The Importance of Research Data Digitization and Its Statistical Analysis-With Examples of Biogas Consumers of Nepal

Jyoti U. Devkota, Binu Hada, Chanda Prajapati, and Swechya Singh

***Abstract*—Renewable energy helps in solving twin problems namely fast depletion of energy sources and climate change. In this paper the results of a statistically sound research with minimum possible white noise have been stated. The data collection scheme was designed by keeping biogas use in the core and getting all the possible information about a typical middleclass Nepalese family inhabiting in rural areas, its economic and social background and change after biogas was used in their household. All sources of error right from the data collection, its digitization and finally its analysis were identified and eliminated. In the stage of questionnaire design all the possible answers to a question were foreseen and given as one of the categories. Pretesting of the questionnaire was also conducted on 30 household. A consumer profile database of biogas consumers is constructed. Then the information collected is stored in an electronic database. An online questionnaire is constructed to store all the information in this database. The digitization of the data and then its statistical analysis can similarly be done for other sources of renewable energy. Statistical analysis is then done for the analysis of impact of renewable energy in extenuating climate change.**

***Index Terms*— Climate change, database, renewable energy, statistical analysis.**

I. Introduction

Renewable energy is energy which comes from natural resources such as sunlight, wind, rain, tides and geothermal heat, which are renewable. It is derived from natural processes that are replenished constantly and replaces conventional fuels in four distinct areas: power generation, hot water/space heating, transport fuels and rural (off-grid) energy services.

A. Climate Change and Renewable Energy

Nepal is hit by long hours of load shedding and is facing acute problems of energy supply, shortage of cheap and efficient fuel, shortage of other many usable commodities and growing environmental pollution. Coupled with the global warming of 1.0 – 6.5 °C by the end of this century Nepalese citizens must start altering their lifestyles and prepare their respective societies for futures with different temperature and precipitation regimes. A switch over to renewable energy provides answer to both the problems of energy scarcity and climate change. The use of fuel wood for cooking by 68.4 % households and agricultural land holdings by 78.4% (Nepal Labor Force Survey, 2010) also make biogas plants suitable for our country. Most of the rural population has the tradition of raising cattle as an integral part of their farm. In addition to draft power and milk, cattle produce them necessary manure in the form of dung. This dung is the most important component of bio fuel. Government of Nepal's NAPA and LAPA (National Adaptation Program of Action and Local Adaptation Program of Action) has outlined various adaptation strategies at the national level and local level respectively. Promotion of renewable energy especially biogas to people most affected by climate change is not only one of our major goals but it also somehow endorses the governmental NAPA and LAPA objectives.

B. Importance of Digitization of Collected Data

In the developing world especially Nepal the quality of the collected data is not so reliable. It is mainly due to the lack of awareness of the various organization and individuals involved in the collection and supply of data. Conduction of well planned surveys and data collections processes result in minimum errors and lack of ambiguity. Data play a vital role in any kind of research. In socio economic studies related to human beings, data are collected through a list of questions called questionnaire. It plays a crucial role in the conduction of a survey. Although data collection is a very important aspect of research its safe handling and safe keeping are also equally important. The digitization of the collected data through the construction of consumer profile database ensures this. Further this database keeps in the entire loads of collected data in an electronic format which ensure ease of future reference.

C. Literature Review

Mathematical modeling is an interdisciplinary branch of science, which helps in understanding, explaining and predicting the state of nature including the changes associated with our surrounding and society [3]. Various aspects of our surrounding such as environment in which we live, biological objects that surround us, and demographical changes that our society undergoes are studied using tools of mathematical modeling [2], [4], [5], [7]. Statistical Modeling can be of great importance for planning and managing resources at both the local and national levels [6].

Climate is a paradigm of complex system. Stochastic Modeling is a scientific effort to bind and predict nature using mathematical and statistical equations. It is widely used in various fields. W. Albers et al have used stochastic

Manufacture received January 30, 2012; revised March 6, 2012. This work is funded by NORAD, SINTEF under Renewable Nepal Project grant number RENP-10-06-PID-379.

