January 2003 ORES SP03-01

Modeling Nitrate Concentration in Ground Water Using Regression and Neural Networks

Nacha Ramasamy
Palaniappa Krishnan
John C. Bernard
William F. Ritter

FOOD & RESOURCE ECONOMICS

FREC Staff Paper

Department of Food and Resource Economics • College of Agriculture and Natural Resources • University of Delaware

MODELING NITRATE CONCENTRATION IN GROUND WATER USING REGRESSION AND NEURAL NETWORKS

N.Ramasamy, Graduate Student, Operations Research Program, Department of Food and Resource Economics, University of Delaware, Newark, DE
531, S.College Avenue, 212, Townsend Hall, University of Delaware, Newark, DE – 19717 – 1303.
Ph: (Office) 302-831-6242 (Home) 302-731-4378 Fax: 302-831-6243 Email: nacha@udel.edu

P.Krishnan, Director, Operations Research Program, Department of Food and Resource Economics, University of Delaware, Newark, DE.
531, S.College Avenue, 212, Townsend Hall, University of Delaware, Newark, DE – 19717 – 1303.
Ph: (Office) 302-831-1502 Fax: 302-831-6243 Email: baba@udel.edu

J.C.Bernard, Assistant Professor, Department of Food and Resource Economics, University of Delaware, Newark, DE.
531, S.College Avenue, 229, Townsend Hall, University of Delaware, Newark, DE – 19717.
Ph: (Office) 302-831-1380 Fax: 302-831-6243 Email: jbernard@udel.edu

W.F.Ritter, Professor, Department of Bioresources Engineering, University of Delaware, Newark, DE.
531, S.College Avenue, 264, Townsend Hall, University of Delaware, Newark, DE – 19717.
Phone : (Office) 302-831-2468 Fax: 302-831-2469 Email: william.ritter@udel.edu

ABSTRACT

Nitrate concentration in ground water is a major problem in specific agricultural areas. Using regression and neural networks, this study models nitrate concentration in ground water as a function of iron concentration in ground water, season and distance of the well from a poultry house. Results from both techniques are comparable and show that the distance of the well from a poultry house has a significant effect on nitrate concentration in groundwater.

INTRODUCTION

Ground water is the major source of water in Sussex County, Delaware. The Coastal Sussex water quality management program found areas in Sussex with excessive levels of nitrate concentration in drinking water supplies. Sussex county is the southern most County in Delaware and is the birthplace of the broiler industry. The headquarters of three of the top twenty broiler producing companies in the country are located in Sussex County. Poultry manure from the broiler industries could be a major cause for having excessive levels of nitrate concentration in ground water [6]. This is because poultry manure is spread on the farmlands and they leach nitrates into ground water.

Nitrates, being extremely soluble in water, move easily through the soil and into the ground water. Ingestion of excessive amounts of nitrates causes ill health effects in infants less than six months old and susceptible adults. It causes "blue baby syndrome" or Methemoglobinemia in infants, which can lead to brain damage and sometimes death. Also, the Maximum Contaminant Level (MCL) for nitrates in public drinking water established by the federal government is 10 mg l^{-1}. About 32% of the wells in coastal Sussex and 21% of the wells in non-coastal Sussex were found to have an average nitrate concentration above the MCL (Ritter and Chirnside, 1982). The areas where those wells are located have numerous broiler production units. Agricultural activities and operations that disturb and aerate the soil enhance the soil nitrogen oxidation, which, along with the fertilizer nitrate components will be leached to the groundwater (Robertson, 1977).

According to published literature, some work has been done pertaining to modeling nitrate concentration in ground water. DRAINMOD-N, a mathematical model, was used to estimate the accumulated nitrate loss in drainage and subsurface water and evaluate different water pollution scenarios [8]. Soil and Water Assessment Tool (SWAT) was used to model the effect of changing land use patterns and practices on nitrate and phosphate loads to surface and ground water [2]. SWAT is a combination of the earlier models Erosion Productivity Impact Calculator (EPIC) and Groundwater Loading Effects of Agricultural Management Systems (GLEAMS). EPIC simulates relevant biophysical processes, modeling cropping systems for long time periods and determines the effect of management on soil erosion and productivity. GLEAMS is a continuous, simulation model and was developed to evaluate the impact of management practices on potential pesticide and nutrient leaching within, through and below the root zone.

