INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)
ISSN 2278 – 7631 (Print)

Data Mining For Social Research: A Study of Nutritious Food Consumption of Indian Households

Mayank Jain
Madras School of Economics, Chennai, Tamil Nadu, India
Pradip Kumar Bala
Associate Professor, Indian Institute of Management, Ranchi, Jharkhand, India

Abstract:
There are various approaches to measure the differences in household income. Income measured from the monetary earnings point of view is called the nominal income. Income measured from the consumption point of view i.e. the basket of consumption goods a household buys is called the real income. In this paper we analyze the income based on the consumption of Highly Nutritious Food. Using tools of data mining (C&R tree and C5.0) we predict the trends in expenditure of households of India in the consumption of Highly Nutritious Food (Pulses, Cereal Products, Meat, egg, milk and milk products) which are not available through the Public Distribution System (PDS). We study the differences in expenditure on Highly Nutritious Food (HNF) across caste, religion, states and demonstrate that multiple factors influence the expenditure on HNF. We also categorize the per capita consumption of HNF as Low, Average and High. The outcome shows disparity in consumption of Highly Nutritious Food and argues that it can be a way to look at income disparity.

Key words*: Highly Nutritious Food (HNF), Per Capita Consumption of Nutritious Food (PCCNF), Households, Caste, Religion, Data Mining, C&R Tree, C5.0*

1. Introduction

Among households of India, there have been inequalities in income since ancient times. This inequality persists in the nominal income, consumption, social status, welfare etc. In this paper we study one such inequality based on the inequality of consumption of Highly Nutritious Food (Pulses, Cereal Products, Meat, egg, milk and milk products) which are not available through the Public Distribution System (PDS). With the rising food prices, food expenditures are increasingly dominating the household budget. The households are forced to reduce their consumption basket and depend solely on the basic food for their nutritional requirement. This issue is more severe and pronounced when we study the differences in expenditure on HNF among the Indian households.

This paper proposes to review the trends in expenditure on HNF's using various input and outcome measures. This is an alternative approach to study the differences in income among households. Households with low incomes would tend to consume less of HNF and depend more on food available through PDS. This paper also identifies the differences in income based on consumption of HNF among households with different sources of income (like agricultural, non agricultural labour, artisan, service, business etc). It tries to investigate whether holding a ration card creates any incentive to consume more of HNF. There is also a difference in consumption of HNF among different regions of India (North, South, East, West, Central and North East) as discovered from the data. There are various other interesting findings which will be discussed subsequently.

Findings of the research incorporated in this article are based on the household data from India Human Development Survey, 2004-05 (IHDS). The IHDS is a nationally representative survey of 41,554 households organised by researchers from the University of Maryland and the National Council of Applied Economic Research. It is a multi-topic multi-purpose survey containing information about a variety of dimensions of social and economic well being of the house-holds. These data are in public domain and at an all-India level, poverty, education, household structure and employment levels recorded in this survey are comparable to those recorded by Census and the National Sample Survey albeit with some exceptions associated with the survey design (Desai et al 2010)

2. Literature Review

The problem of differences in income has been addressed by various authors, but very less literature is available about the differences in consumption of HNF. Most of the study in this area is based on earnings point of view and on per capita calorie consumption from the expenditure point of view. Few of them are "Income inequality in village India: The role of caste" by Swaminathan and Rawal (2011), which examines the role of caste in understanding inequality in incomes in rural India. In the paper "Regional Heterogeneity in Food Consumption and Nutrition Intake in India" Srivastava et al, the authors raise the issue of difference in consumption of cereal crops across Indian states. In their article "Food and Nutrition in India: Facts and Interpretation" Deaton and Dreze (2009) analyze per capita consumption from the calories perspective.

3. Understanding The Problem

The Indian social system is characterized by various social groups or Caste. The New Shorter Oxford Dictionary defines caste as "a Hindu hereditary class of socially equal persons, united in religion and usually following similar occupations, distinguished from other caste in the hierarchy by its relative degree of purity or pollution" [Ed Lesley Brown, Clarendon Press, Oxford, 1993]. In the data the caste is categorized as Brahmin, Scheduled Caste (SC), Scheduled Tribe (ST), Other Backward Class (OBC), others.