Authors are with Kathmandu University, Nepal (e-mail: devkotajb@gmail.com, drjdevkota@ku.edu.np).

modeling to predict the growth of academic knowledge among university students [2]. Helle Aagaard-Hansen and G. F. Yeo [1] applied stochastic modeling to find a discrete generation birth, continuous death population growth model and its approximate solution. L.M. Ricciardi [19] constructed discrete stochastic models to study the effects of environmental variance on the asymptotic behavior of a population subject to logistic growth in random environment. Kenneth G. Manton [18] forecasted the health status changes in an aging U.S. population. Similarly Mark S. Boyce [5], Desharnais [6], James Matis et al [17] have analyzed statistically problems from Ecology. Statistical modeling has been applied in various fields such as mathematical demography in Jyoti Upadhyaya Devkota [7, 8, 9, 14, 15, 16, 17], Biotechnology in Jyoti Upadhyaya Devkota [22], Environmental pollution in Jyoti Upadhyaya Devkota [3, 4, 11, 12], Burn Injuries in Jyoti Upadhyaya Devkota [13], Queuing Theory in Jyoti Upadhyaya Devkota [10, 21].

In this paper firstly different sources of errors from data collection to its statistical analysis are identified. Measures in their identification and minimization are discussed. This is followed by details of the questionnaire and the survey collecting information on socioeconomic parameters of consumers of renewable energy and then the consumer profile database constructed thereafter. The highlights of consumer profile database stresses on the importance of data digitization and its safe handling. These are all relevant to all the sources of renewable energy, but here the results are mainly on the data of 400 households consuming biogas as a source of renewable energy. It is followed by the statistical analysis of the parameters measuring the impact of a source of renewable energy in mitigating climate change.

II. Methodology

A well planned survey and a well tested questionnaire are very important pre requisites for a statistically sound research. A judiciously designed questionnaire pre tested and refined with all the possible answers classified and mentioned in several categories, collects information with minimum possible error. Error could creep into the research at various stages from filling up of the questionnaire to entry of the data and its analysis. If all the answers to the questionnaire are worked out properly categorized into several categories which are given as possible options to the interviewee this minimizes the chances of error due misunderstanding between interviewee and interviewer. And if the database is designed carefully then possibility of errors in data entry can be carefully eliminated by the use of input masks, default values and formats in the field properties. So a well planned questionnaire results in categorical data which has to be analyzed by categorical data analysis.

A. Questionnaire

A questionnaire comprising of 59 questions and requiring 40-45 minutes to complete it was designed. The draft questionnaire comprised of 62 questions before it was tested on 30 households in Sudal VDC, Bhaktapur. Then this questionnaire was refined according to the responses of the interviewee to 59 questions. This questionnaire collects information in details about the degree of change in their life style after biogas was used by them as a source of renewable energy. It also enquires about the role of gender in different activities of the household and their empowerment after the use of renewable energy. Thus the questionnaire was designed with an objective of keeping biogas use in the core and getting all the possible information about a typical middleclass Nepalese family inhabiting in rural areas, its economic and social background and change after biogas was used in their household. In these 59 questions information was collected on very relevant topics such as the age distribution of 400 households comprising of 2272 individuals of different age groups, amount of landholdings, livestock, their fuel wood expenses before and after the installation of plants etc. So with the structure of the questions information can be obtained about the households that haven't installed biogas as a source of renewable energy.

Firstly the age distribution of the family is studied by asking questions related to age and educational status of the family members. Then the socio economic status of the consumer is assessed through questions on the amount of livestock, occupational status, extent of agricultural holdings, type of house, ownership of the house(own or rented) and source of water. Then the impact of renewable energy especially biogas in their life is studied through questions on source from which they first heard of biogas, major purpose of its installation, name of organization from which they got support in cash or kind in its construction, their expense at installation. Then the frequency of operation of the biogas plant is studied by asking about frequency of feeds. This is followed by questions on the winter summer differential in the in the gas produced by biogas plant. The collection of firewood, family members involved in the collection of firewood amount of firewood used before and after, fuel expenses before and after, frequency of hospital visits before and after, use of spare time, details of cooking activities through the biogas plant are all asked in details. Before and after questions aim at knowing in details about the change in the lifestyle after the consumers started using biogas as an energy source. To reduce ambiguity of answers all the possible answers were categorized into several categories thus ensuring minimization of errors due to ambiguous answers. Questions on who keeps the profit, who decided for the plant, who is responsible for its operation, under whose name it is registered are questions aimed at studying the female empowerment.