The spatial distribution of nitrates in groundwater was assessed when conventional management and best management practices were applied [4]. GMS and GLEAMS was used to apply particular better management practices. Artificial neural networks was used to predict the pesticide and nitrate contamination in

rural private wells [5]. Depth to aquifer materials from land surface, well depth and distance to cropland were used as input parameters and concentration of pesticides or nitrates were the outputs. A set of neural networks was used to predict soil water content at a given depth as a function of soil temperature and soil type and was compared with a multiple regression model [1]. Neural networks were generally able to predict the soil water content over time but the regression model did not perform well in following the trends in the data over time. Most of the above models were physical models. They are non stochastic and are based on physical and chemical reactions and mathematical equations. Very little work has been done so far in building stochastic models to predict nitrate concentration in groundwater using regression and neural networks.

The objective of this study was to model nitrate concentration in ground water using regression and neural networks. In our study, we assume that nitrate concentration in ground water depends on iron concentration in ground water, season of the year and distance of the well from a poultry house.

DATA AND METHODOLOGY

The data set used for this study was obtained by Ritter and Chirnside (1982). No data has been collected since and no modeling work has been done with this data. This data was collected from 119 wells in Sussex County, Delaware. A total of 627 observations was collected from these wells during the different seasons of the year. The data set was divided into two groups: 484 observations (77% of the data set) for building the model (training data set) and 143 observations (23% of the data set) for validating the model (validation data set). Data is available upon request.

Regression

Nitrate concentration (*NO3* measured in mg l^{-1}) was the dependent (response) variable and iron concentration (*FE* measured in mg l^{-1}), season and distance of the well from a poultry house (measured in meters) were the independent (explanatory) variables. Iron and nitrogen together form a compound called ferrous nitrate, which is very unstable. It decomposes and liberates nitrogen gas, which escapes into the atmosphere. So as iron concentration increases, nitrate concentration will decrease as seen in Fig 1. Therefore, *FE* is expected to have a negative coefficient of regression. Sometimes, iron concentration (*FE*) was zero. For season, dummy variables *spring*, *summer* and *fall* were created. Winter was taken as the base. During spring and summer, plants tend to absorb the rainwater and a very minimal amount of rainwater only seeps through the soil with nitrates into groundwater. In fall, the plants are harvested and the rainwater seeps through the soil with nitrates directly into groundwater. Therefore, *fall* is expected to have a more positive coefficient of regression than that of *spring* and *summer*. The area around a poultry house was divided into three zones as the effect of the poultry house was studied in these zones. Zone 1 is the region where the distance of the well from a poultry house is less than 150 m, zone 2 is the region where the distance of the well from a poultry house is between 150 m and 300 m and zone 3 is the region where the distance of the well from a poultry house is greater than 300 m. For distance of the well from the poultry house, dummy variables *D2* and *D3* were created. When *D2* and *D3* were zero, the distance was less than 150 m (zone 1). When the distance was between 150 m and 300 m (zone 2), *D2* was 1, otherwise 0. When the distance was greater than 300 m (zone 3), *D3* was 1, otherwise 0. Distance less than 150 m (zone 1) was taken as the base. As the distance of the well from the poultry house increases, nitrate concentration generally decreases. Therefore, *D2*and *D3* are expected to have negative coefficient of regression and *D3* must be more negative than *D2*.