Caste system dominates a larger part of decision making in the Indian household scenario. A person of the Brahmin caste enjoys the supreme position and is generally well-off. The others are the people who are not Brahmins but club together with Brahmins to form the General Category. The Scheduled Caste are people who were historically disadvantaged and were considered untouchables. Scheduled Tribes are the ethnic tribal groups. Other Backward Class are socially and educationally backward communities.

India is a secular country. People following various religions, Hindu, Muslim, Sikh, Christians, Jains, Buddhists, Tribals etc all live in harmony. The vast landscape of India is divided into 28 states and seven union territories. No socio-economic study on India can be complete without taking in account the influence of Caste, Religion and Region in their decision making.

This study is important and unique in various aspects. Firstly, No study has been done to assess the differences in consumption expenditure across household based on highly nutritious food consumption. Most studies till now have focussed on only calorie intake and minimum calorie requirement. Here, we argue that calorie intake is necessary, but getting calories only from rice and wheat makes consumption monotonous. Even the poorest of the poor household consumes rice. But its consumption of HNF is limited due to its inability to purchase them. As income rise the consumption decisions also change and the expenditure on HNF's increase.

Having focussed on the Indian households, the point to investigate is whether the rise in income is treated as the same by people of all caste, religion, and states. Prima facie it appeared that there is a difference in expenditure decisions across them. What might be the reasons contributing to these? Do the people living in North India spend the same on HNF as compared to South India? Do the OBC's increase their expenditure on HNF in the same proportion as the Brahmins? Do the Hindu's and Muslim's record the same increase in their consumption of HNF's? Why not?

4. Research Method

Household data from IHDS contained recorded data of 41,554 households from 1503 villages and 971 urban neighborhoods across India. This data contained 924 variables. Out of these, we used 30 variables for our research. States were categorized in North, East, West, South, Central and North East.

North included states of Delhi, Haryana, Himachal Pradesh, Jammu & Kashmir, Punjab, Rajasthan, Uttar Pradesh, Uttaranchal (presently Uttarakhand) South included states of Andhra Pradesh, Karnataka, Kerala, Pondicherry, Tamil Nadu. East included states of Bihar, Jharkhand, Orissa, Sikkim, West Bengal. West included states of Dadra & Nagar Haveli, Daman and Diu, Goa, Gujarat, Maharashtra. Central included states of Chhattisgarh and Madhya Pradesh. North East included the states of Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland, Tripura. The states were divided into these 6 zones so as to capture the vast geography of India in a condensed form.

Two new variables were created, 1) "Nutritious food" in which the expenditure on 6 highly nutritious food items (Pulses, Cereal Products, Meat, egg, milk and milk products) were taken and summed together. 2) PCCNF which recorded the per capita consumption on nutritious food by dividing Nutritious Food by the number of persons in that particular household. The remaining variables were changed so that it is more clearly understood.

Using data mining to predict the trends in the household data was an interesting idea. Data Mining was applied on this data set using SPSS Clementine 12.0. PCCNF was taken as the output variable and was treated with various individual input variables (religion, caste, ration card, no. of meals per day, state zone etc) and combination of input variables using C&R Tree. "The Classification and Regression (C&R) Tree node is a tree-based classification and prediction method. This method uses recursive partitioning to split the training records into segments with similar output field values. The C&R tree node starts by examining the input fields to find the best split, measured by the reduction in an impurity index that results from the split. The split defines two subgroups, each of which is subsequently split into two more subgroups, and so on, until one of the stopping criteria is triggered. All splits are binary (only two subgroups)." Clementine Help.

