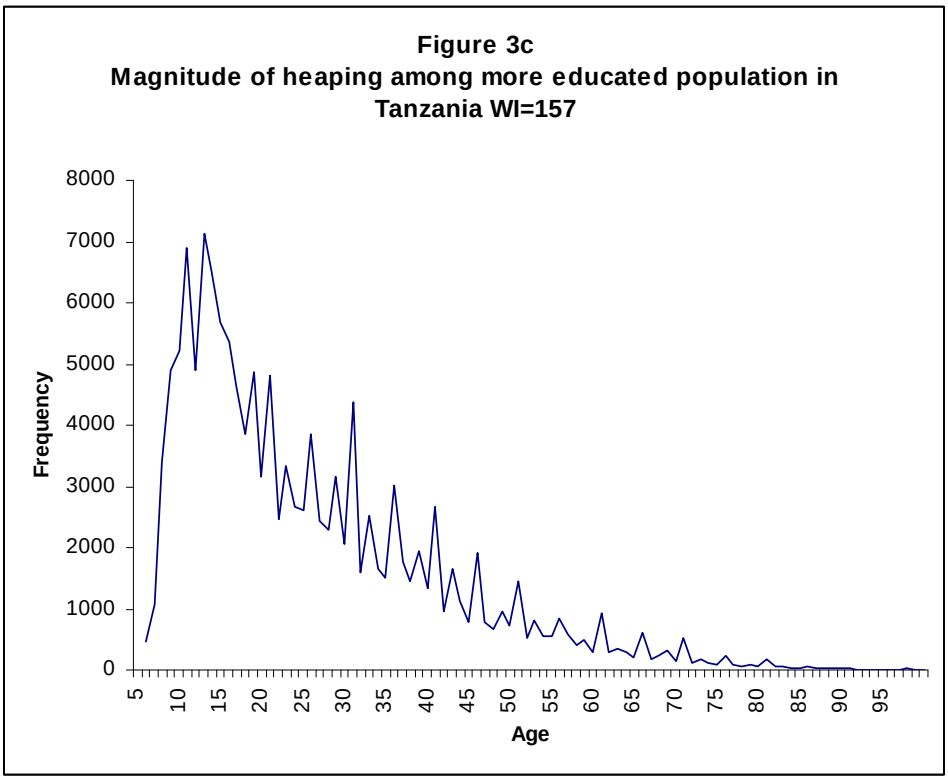
Figures 3a and 3b show the magnitude of age heaping of males and females along with the Whipple Index.