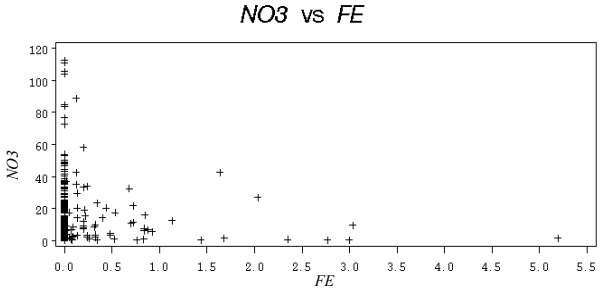

FIG 1

Various functional forms are present and care must be taken in choosing the appropriate form. For this study, a Log-Level model was assumed. The nitrate concentration was never zero. When the iron concentration was zero, season was winter and distance was less than 150 m, all the explanatory variables become zero, which prevented us from assuming a log form for the explanatory variables. The nitrate concentrations were high and taking a log for this is quite acceptable. So a Log-Level model was chosen.

The econometric model was given as:

$$LNO3 = b_0 + b_1 (FE) + b_2 (spring) + b_3 (summer) + b_4 (fall) + b_5 (D2) + b_6 (D3) \quad (1)$$

where *LNO3* was the log of the nitrate concentration in ground water (*LNO3*=Log(*No3*)), *FE* was the Iron concentration in ground water, *spring*, *summer* and *fall*

were the dummy variables created for season and *D2* and *D3* were the dummy variables created for the distance of the well from a poultry house. The SAS software (SAS inc., 1990) was used as a tool to build the model.

Neural Networks

Neural networks are composed of a large number of highly interconnected processing elements called neurons. Their function is determined by the structure of the network, strength of the connections and the processing performed at each computing node or neuron. Neural networks are well suited for problems that are too complex for conventional technologies. These problems include the ones, which do not have an algorithmic solution but can easily be solved by humans. Neural networks are used for pattern recognition and classification. A neural network can generalize in making decisions using imprecise input data. Humans apply knowledge they gained from past experience to solve problems. Neural networks mimic this problem solving process of the human brain. They cannot be programmed to do a specific task; but rather, learning is done with an example, by training or exposing to a truthed set of input/output data where the training algorithm iteratively adjusts the connection weights. These connection weights show the strength of input and the knowledge necessary to solve the problem.

All networks consist of at least three hierarchical layers, namely the input layer, one or more hidden layers and the output layer. All these layers are fully connected. Inputs or patterns are fed to the network via the input layer, which communicates with one or more of the hidden layers where the actual processing and weight adjustments are performed. The hidden layers then pass this information to the output layer where the actual output is received.

The most common method used for adjusting the weight is back propagation; these networks are called Back Propagation Neural Networks (BPNN). For every observation, the difference between the desired and actual output, called the network error, is calculated and propagated backwards through the network. In BPNN, weights are adjusted so that the square of the network error is minimized. This process is repeated until the error is within the acceptable range. Once the network is trained to a satisfactory level, it can be used to predict the output for the input, which the network has not seen before. But this time, the network works only in the forward propagation mode and the output is retained. For this study we used NeuroShell Easy Predictor (Ward Systems Group Inc., 1997).

RESULTS AND DISCUSSION

Regression

Regression results for training data set

The regression equation was given as:

$$LNO3 = 2.7628 - 0.2411\ FE - 0.0088\ spring + 0.0721\ summer + 0.4717\ fall - 2.0312\ D2 - 1.8524\ D3 \quad (2)$$

Other regression outputs are given in Table 1. The R-square value was 0.4476. This value of R-square shows that 44.76% of the variation in *LNO3* can be explained by the model. At a 5% significance level, the F-value for the model was 64.41. As the p value was less than 0.0001, the above regression model was statistically significant. The Variation Inflation Factor (VIF) was less than 10 for all explanatory variables and also the Condition Index from the Collinearity Diagnostics was less than 30 for all explanatory variables. Thus, multicollinearity was not a problem. The Chi-Square statistic for the heteroskedasticity was 45.21 and the p-value was 0.0006. At a 5% significance level, we reject the null hypothesis of homoskedasticity. This result implies that the model has heteroskedasticity, causing the variances to be biased. So t-values and p-values calculated for the regression coefficients by SAS will be misleading. Feasible GLS and other usual methods were employed to remove heteroskedasticity. Still only the signs of the regression coefficients changed and heteroskedasticity was not removed. So the heteroskedastic robust t-values were calculated for the regression coefficient from White's heteroskedasticity consistent variances, which is shown in Table 2. From this, we see that all the regression coefficients except *spring* and *summer* are statistically different from zero. However, intuitively, we still retain *spring* and *summer* in the model; removing this might give misleading results.