Later on, the per capita consumption of Nutritious food was categorized into 3. PCCNF in the range of INR

- [0-100): Low
- [100-200): Average
- >200: High

This is the discreetisation of range variable. We categorize this as per our own discretion in order to find a better and clear picture of the results which is more clearer and has less of numeric values. On the discreet and categorized values, we use the C5.0 modelling and note the results. "The C5.0 node builds either a decision tree or a rule set. The model works by splitting the sample based on the field that provides the maximum information gain at each level. The target field must be categorical. Multiple splits into more than two subgroups are allowed." Clementine Help

5.Empirical Results Using C&R Tree

- **Across Caste**
  The predicted (average) Per Capita Consumption of Nutritious Food of households across India in Rupees at 2005's current prices was INR 134.87. The caste was categorized into two nodes. The first node Brahmin, Others (32.48%) showed a predicted PCCNF of INR 178.20. Further sub dividing Brahmins (5.83%) had a predicted PCCNF of INR 190.74 while others (26.65%) consumed INR 175.38. Among the OBC, SC and ST, the OBC's (39.19%) showed a consumption of INR 121.19 followed by SCs (20.05%) at INR 109.89 and STs (8.28%) at INR 90.31 (Refer Fig 1)
- **Across Religion**
  Across religion, the Sikh and Jain community 2.72% of the whole sample, had the maximum consumption on HNF predicted at INR 204.5. Next to them are Christians who were 3.32 % of the sample consumed INR 158.68. Hindus formed the major part of the sample (81.36 %) and spent INR 132.35 per capita on consumption of HNF, which is slightly below the national average. The tribal population representing only 1.03% of the sample showed the drastic fall in HNF consumption, thereby spending only INR 72.70
- **Across Regions/States**
  Across regions, the households of North India spent on an average INR 187.65 per capita. We further sub divide North India and club Punjab & Haryana together and keep rest of the North Indian States together we notice an interesting result. Punjab and Haryana have a per capita consumption of HNF at INR 228.90. South Indian households have a per capita consumption of INR 124.02 while the West India has a predicted consumption of INR 119.53 . The Central and East Indian States show a PCCNF of INR 72.74 and INR 97.76 respectively. The North Eastern households, though 5% of the sample spent INR 148.35 which is above the national average of INR 134.88.
- **Across Income**
  Consumption of HNF also shows high correlation with income. Majority of households had their annual income between INR 0 to INR 27553.96 (44.28 % of households). These low incomes were clearly reflected in the expenditure of HNF's where the households spent an average of INR 92.52 per capita per month. The households who with annual income between INR 27553.96 to INR 35834.67 (9.83 % of households) had a PCCNF of INR 117.57 which is still below the National Average. Households with annual income above INR 35834.67 to INR 47773.70 (10.3 %) spent near the national average. Their average PCCNF was 131.23. Thus, we see that 64.41 %, i.e., around 2/3rd Indian households consumed nutritious food below the national average. This suggests that National Average is affected by the extreme values of the high consumption of the higher income households. Households with higher income subsequently showed proportionately higher consumptions of HNF's. All households with income greater than 47773.70 had a predicted consumption of HNF more than the national average. The highest average consumption of INR 285.27 was shown by households with income greater than INR 237665. Though they were just 2.28% of the sample.