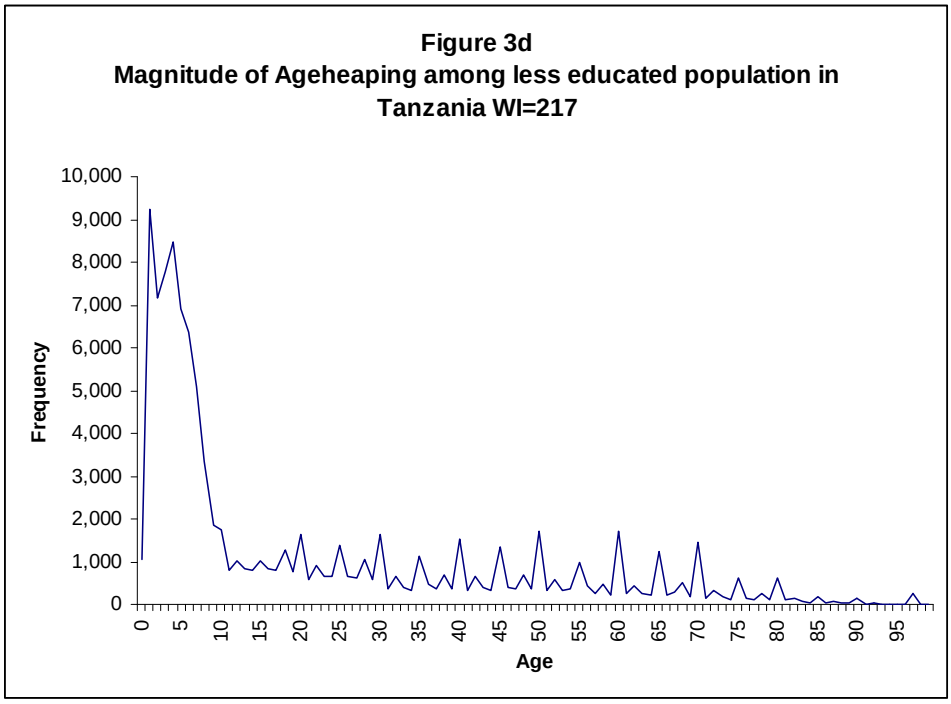




The above figures show extreme pattern of digit preference especially those that end with zero. As expected the distortion is more pronounced among females than males as the Whipple Index for males and females are 168 and 194 respectively**.**

One would expect that the educational level would have an effect on the magnitude of digit preference or age heaping; respondents with higher level of education are expected to provide a more correct response. In order to verify this expectation, respondents were classified into two; those who do not know how to read and write and the others. The results are given in Figure 3c and 3d .





It appears that the effect of education on the magnitude of age heaping as being substantial; there is a difference of 60 points between the two indices. If education of the respondent is to have significant impact in reducing age distortion, the educational achievement ought to be high (probably 12 or more years of education).

4. Digit preference in other variables

We have already noted that digit preference is also common in other variables. In this section we consider the magnitude of digit preference in selected variables or responses. The choice of some of these variables is predicated by the subsequent objective, that is, the aim of estimating Production Functions. Three discrete and three continuous variables are considered. It should be highlighted that there are many discrete and continuous survey variables with responses that tilt towards digit preference. The variables selected for analysis are

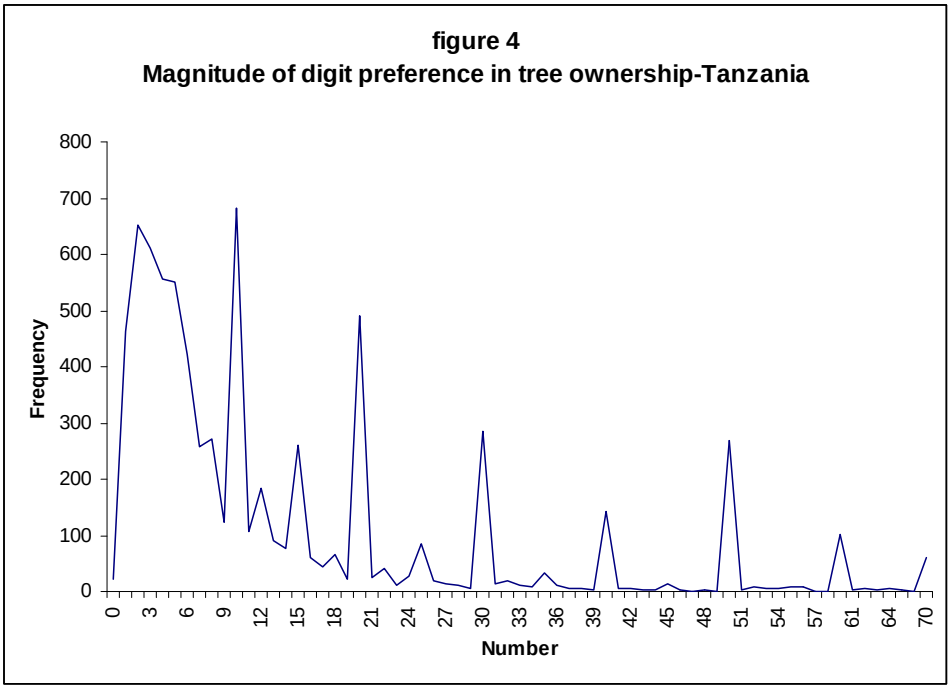
- Number of trees owned
- Number of cattle owned
- Number of goats owned
- Time to fetch water
- Quantity harvested