The coefficient for *FE* was –0.2411, which implies that with a unit increase in *FE*, there will be a 24.1% decrease in *LNO3*. The coefficient for *fall* was 0.4717 which implies that there will be 47.1% increase in *LNO3* in fall when compared to winter. The coefficient of *D2* was –2.0313 which implies that there will be a 20.3% decrease in *LNO3* when the distance changes from less than 150 m (*D1*) to a distance between 150 m and 300 m (*D2*). The coefficient of *D3* was –1.8524 which implies that there will be a 18.5% decrease in *LNO3* when the distance changes from less than 150 m (*D1*) to a distance greater than 300 m (*D3*). From the scatter plot of the residual vs predicted *LNO3* (Fig 2), we see that there is no special pattern or trend in this distribution. They are randomly distributed about zero.

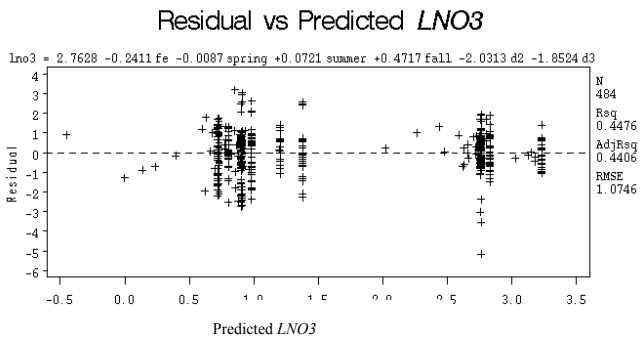

FIG 2

From the plot of the predicted *LNO3* vs actual *LNO3* (Fig 3), we can see that there are two separate clusters of points. This pattern is because we have used the dummy variables for season and distance. Also, a paired t-test was done to check if any statistically significant difference existed between the actual *LNO3* and predicted *LNO3* for the training data set. The probability of the calculated t-value exceeding the tabulated t-value (for α = 0.05 and df=477) was 1. Thus, no statistically significant difference exists between the actual *LNO3* and predicted *LNO3* for the training data set.

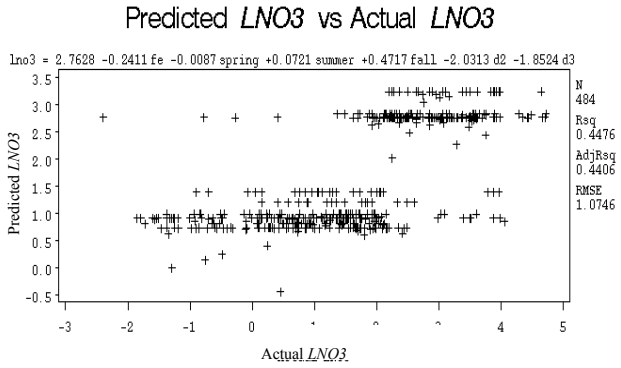

FIG 3

Model validation using validation data set

For model validation purposes, 143 observations (23% of the data set) were used. The mean sum of squared errors (mean SSE) was calculated for the validation data set and then compared with the mean SSE of the training data set. The mean SSE for both the training and validation data sets are shown in Table 3. The mean SSE for the validation data set is less than that of the training data set, implying that the predictions are better for the validation data set than the training data set. From the plot of predicted *LNO3* vs actual *LNO3* for the validation data set is shown in Fig 4, we can see that there are two separate clusters of points. This pattern is because we have used the dummy variables for season and distance. A paired t-test was done to check if any statistically significant