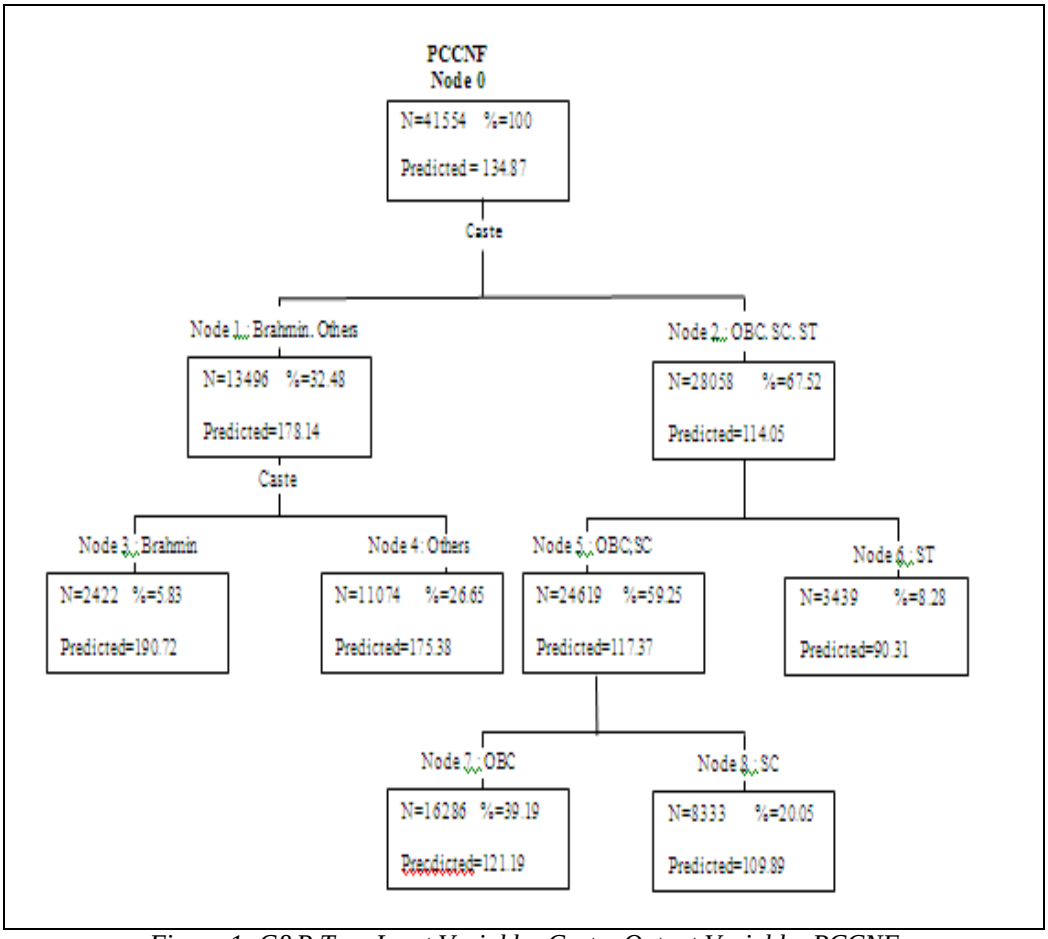

Figure 1: C&R Tree Input Variable: Caste; Output Variable: PCCNF

- **Across Sources Of Income**
  Comparing from the point of view of main income source of Indian households, agricultural and non agricultural labourers are the worse off in terms of consumption of HNF. Household with Agricultural labourers (13.60% of the total households) spent INR 80.36 on HNF while those with Non Agricultural labourers (17.58 %) spent INR 100.29 per capita. Artisans (5.96 %) spent on an average INR 125.27 while those involved in some type of cultivation (23.58%) spent INR 124.70. Households with other professions showed a consumption above the national predicted value of expenditure on HNF i.e. INR 134.88. Households involved in petty trade (4.53 %) spent INR 141.84. When the main income source was allied agriculture, business or some profession (7.78%) the expenditure on HNF per head was INR 172.99 whereas in case of Salary as the main source of income for the household (20.40 %) expenditure per head of HNF increased to INR 185.01. The highest consumption was shown by household whose main source of income was rental income from property or pension (3.93%). They consumed as high as INR 208.64 worth of HNF.
- **Across Other Variables**
  Applying the data mining techniques on other variables yielded many interesting results. A Variable POOR was assigned a value YES if the household was poor according to the 2005 poverty estimates which differed across states. Taking POOR as input variable when data mining was performed, we found that 19.70 % households which came in the category of poor; their average consumption of HNF was as low as INR 46.28. The households which were not poor (80.30 %) showed a higher average consumption of INR 156.61 . If the poor belonged to the ST community, their condition was all the more pathetic as they could consume only INR 29.80 worth of HNF's. The poor belonging to SCs and OBCs had a predicted average consumption of 47.37. Brahmins and Others of the poor category consumed better than the SCs and STs. Their consumption per capita was INR 59.68.
  Mining with multiple input variables, some interesting results were noted. Agricultural labourers of the North India were the best in consumption, spending INR 123.54 in HNF while those of Central & East India were the worst, consuming only INR 45.65. Among the Brahmins, the North Indian Brahmins show the maximum consumption consuming HNF worth INR 225.20 while the Brahmins of central, east and west India, consumed only INR 151.5 . The OBC and STs of North India consumed the maximum HNF INR 163.53. Among the regions of higher ST population, interestingly, the states of North East performed better than all others consuming INR 153.82 worth of HNF. While the tribals of Central India were the worst

performers consuming only INR 36.60. Even though the agricultural labourers consumed the least, the very few Brahmin agricultural labourers had a higher consumption of more than INR 100. While agricultural labourers coming under ST category consumed the least, INR 63.16. Similar was the result for Non-Agricultural labourers.