- Area planted

4.1 Digit preference for discrete variables

4.1.1 Number of trees owned

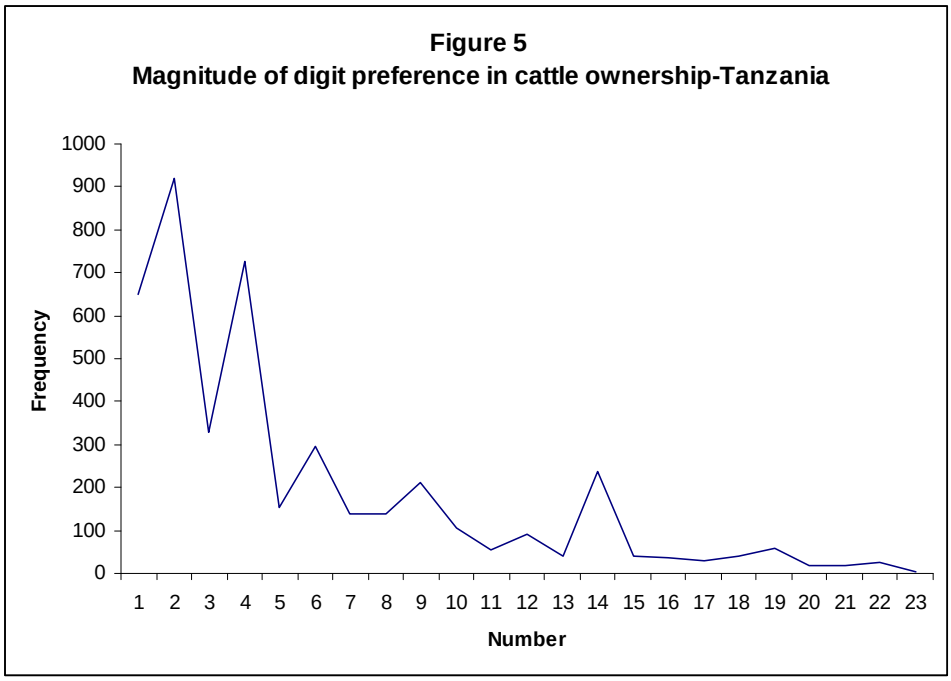
The amount of tree owned ranges from zero to 9000. Fifty percent of the respondents own less than 10 trees. Chances are that people with less than 10 trees report the exact number of trees they own. Similar conclusion may apply to cattle and goat ownership. With regards to tree ownership we considered those respondents with number of trees between 10 200 (these constitute 75% of respondents). The results are given in Figure 4 where one observes a clear pattern of digit preference for numbers ending with zero. Preference for numbers ending with five is also prevalent



4.1.2 Number of cattle owned

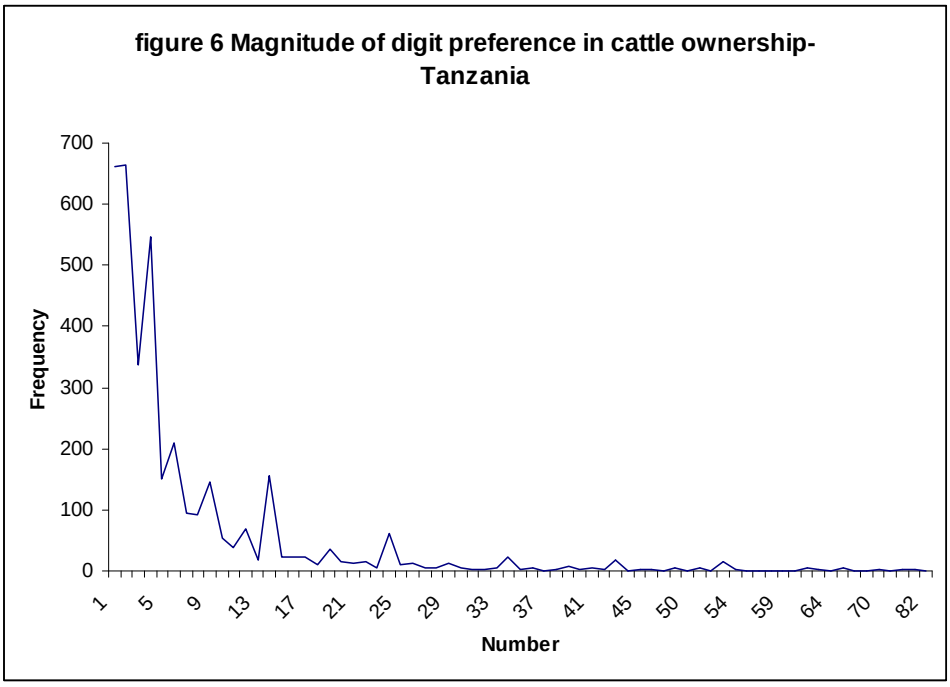
About 45 percent of respondents stated that they owned less than 10 cattle. Again these range of values are assumed to be relatively correct and are not subject to digit preference. The prevalence of digit preference applies to about 50 % of respondents who