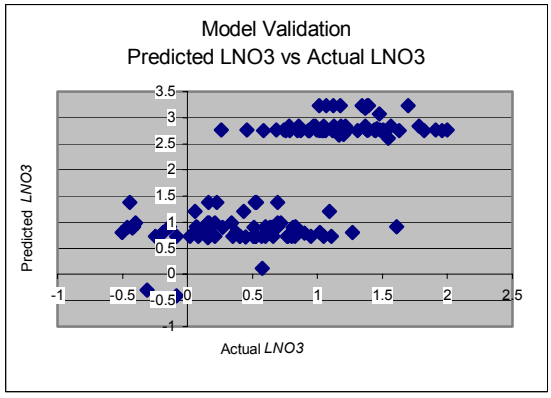

FIG 4

difference existed between the actual *LNO3* and predicted *LNO3* for the validation data set. The probability of the calculated t-value exceeding the tabulated t-value (for α = 0.05 and df=136) was 1. Thus, no statistically significant difference exists between the actual *LNO3* and predicted *LNO3* for the validation data set.

Neural Networks

Neural networks results for training data set

The same 484 observations used for building the regression model was used to train a neural network. The R-square value of the training data set was 0.4470. A paired t-test was done to check if any statistically significant difference existed between the actual *LNO3* and predicted *LNO3* for the training data set. The probability of the calculated t-value exceeding the tabulated t-value (for α = 0.05 and df=477) was 0.3179. Thus, no statistically significant difference existed between the actual *LNO3* and predicted *LNO3* for the training data set.

The importance of each input is shown in Fig 5. *D3* was expected to have greater importance than *D2* because nitrate concentration would decrease with increasing distance. However, from Fig 5, we see that *D2* (Zone 2) has the highest importance. This result might be due to the fact that although the nitrate concentration decreases as the distance increases, other factors such as rainfall and agricultural activities may tend to increase the nitrate concentration in zone 3. Heavy rainfall causes nitrates to be leached. It takes a long time (1.5 to 2 years) for the nitrates to reach the groundwater depth where the sampling wells were located. From Fig 5 we see that, among the seasons, fall is significant. This result is because, in fall two years before the data was collected, there was a considerable amount of rainfall and this rainwater has taken around two years to reach the ground

water depth, which happens to be the fall of the year during which this data was collected. Also the whole of the previous two years have been wet and there was a total rainfall of 264.55 cm in the previous two years [3].

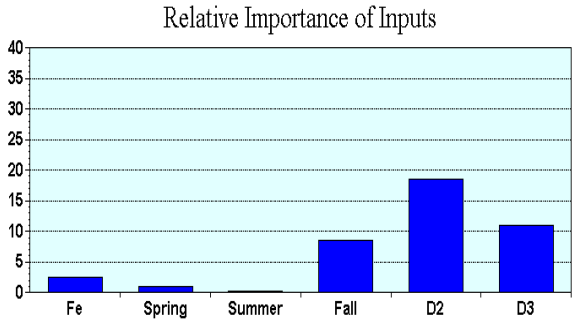

FIG 5

Model validation using validation data set

The same 143 observations used for validating the regression model was used here. The mean SSE for both the training and validation data sets are shown in Table 3 and they are used to compare the predictions of both data sets. From Table 3, we see that the mean SSE is less for the validation data set than that of the training data set. This result implies that the predictions are better for the validation data set than for that of the training data set. A paired t-test was done to check if any statistically significant difference existed between the actual *LNO3* and the predicted *LNO3* for the validation data set. The probability of the calculated t-value exceeding the tabulated t-value (for α = 0.05 and df=136) was 0.8539. Thus, no statistically significant difference exists between the actual *LNO3* and predicted *LNO3* for the validation data set.