The questionnaire was tested on 30 households in Sudal VDC, Bhaktapur and finalized. The response of the consumers was noted and the answer options were accordingly refined and updated to remove errors, ambiguity of answers and smoothness in the flow of answers. Research assistants were trained on the importance of each question and possible setbacks if the answers were not recorded correctly. They were trained to be polite and not easily offended in case of lack of cooperation by the respondents. But on the whole the consumers cooperated in the entire data collection process. The data was collected on 400 households, where 370 households were the consumers of Rapti Renewables and Energy Services. Two research assistants assisted in the pretesting of the questionnaire and five research assistants helped in the entire data collection process.

The entire data collection process of 400 households comprising of 2272 individuals was completed in 15 days.

After the data was collected in filled up printed questionnaire then it was digitized into an electronic form in a database. Different variables in the questionnaire were carefully analyzed and the design of the database was carefully worked out.

B. The Survey of 400 Households

The survey was conducted in three regions firstly in Sudal VDC, Bhaktapur where the information was collected on 30 households and then Simara, in Bara with 300 households and lastly in Sarlahi with 70 households. The entire data collection process took 15 days. In Sudal VDC the questionnaire was tested refined and finalized. The time taken to fill up the entire questionnaire was noted (40 – 45 minutes) and it helped in planning for the recruitment of the field workers for the two major surveys in Sarlahi and Simara. The approximate time for questionnaire per household was found to be 45 minutes initially. Due to several practice during the pre-test, the time decreased down to 35-40 minutes. The questionnaire was tested only on two households on the first day of the pretest and then on the second day after making necessary changes in the questionnaire it was tested on five households and refined further thereafter. The pre-test helped us to explore different types of idea to cross-check and seek the required information from the interviewee. The pre-test also helped the data collection team to mix up well with the locals and build up confidence in the ability to handle the locals which was very beneficial for the other two major surveys on 300 and 70 households respectively.

After that the data was collected from Sarlahi and Simara districts with information collected on 70 and 300 households respectively. A team of two research assistants was recruited for the pilot testing of the questionnaire in Sudal VDC and then five research assistants collected data on 370 households from Simara and Sarlahi.

C. The Consumer Profile Database of Renewable Energy specially Biogas

Collection of statistically sound, minimum error data leads to its digitization through the construction of database. This ensured the safe handling of data and the ease in its statistical analysis by using statistical software. MS Access was used for construction of the Consumer Profile Database. To ensure the accuracy of the data entry an online questionnaire was prepared where each entry was controlled by formatted structures like input masks and default values. The answers in the filled up questionnaire were looked in details and repetitive same answers were coded under default values and input masks. This ensured correct and fast entry of the data.

Thus to summarize the whole research methodology was split into a number of phases. The phases were sequentially the design of the draft questionnaire, identification of sources of error during data collection, then pretesting of the draft questionnaire, training of the field assistants on art of collecting error free data, conduction of main survey, thorough study of the filled up questionnaire for designing the database, design of the database, construction of online forms with input masks and default values for error free and fast data entry, statistical analysis of the data.

D. The Statistical Analysis of Impact of Biogas in Mitigating Climate Change

The table I shows the relationship between the maintenance of special shed for storage of firewood and firewood savings. Out of 210 houses that has maintained a shed, in 14 houses up to 30 kg firewood is saved, in 43 houses 30-50 kg firewood is saved and in 153 houses above 50 kg firewood is saved. Among 190 houses that do not maintain shed in 25 houses up to 30 kg is saved, in 72 houses 30-50 kg is saved and in 93 houses above 50 kg firewood is saved. Also out of 400 household using biogas plants of different sizes 39 households save up to 30 kg, 115 households saved between 30 – 50 kg and 246 households saved more than 50 kg of firewood per year. Table 2 shows the 95 percent confidence interval for the amount of firewood saved.