own more than 10 cattle. The pattern of digit preference is given in figure 5. Here the preference for a number ending with five appears to be common.



4.1.3 Number of goats owned

The number of goats owned range between and zero and 700. Fifty percent of respondents own less than five goats. Only 80 percent of the respondents are included in the estimation process. The results are given in Figure 6. Like in case of cattle ownership the preference for digits ending with five appear to be more prevalent.

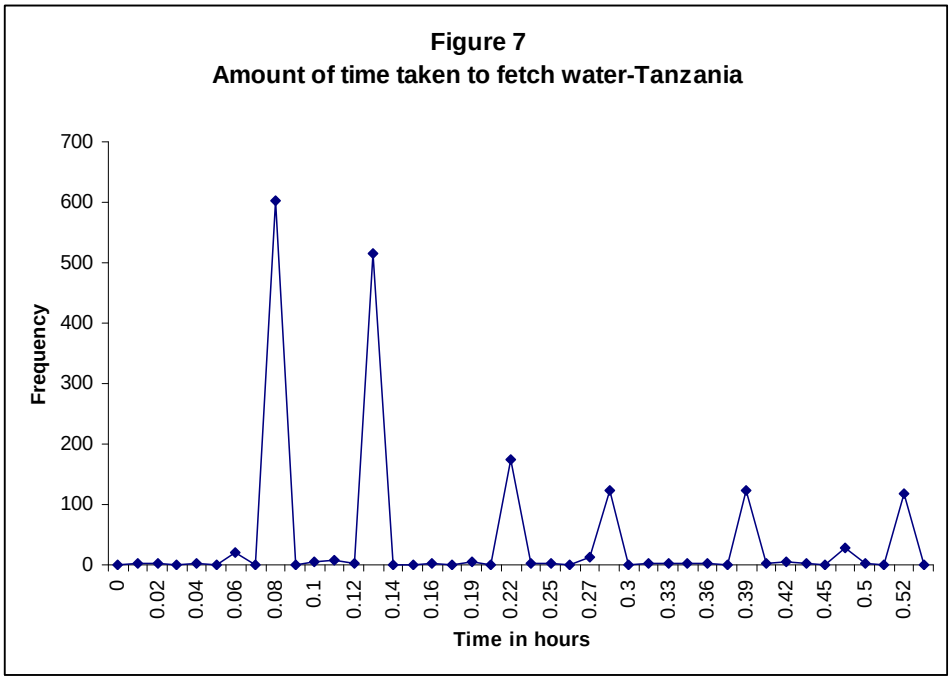


4.2 Digit preference for continuous variables

Even though there is digit preference in continuous variables it will be difficult to include all possible values. For example the total amount harvested ranges between 60 Kg. and 10500 kg. and there exist more than two thousand digits ending with zero and five. We will consider the fifty percentile of the total for graphical presentation. It should also be noted that it may not be fair to ask a peasant farmer the exact length of time or the exact area planted. His knowledge of area or time is a function of his level of education.