COMPARISON OF REGRESSION AND NEURAL NETWORK TECHNIQUES

In neural networks and regression, among the seasons, only *fall* has a significant effect on *LNO3*. This result is because, in fall two years before the data was collected, there was a considerable amount of rainfall and this rainwater has taken around two years to reach the ground water depth, which happens to be the fall of the year during which this data was collected. Also, *D2* and *D3* are highly significant and *D2* is more negative than *D3* in neural networks and regression. This result might be due to the fact that although the nitrate concentration decreases as the distance increases, other factors such as rainfall and agricultural activities may tend to increase the nitrate concentration in zone 3. In Neural networks, *FE* has very little effect on *LNO3,* in contrast to regression, where *FE* has a significant effect on LNO3. *Spring* and *summer* have practically no effect on *LNO3* in neural networks and regression. The mean SSE is less for neural networks than for regression for both training and validation data sets (Table 3). Also, the R-square values from regression and neural networks are close to each other for both data sets, indicating that both techniques perform equally well. The Mean Absolute Error (MAE) was calculated to compare the predictions of regression and neural networks. For both data sets, MAE is less for neural networks than for regression, implying that the predictions of neural networks are better.

Two paired t-tests were done to check if any statistically significant difference existed in the predictions from regression and neural networks for both the training and validation data sets. For both the training and validation data sets, the probability of the calculated t-value exceeding the tabulated t-value (for α = 0.05 and df=477 and 136) was 1. Thus, no statistically significant difference exists in the predictions of regression and neural networks for both the data sets. These tests shows that both regression and neural networks are performing equally well. Regression over predicts by 3% for 90% of the observations in the validation data set. Predictions from neural networks lies within ± 5% of the actual output for 90% of the observations in the validation data set.

CONCLUSION

Two models have been built using regression and neural networks to predict the nitrate concentration in ground water as a function of iron concentration in ground water, season and distance of the well from a poultry house. Results from both techniques were comparable and show that the distance of the well from a poultry house has a significant effect on nitrate concentration in ground water. All the tests showed that there was no statistically significant difference between the predicted LNO3 and actual LNO3. Though the statistics from neural networks were better than that of the statistics from regression, neural networks under predicted LNO3. But regression over predicts LNO3. It is safer to use the LNO3 predicted by regression, as it might be a good margin of safety. As a further extension, some new methods can be employed to remove heteroskedasticity from the model.

REFERENCES

[1] Altendrof, C.T., Elliot, R.L., Stevens, E.W. and Stone, M.L. *Development and Validation of Neural Networks Model for Soil Water Content Prediction With Comparison to Regression Techniques*. ASAE, 1999, Vol 42 (3): 691-699.

[2] Flay, R.B. (2001). *Modeling Nitrates and Phosphates in Agricultural Watersheds with the Soil and Water Assessment Tool.* Available online at **http://cook.berkeley.edu/pubs/swat/SWAT_Review.PDF**

[3] Leathers, D.J. (2002). *Delaware Climate Data* available online at **http://www.udel.edu/leathers/declim.html**

[4] Mason, E., Neiber, K.W.J., Misra, D. and Nguyen, H.V. *Assessment of the Impact of Non-Point Source Pollution on Groundwater Quality*, 1996. Available online at **http://www.arc.umn.edu/education/archives/1996/mason**

[5] Ray, C. and Klindworth, K.K. *Neural Networks for Agrichemical Vulnerability Assessment of Rural Private Wells.* Journal of Hydrologic Engineering, 2000, Vol 5 No.2 April 2000.

[6] Ritter, W.F. and Chirnside, A.E.M. *Groundwater Quality in Selected Areas of Kent and Sussex Counties, Delaware.* Report prepared for the Division of Environmental Quality, DNREC, Dover, Delaware,1982, 19901-229 pages.

[7] Robertson, F.N. *The Quality and Potential Problems of the Groundwater in Coastal Sussex County, Delaware.* Report prepared by the Water Resources Center, University of Delaware, under the sponsorship of Sussex County Council, Coastal Sussex County 208 Program, Contract no. WRC-208-06-76, 1977.