TABLE I: RELATION BETWEEN FIREWOOD SAVED AND STORAGE OF FIREWOOD IN A SEPARATE SHED

		Firewood Saving			
		Up to 30Kg	30-50 Kg	Above 50 Kg	Total
Maintain Firewood Storage in a special shed	Yes	14	43	153	210
	No	25	72	93	190
	Total	39	115	246	400

TABLE II: 95 PERCENT CONFIDENCE INTERVAL FOR AMOUNT OF FIREWOOD SAVED

Amount of Firewood Saved	p and q	Confidence Interval
Up to 30 kg	p = 39/400 = 0.09 q = 361/400 = 0.91	$.07 \le p \le .11$
30 – 50 kg	p = 115/400 = 0.28 q = 1- p = 0.72	$.24 \le p \le .32$
Above 50 kg	p = 246/400 = 0.61 q = 1- p = 0.39	$.57 \le p \le .65$

In table II

0.09-1.96 $\sqrt{\frac{0.09*0.91}{400}} \le p$ 0.09+1.96 $\sqrt{\frac{0.09*0.91}{400}}$ **is equivalent to** $.07 \le p \le .11$ for up to 30 kilograms, 0.28-1.96 $\sqrt{\frac{0.28*0.72}{400}} \le p$ $\le$ 0.28+1.96 $\sqrt{\frac{0.28*0.72}{400}}$ **is equivalent to** $.24 \le p \le .32$ for 30 – 50 kilograms and 0.61-1.96 $\sqrt{\frac{0.61*0.39}{400}} \le p$ $\le$ 0.61+1.96 $\sqrt{\frac{0.61*0.39}{400}}$ **is equivalent to** for $.57 \le p \le .65$ more than 50 kilograms.

III. RESULT AND DISCUSSION

This research was conducted with an aim of having minimum possible error in data and subsequently in its analysis. All the sources of error at various stages of research were identified namely during questionnaire design through ambiguous questions and ambiguous answers, conduction of survey through ill trained and impatient interviewers, confusing answers of the interviewee due to failure in

mentioning all possible answers as a choice, slow and faulty data entry due to not well planned database and erroneous conclusions due to the use of inappropriate statistical tools. The chances of ambiguous questions and answers were eliminated by conducting the pre test of the questionnaire and by working out all the answers of each question before hand and categorizing it and using categorical data analysis as one of the statistical tools. Pretesting of the draft questionnaire improved the interviewing and socializing skills of interviewers. These two interviewers passed their skills and experiences to the rest of the team in a training program conducted for the rest of the three research assistants involved in two major surveys. Here the entire questionnaire along with their experiences and response of the interviewee was discussed in detail. This improved the interviewing skills of the rest of the team and also enhanced the speed of data collection with no compromise to the accuracy. Then the filled up questionnaire was analyzed in detail and the design of the database was carefully planned out. This ensured fast and correct data entry.

Initial study during the data collection on 30 households in Sudal revealed that biogas plant was mostly utilized for purpose such as cooking, lighting and fertilizer. Six visits from 13 th August to 17 th September 2010, with questionnaire refined, edited and revised after each visit resulted in data collected from 30 households. Further the data collection skills, interviewing skills of the team responsible for data collection were also developed. This also resulted in confidence building of the project team which further prepared them for two major surveys held during October and End of November in two different regions of Nepal located in terai region. Sudal is located in hills of Nepal unlike Sarlahi and Simara which are located in Terai. Sudal is a village and Village Development Committee in Bhaktapur District in the Bagmati Zone of central Nepal.

Data was collected on 370 households using biogas as a source of renewable energy and located in the villages Phattepur, Dummarwana, Simara, Inaruwasira, Amlekhganj, Haraiya in Bara district. This survey was conducted from October 2 to October 11, 2010.

The information collected on 400 households using biogas as a source of renewable energy was stored in a consumer profile database, specially designed for this study. This highlights the fact that not only the data has to be collected in a systematic manner but it also has to be kept in a database in a digitized form so that it can be used easily for future reference. Fig.1 gives an overview of the consumer profile database. First 386 variables necessary for research were identified during the design stage of the database. Accordingly 386 field names, their descriptions, default and input masks were fixed. So the entire information was coded according to question numbers into 23 tables. All the information collected from 59 questions of the questionnaire was included in 386 variables and 23 tables of the consumer profile database named Edited Biogas Consumer Profile Database. The data collected from 400 households was entered into an online form. This made the data entry process fast and accurate. As the entire questionnaire was classified into 23 tables so accordingly we had 23 sub forms which made the data entry process smooth and fast. Fig. 2 gives the form view of the Database specially the sub form Biogas questions. An electronic version of the questionnaire in the form of online questionnaire was made. This ensured ease in entry of the collected data. Further the accuracy of the information being entered into the database was made sure right during the design of the database by using input masks and default values.