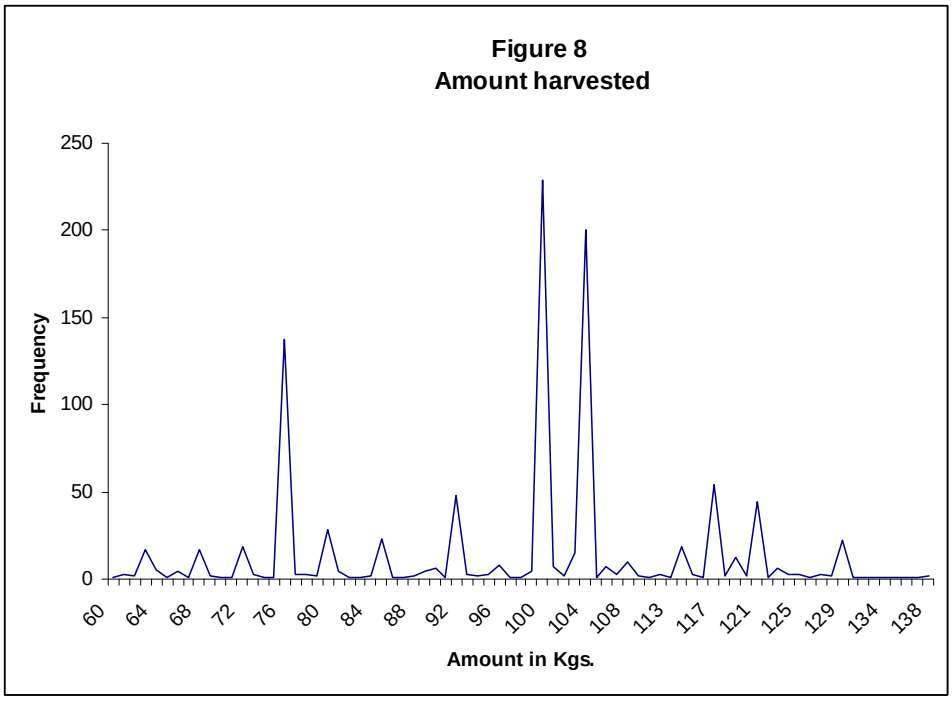
4.2.1 Time taken to fetch water

In rural Tanzania it is mostly the females fetch and carry water. There are two points that need to be emphasized. First the respondents are mostly male headed households and it is unlikely for them to give exact response to an activity that they are not involved. Secondly one should not expect the exact value of a continuous variable such as time to fetch water. The time range of response was between zero and fifty hours! There are too many values between the two extremes . Figure 7 shows a clear case of digit preference especially those ending with zero.



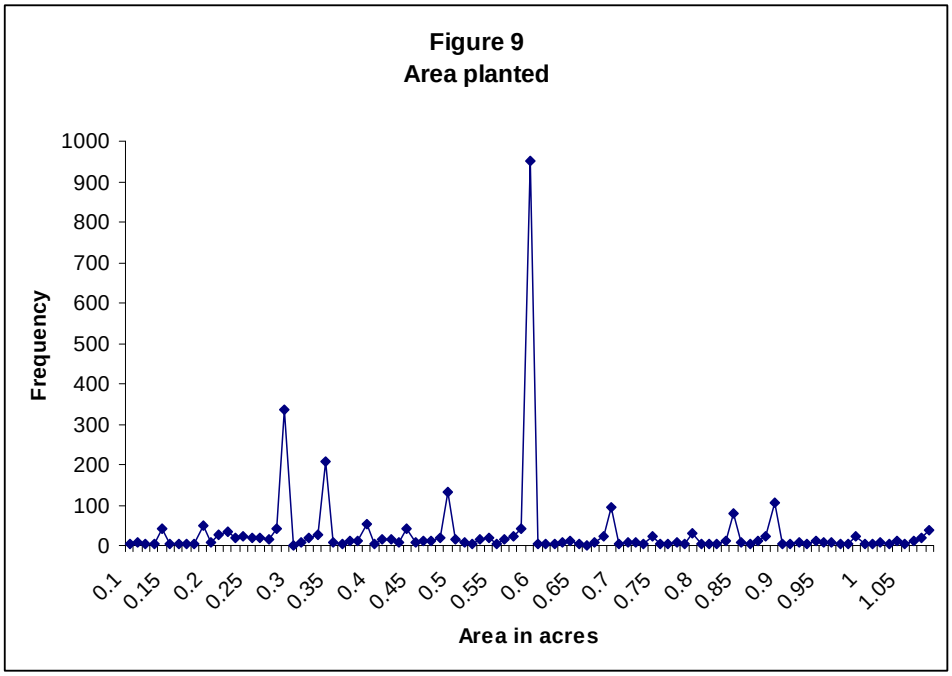
4.2.2 Amount harvested

We have already noted that the range of value for harvest is between 60 and 10500 Kgs. Again there is an extreme pattern of digit preference especially around the first and second quartile.