6. Results Using The C5.0 Modelling

In the C5.0 modelling, we take the discreet PCCNF as the output variable and take multiple input variables, like caste, religion, income; state zones, ration card etc are taken. We note various results which are in resonance with the results obtained through C&R tree. The noted results are as follows:

- 57.29% of the households living in central, east, south and west India had low consumptions of HNFs.
- 38.96% of households of North East had low consumptions while same percentage of households had average consumption. Only 22.08% of North Eastern household had high consumptions of HNF. In north India almost equal percentages of population where in the Low, Average and High category
- 49.14% of Brahmins and Others in North India had High consumption while 32.76% of them had average consumption. This shows that North Indian Brahmins and Others are comparatively well off than OBC, SC, ST who are just 25.35% in High and 34.09% in Average category..
- The Sikh are the most well off communities with more than 40% of them showing High consumptions of HNFs. They are followed by Jain and Christians with 28.28% of them consuming High HNF's.
- Among all other religions, only around 20% of them show High consumptions of HNF while almost 50% of them are in the Low consumption category.
- 61.17 % Households across the country with annual income less than 47810 have Low consumptions of HNFs. While households above this income level only 25.24% have Low consumptions
- Among all income groups, the Brahmins and Others show high consumption. All other communities are far behind the average consumptions of Brahmins and others.

7. Conclusion

Disparity in income is clearly seen in the empirical results and categorized results when we look at it from the point of view of per capita consumption of nutritious food. Interpreting this consumption as Income, the Brahmins and Others across India are better off than all other communities while the ST community is the poorest. A large income gap is also seen between the Brahmins and STs. The Sikh and Jain are the higher income earning religious communities, so their consumption of HNF is also higher compared to all other religions. The households of North India are better off than all others; this might be due higher agricultural productivity and production & consumption of milk and milk products in states of Punjab, Haryana and western Uttar Pradesh.

PCCNF highly correlates with increase in Income. Increases in Consumption of highly nutritious food can also be observed as increase in real income. It is important to note that agricultural labourers are the poorest in consumption of nutritious food and thus income. One of the reasons for this can be the disguised unemployment in agriculture. Renting a property and a salaried job generates higher income and thus higher consumption. Thus, summing up, we may argue that disparity in consumption of nutritious food provides a better picture of the income disparity of Indian households. Highly Nutritious Foods are the food items which a household has to buy at market prices, thus its basket is suited for estimating the real income.