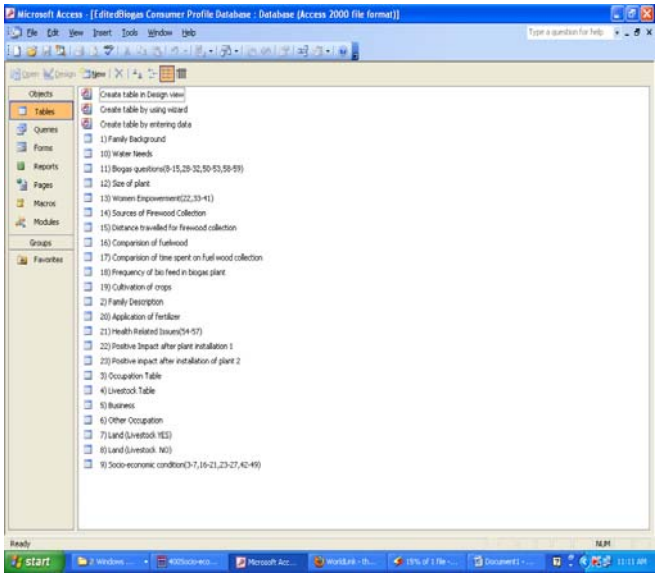

Fig. 1. Overview of the Consumer Profile Database

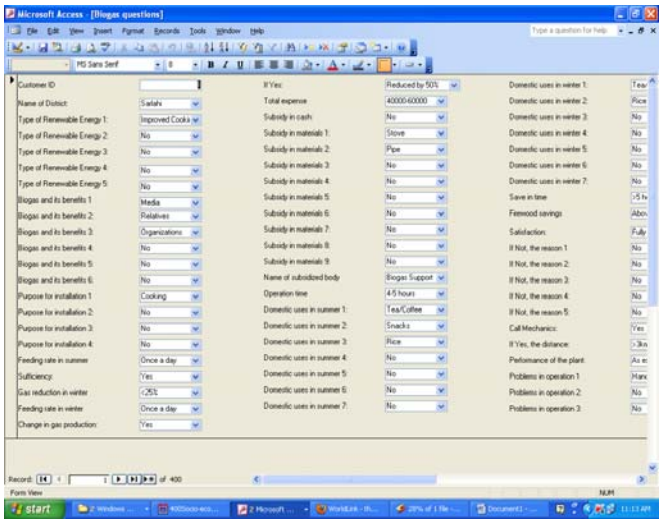

Fig. 2. The Form View of the Database

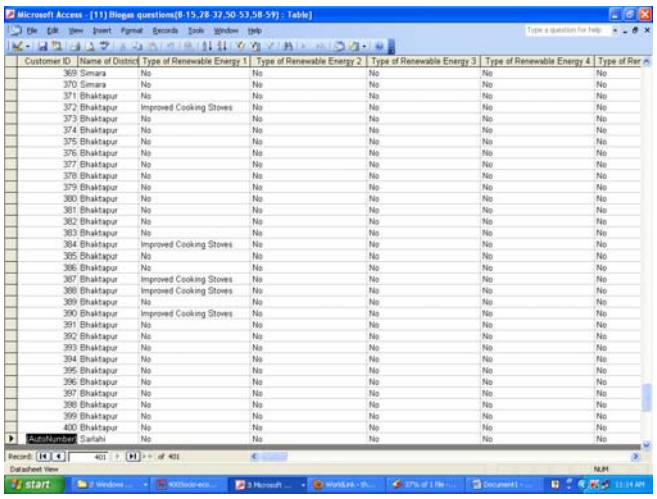

Fig. 3. Data Entry in Table View

Fig. 3, below gives the table views of the entered data of the sub form named Biogas questions. This table view of the data enables the export of the data into other database for advanced statistical analysis. Fig. 4 gives the detailed description of field names in sub form Biogas Questions along with the input masks and default values. For example the field names Purpose of Installation 1 gives cooking as a default value. This was decided after thorough study of the filled up printed form of Questionnaires obtained the after completion of two major surveys.