4.2.3 Area planted

In many traditional agrarian societies like Tanzania the area of land owned by peasants is known and limited to few numbers such half an acre, one acre, two and more acres. However the whole area may not be planted; the area owned may include farmers' houses, barns for animals, area allotted for gardens and for perennial crops. Area planted for annual crops may not be affected by digit preference only; the response may also be an overestimate. The reported values (Figure 9) for area planted show a glaring pattern of digit preference around the median and around lower values.



4.3 Selected numerical indicators of digit preference

The 2002-2003 Tanzanian national sample census of agriculture is based on more than 51000 households – one of the largest to be conducted in sub Sahara Africa. The survey was well budgeted and contained hundreds of questions. Out of these only six variables are included in this study. The results are given in graphical presentation. It would be too detailed and cumbersome to prepare a tabular presentation. Instead we present a summary of the magnitude of digit preference for selected numbers. This is given in Table 2 It appears that digit preference for age is relatively milder when compared to various types of wealth ownership such as cattle, trees and acreage.

Table 2 Preference for selected digits				
Variable/range	**Digit ending with zero and adjacent value**		**Digit ending with five and adjacent value**	
Age (0-99)*	**Digit**	**Frequency**	**Digit**	**Frequency**
	19	106	24	161
	20	**352**	**25**	**1216**
	21	281	26	826
	.	**.**	**.**	
	29	908	34	893
	30	**2118**	**35**	**1947**
	31	870	36	1045
Tree ownership (0-99)*	19	22	14	77
	20	**490**	**15**	**259**
	21	24	16	60
Cattle ownership (0-639)*	19	40	14	140
	20	**236**	**15**	**213**
	21	41	16	107
Goat ownership (1-700)*	19	19	14	77
	20	**156**	**15**	**259**
	21	22	16	60
Time to fetch water (0-10.5 hours)*	0.26-0.29	7		
	0.30	**1855**		
	0.31-0.39	28		
Acreage (0-50 acres)*	0.91-0.99	914		
	1.0	**4233**		
	1.01-1.09	656		
Harvest (0-10500 kgs)*	91-99	121		
	100	**2210**		
	102-109	107		

*Minimum and maximum values

5. Estimates of an agricultural production function

In an attempt to compare parameter estimates based on unsmoothed (unadjusted) and smoothed (adjusted for digit preference) variables, two types of production functions- linear and Cobb Douglas- were estimated. We assume that crop output is a function of area planted or acreage, number of cattle (cattle are major inputs in small scale peasant agriculture) as well as the age of household head (proxy for productivity of labour). There are different methods of smoothing data with digit preference (Chatfield 2001**,** Montgomery and Johonson1990**).** In this exercise we have applied a nine point moving average. It should be noted that the smoothing or filtering methods have common

application in time series with rare application in cross section. It should also be highlighted that taking logarithms is part of smoothing process even though it does not completely reduce the magnitude of digit preference.

The results show significant difference between production functions based on unsmoothed and smoothed data. As expected the Cobb Douglas production estimates gave much better results compared to linear models. With regards to linear estimate (Table 4a) smoothed variables gave better estimates than unsmoothed variables. Cattle input was significant under smoothed and not so under unsmoothed variables.

A much better result was obtained when the Cob Douglas estimates were made for unsmoothed and smoothed variables (Table 4b). Results based on smoothed variables showed a much improved result All explanatory variables including the age variables were found to be significant with the desired sign. In all cases the adjusted coefficient of determination is quite low. This is to be expected from survey data.

It should be noted that the aim of this exercise is to compare results between variables with digit preference and the ones smoothed for the same. In other words we are trying to compare whether parameter estimates based on smoothed data give different results. In this exercise the results were not only different, but smoothed variables provided better and predictable results.

Table 4a

Linear production function under unsmoothed and smoothed variables

Dependent variable-Harvest in Kgs.