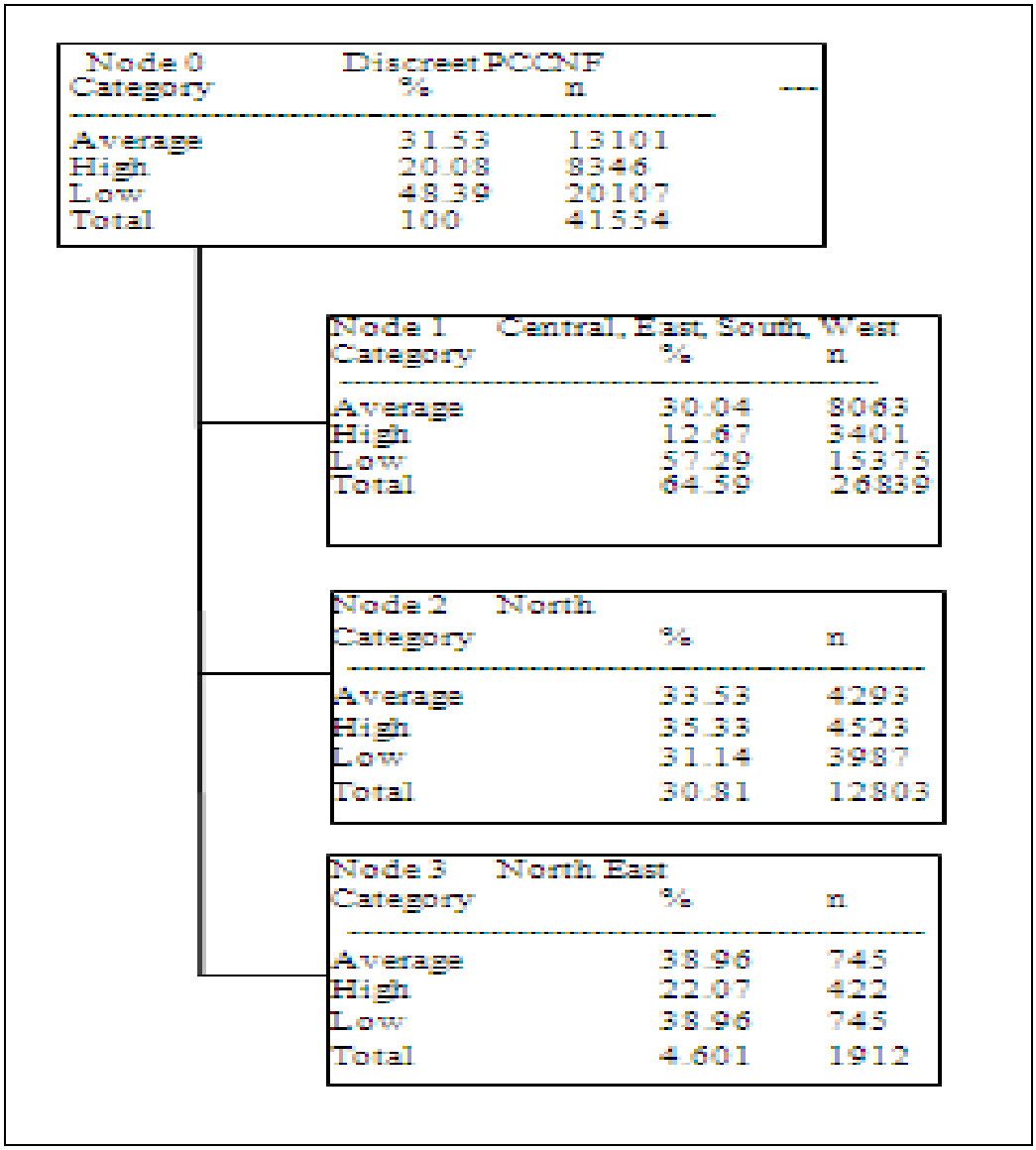

Figure 2: C5.0 Modelling

Input Variable: NEWS; Output Variable: Discreet PCCNF

8. References

1. Angus Deaton, Jean Dreze (2009): Food and Nutrition in India: Facts and Interpretation (Economic and Political Weekly, 14 February, 2009, Vol XLIV, NO 7, Pages 42-65
2. Desai Sonalde, Reeve Vanneman "India Human Development Survey" 2005, IPSCR 22626
3. Desai Sonalde, Amaresh Dubey, B L Joshi, Mitali Sen, Abusaleh Shariff and Reeve Vanneman (2010) : Human Development in India: Challenges for a society in Transition (New Delhi: Oxford University Press)
4. C&R tree, Help , Clementine 12.0
5. Pujari, Arun K: Data Mining Techniques (Universities Press, Hyderabad, India, Second Edition)
6. Madhura Swaminathan, Vikas Rawal (2011): Income inequality in village India: The role of caste ECINEQ WP 2011-207
7. Shraddha Srivastava, Amarnath Tripathi, A R Prasad: Regional Heterogeneity in Food Consumption and Nutrition Intake in India (Banaras Hindu University)
8. Mahendra Dev, S (2005): Calorie Norms and Poverty (Economic and Political).