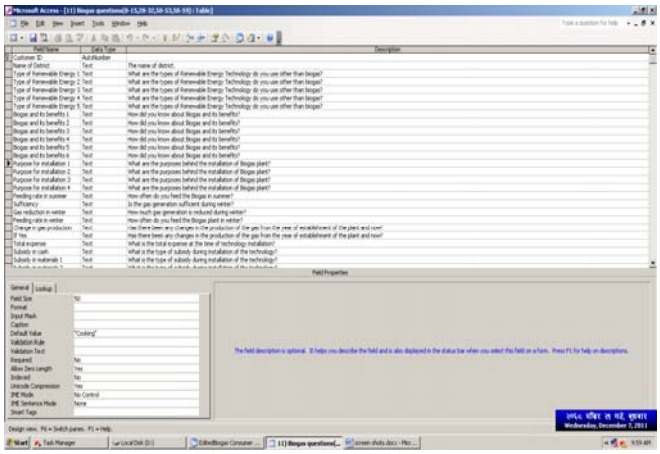

Fig. 4. Properties of the Field Names and their Descriptions

Table I shows the cross tabulation of 400 household between firewood storage in a special shed and firewood saving. Firewood saving is directly correlated to the decrease in deforestation as forests are the main source of fuel wood in Nepal. From table 2 we can be 95 % confident that if a source of renewable energy especially biogas is installed in all the potential households of Nepal between 57 % to 65 % will save more that 50 kg, 24% to 32 % save between 30 to 50 kg and 7 % to 11% save up to 30 kg. This will reduce the pressure on forests as they are the major source of firewood in Nepal and hence resulting in the reduction of green house gas emission. This statistically justifies the fact that a switch over to a source of renewable energy is one of the most suitable adaptation strategies to climate change.

IV. Conclusion

Renewable energy can play a crucial role in our part of the world where the energy resources are depleting at a very fast rate. In developing countries firewood is one of main sources of energy. Forests are main sources of firewood. In this paper it has been statistically verified that a switch over to a source of renewable energy reduces the dependence of households on firewood and hence on forest. This in turn reduces the carbon dioxide emission and is a useful adaptation strategy for climate change. 95 percent confidence intervals have been constructed to statistically validate the results. The statistical analysis is done on the data collected from a carefully designed study aimed to have minimum error from all the possible sources. The questionnaire was designed, tested and revised. All the possible answers to the questions were worked out and mentioned as categories to reduce the error due to ambiguity of questions asked. Then the database was carefully designed. Further input masks and default values were identified after a detailed study of the filled up questionnaire. This prevented error in data entry. The database comprises of information on 386 variables collected from 59 Questions.

Acknowledgment

The authors gratefully acknowledge the contributions of Rapti Renewable and Energy Services and other members of the data collection team.

References

[1] H. Aagaard-Hans and G. F. Yeo, "A stochastic discrete generation birth, continuous death population growth model and its approximate solution", *J. Math.Biology*, Vol. 20, no. 1, pp. 69-90, 1983.

[2] W. Albers, R. J. M. M. Does, Tj Imbos, M. P. E. Jansen, "A stochastic growth model applied to repeated tests of academic knowledge", *Psychometrica*, Vol. 54, no. 3, pp. 451- 466, 1989.

[3] B. Devkota, B. B. Panthi, J. U. Devkota, "Pesticidal Activities of Mixture of certain Botanicals against coffee stem Borer", *Abstracts of Fifth National Conference on Science and Technology*, Nov. 10 – 12, 2008, pp. 69, 2008.

[4] B. B. Panthi, B. Devkota, J. U. Devkota, "Effect of botanical pesticides used to control coffee pests on the soil fertility of coffee orchards", *J. Agriculture and Environment*, MOAC, Vol. 9, 2008.

[5] B. Mark, C. V. Haridas and C. T. Lee, "Trends in ecology and evolution", *Science Direct*, Vol. 20, no. 20, pp 96-104, 2005.