[8] Yang, C.C., Prasher, S.O., Wang, S., Kim, S.H., Tan, C.S and Drury, C. *Simulation of Nitrate-N Pollution in Southern Ontario with DRAINMOD-N.* NABEC, 2002, Paper # 02-028.

Table 1
Regression Output for training data set

Independent Variables	*FE, spring, summer fall, D2, D3*
Dependent Variable	*LNO3*
No. of observations	484
R^2 value	0.4476
F-statistic for overall significance	64.41
p-value for F-statistic (α=0.05)	< 0.0001
χ^2 statistic for heteroskedasticity	45.21
p-value for χ^2 statistic for heteroskedasticity	0.0006
VIF for all the independent variables	< 10
Condition index for all the independent variables	< 30

Table 2
Heteroskedasticity Robust t-values

Variable	Coefficient	Standard Error	Calculated t-value	p-value
FE	-0.24113	0.10811	-2.23041	0.0262 *
Spring	-0.00875	0.14219	-0.06154	0.9510
Summer	0.07207	0.14026	0.51383	0.6076
Fall	0.47173	0.15382	3.06677	0.0023 *
D2	-2.03128	0.11100	-18.29982	< 0.0001 *
D3	-1.85238	0.11589	-15.98395	< 0.0001 *

* Significant at 5% level

Table 3

Comparison of regression and neural networks

	Regression	Neural Networks
R^2 value		
1) Training data set	0.4476	0.4470
2) Validation data set	0.5638	0.5541
Mean Sum of Squared Errors		
1) Training data set	1.1547	0.2176
2) Validation data set	1.0062	0.1651
Mean Absolute Error (MAE)		
1) Training data set	0.9755	0.3536
2) Validation data set	1.1132	0.3115

The Department of Food and Resource Economics
College of Agriculture and Natural Resources
University of Delaware

The Department of Food and Resource Economics carries on an extensive and coordinated program of teaching, organized research, and public service in a wide variety of the following professional subject matter areas:

Subject Matter Areas

Agricultural Finance
Agricultural Policy and Public Programs
Environmental and Resource Economics
Food and Agribusiness Management
Food and Fiber Marketing
International Agricultural Trade
Natural Resource Management
Operations Research and Decision Analysis
Price and Demand Analysis
Rural and Community Development
Statistical Analysis and Research Methods

The department's research in these areas is part of the organized research program of the Delaware Agricultural Experiment Station, College of Agriculture and Natural Resources. Much of the research is in cooperation with industry partners, other state research stations, the USDA, and other State and Federal agencies. The combination of teaching, research, and service provides an efficient, effective, and productive use of resources invested in higher education and service to the public. Emphasis in research is on solving practical problems important to various segments of the economy.

The department's coordinated teaching, research, and service program provides professional training careers in a wide variety of occupations in the food and agribusiness industry, financial institutions, and government service. Departmental course work is supplemented by courses in other disciplines, particularly in the College of Agriculture and Natural Resources and the College of Business and Economics. Academic programs lead to degrees at two levels: Bachelor of Science and Masters of Science. Course work in all curricula provides knowledge of tools and techniques useful for decision making. Emphasis in the undergraduate program centers on developing the student's managerial ability through three different areas, Food and Agricultural Business Management, Natural Resource Management, and Agricultural Economics. The graduate program builds on the undergraduate background, strengthening basic knowledge and adding more sophisticated analytical skills and business capabilities. The department also cooperates in the offering of an MS and Ph.D. degrees in the inter disciplinary Operations Research Program. In addition, a Ph.D. degree is offered in cooperation with the Department of Economics.

For further information write to: Dr. Thomas W. Ilvento, Chair
Department of Food and Resource Economics
University of Delaware
Newark, DE 19717-1303

FREC Research Reports are published as a service to Delaware's Food and Agribusiness Community by the Department of Food and Resource Economics, College of Agriculture and Natural Resources of the University of Delaware.