Explanatory variables	**Linear model unsmoothed**		**Linear model-smoothed**	
	coefficient	standard error	coefficient	standard error
Acreage	221.7*	2.54	249.20*	2.76
Cattle input	0.040	0.23	0.51*	0.25
Age**	0.299	0.53	-0.051	0.21
Age squared	-0.007	-0.93	0.00	0.00
Constant	109.80*	7.70*	77.83*	5.72
$\bar{R}^2$	**0.22**		**0.21**	
F(4,26513)	**1911.5**		**2036.9**	
Prob>F	**0.0000**		**0.0000**	
N	**26518**		**30958**	

*significant results

**In peasant agriculture, the household head constitutes the main labour input. Age is expected to be a major determinant of labour productivity.

Table 4b **Cob Douglas (CD) production function under unsmoothed and smoothed variables** **Dependent variable-Harvest in Kgs.**				
Explanatory variables (log)	**CD model unsmoothed**		**CD Linear model-smoothed**	
	coefficient	standard error	coefficient	standard error
Acreage	270.38*	5.30	0.769*	0.020
Cattle input	20.67*	3.91	0.155*	0.015
Age**	Dropped	-	-0.067*	0.02
Age squared	-1.02	1.70	0.010*	0.005
Constant	432.75*	10.55	5.32*	0.07
$\bar{R}^2$	**0.17**		**0.052**	
F(4,26513)	**1743.33**		**409.00**	
Prob>F	**0.0000**		**0.0000**	
n	**26295**		**29577**	

*significant results
**In peasant agriculture, the household head constitutes the main labour input. Age is expected to be a major determinant of labour productivity.

6. Conclusion

Many African and other developing countries depend on survey data to gather vital micro and macroeconomic as well as demographic variables. Some of the surveys including the one considered in this paper involve a high cost both in finance and time. The questions included in such surveys are too many. The 2002-2003 Tanzanian agricultural survey includes more than 300 questions and more than 1500 answers! Some of the questionnaires take more than three hours to fill. Under this condition it is not unexpected to report results that are not accurate. Because of a “respondent fatigue” it is not unusual that questions towards the end of the questionnaire to be not only wrong but misleading. Also one is not sure whether some of the responses are analyzed or even whether it is worth analyzing them.

The variables and the results discussed in this paper show digit preferences of various magnitude. Digit preference among continuous variables is more prevalent and more pronounced compared with the discrete variables. When two types of production function were estimated using unsmoothed and smoothed variables one was able to get different and better results with the latter. The smoothing of variables for model estimation appears to be the preferred option. If the variables are to be summarized in a form of a table then grouping the data will partially solve the problem of digit preference.

It may be argued that there are various techniques of smoothing and the outcome may vary with the type of smoothing. Choice of smoothing techniques is unlikely to have significant effect. There is a difference on the purpose of smoothing depending on whether a data set is time series or survey based. In time series data smoothing is applied to eliminate seasonal or irregular component to the original data set which are assumed to be correct. On the other hand when we apply smoothing to survey data to eliminate digit preference we are trying to correct data with wrong values.

A questionnaire with much less questions and with scientific method of sampling will go a long way towards obtaining more reliable and error free observations. Also physical measurement of variables (such as acreage, distance, time etc) by enumerators on a fraction of the surveyed households may help one to develop a method of correcting digit preference; this may serve as an alternate to smoothing.

7. References

Africa Commission (2004) Strengthening the quality and use of data in Africa. **(monograph)**

Chatfield, C. (2001) **Time Series Forecasting** . London, Chapman and Hill

Coale, A.J. and P. Demeney (1976) **Model Life Tables and Stable Populations** New York, Academic Press

DHS (2004-2005) **Tanzania: Demographic and Health Survey**

Kidane, A. (2000) African Macroeconomic and Price Data: Quality, Reliability and Internal Consistency **(monograph)**

Kidane, A.(1996) The Quality Reliability and Internal Consistency of African Macroeconomic Data: Some Methodological Issues **African Economic Research Consortium**, Nairobi

Montgomery, D. C., L.A. Johnson and J.S. Gardiner, (1990) **Forecasting and Time Series Analysis.** 2nd. ed. New YorkMcGraw Hill

Marrakech, 2004: Better Data for Better Results: An Action Plan for Improving Development Statistics, Marrakech Feb 2004 **(Proceedings**)

Tanzania NBS (2005) Tanzania Agriculture Sample Census: Main findings **(various issues)**