[6] R. A. Desharnais, R. F. Costantino, J. M.Cushing, "Experimental support for scaling rule for demographic stochasticity", *Ecology Letters*, Vol. 9, pp. 537 – 547, 2006.

[7] J. U. Devkota, R. S. Singh, "Mathematical modeling of mortality for countries with limited and defective data", *J. Applied Statistical Sciences*, Vol.19 no.1, 2011.

[8] J. U. Devkota, R. S. Singh, "Deterministic and probabilistic models with applications to modeling fertility data", *J. Applied Statistical Sciences*, Vol.18 no.2, 2010.

[9] J. Upadhyaya, "Mathematical Analysis of Age Structured Population data", *Abstracts of Fifth National Conference on Science and Technology*, Nov. 10 – 12", pp 390, 2008.

[10] J. Upadhyaya, R.P.Ghimire, T. Upadhyaya, " Finite Buffer Non-Preemptive Priority Queue with Heterogeneous Servers under Randomized Push out scheme", *Abstracts of Fifth National Conference on Science and Technology,* Nov. 10 – 12, pp 382, 2008.

[11] J. U. Devkota, S. Aryal, "Mathematical modeling of residual effects of distillery effluents application by finite difference method", *Kathmandu University Journal of Science, Engineering and Technology*, Vol. 1, no. 2, 2006.

[12] J. U. Devkota, "*Report on mathematical modeling of post methanated effluents*", Kathmandu University 2006.

[13] J. U. Devkota., V. P. Saxena, "Mathematical modeling applied to burn injuries", *Kathmandu University Journal of Science, Engineering and Technology*, Vol. 1, no. 1, 2005.

[14] J, U. Devkota, "Rural and urban effects on mortality and fertility", *Archives of the Himalayan Environment*, Vol. 1, issue 1, pp16-24, 2002.

[15] J. U. Devkota," Some models reflecting and projecting Nepal's fertility scene", *The Nepali Mathematical Sciences Report*, Vol. 19 (1 & 2), pp. 101-116, 2001.

[16] Jyoti Upadhyaya Devkota, "Calculation and Modeling of Parity Progression for Nepal", *The Nepali Mathematical Sciences Report*, Vol. 18 (1 & 2), pp. 79-102, 1990-2000.

[17] J. U. Devkota, "*Modeling and Projecting Nepal's Mortality and Fertility*", unpublished PhD. thesis submitted in the Department of Informatics and Mathematics, University of Osnabrueck, Germany, 2000.

[18] M. James, T. R. Kiffe , "Stochastic population models", Lecture Notes in Statistics, Springer Verlag, 2005.

[19] K. G. Manton," Forecasting health status changes in an aging U.S. population", *Climate Change*, Vol. 11, no. 1-2, pp 179-210, 1987.

[20] L. M. Ricciardi , " On a conjecture concerning population growth in random environment", *Biological Cybernetics*, Vol. 32, no. 2, pp. 95-99, 1979.

[21] R. P. Ghimire, J. Upadhyaya, Tulsi Upadhyaya, " Priority Queues with Balking and Reneging", *Abstracts of Fifth National Conference on Science and Technology*, Nov. 10 – 12, 2008, pp. 383, 2008.

[22] S. M. Dixit, J. Upadhyaya, et al., " Assessment of Aqueous extracts of selected nepali high altitude medicinal and aromatic herbs as antimicrobial agents against selected gram – negative and gram-positive opportunistic pathogens", *Research Journal of Biotechnology*, Vol. 5, no. 2, pp 37 – 42, 2010.

Jyoti U. Devkota, PhD., who is working as an Associate Professor of Statistics at Kathmandu University**,** has obtained PhD.(Dr. rer. Nat.) from Germany under DAAD fellowship. She is also the project leader of Renewable Nepal Project RENP-10-06-PID-379. Her research interest lies in interdisciplinary applications of Statistics and Data Analysis tools into various fields. She is also a member of Multi-stakeholder Climate change Initiatives Coordination Committee (MCCICC) under the coordination of Ministry of Environment, Government of Nepal. Her contact e mail address is devkotajb@gmail.com and drjdevkota@ku.edu.np.