# Predicting Fish Ecological As Indicator of River Pollution Using Decision Tree Technique

# Predicting Fish Ecological As Indicator of River Pollution Using Decision Tree Technique

**Che-Yu Hsu [1], Sheng-Jung Ou [2] & Wei-Fan Hsieh[3]**

1.Adjunct Assistant Professor, Department of Landscape and Urban Design, Chaoyang University of Technology, 168, Jifeng E. Rd., Wufeng District, Taichung, 41349 Taiwan
2.Professor, Department of Landscape and Urban Design, Chaoyang University of Technology, 168, Jifeng E. Rd., Wufeng District, Taichung, 41349 Taiwan
3. Assistant Professor, Department of Landscape and Urban Design, Chaoyang University of Technology, 168, Jifeng E. Rd., Wufeng District, Taichung, 41349 Taiwan

sjou@nchu.edu.tw, tony428tw@hotmail.com (Corresponding author).

**Abstract.** The Goal of the research is to introducing the principle of Decision Tree that is being used to forecast river pollution, it provides a new method to evaluate the river pollution based on its water quality. We collected monthly monitoring data of water quality from Dezikou River basin of Yilan County, and the data of fish ecology obtained from ecological survey and report, in which to build a water quality and ecology resources database through an actual field investigation. By using data mining software, IBM SPSS Modeler 14.1's decision tree, conducting the River Pollution Index. Shannon-Weaver diversity, Pielou's Evenness Index, Margalef's Species Richness Index, Fish Tolerant Index and Simpson's Index of Diversity′ s classification and prediction, to build a model for river pollution prediction, and to compare this with the Multiple Logistic Regression Analysis. The results showed that the model for river pollution prediction built under the Decision Tree can obtain a better forecast result. The following are the accurate rates of Decision Tree: 88% for CART, 90% for CHAID, 91.67% for C5.0, and 86.11% for Multiple Logistic Regression Analysis. Therefore, the Decision Tree's algorithm shows a better result in forecasting than the Multiple Logistic Regression Analysis.

## 1. Introduction

Rivers serve many societal functions and belong to the most intensively human influenced ecosystems worldwide [1]. By monitoring the fish in the river, it not only reflects the water quality of the river by the presence of the migratory fish and the fish biology indicators, but also it can further reflect the changes in the environment underwater by RPI. Fishes have a number of advantages as biological-integrity indicators of a watershed system [2]. Some species of fish are very sensitive to changes in water chemistry, such as pH or dissolved oxygen, which may be caused by pollutants. After the monitoring outcome of the success of restructure the river in Singapore and Guyahoga River in America, it is not just to pay attention to the improvement of the water quality, but also it is to concern the effectiveness of the environmental ecology in rivers, meaning fish is one of the main targets for monitoring. This research applies the Shannon-Weaver diversity, Pielou's Evenness Index,

Margalef's Species Richness Index, Fish Tolerant Index and Simpson's Index of Diversity to evaluate the analysis. IBM SPSS Modeler is a data mining and text analytics software application built by IBM. Classification and regression trees are ideally suited for the analysis of complex ecological data [3]. An ensemble algorithm of data mining decision tree (DT)-based CHi-squared Automatic Interaction Detection (CHAID) is widely used for prediction analysis in variety of applications. [4]. The main goals of this study were to introduce the principle of Decision Tree that is being used to forecast river pollution, it provides a new method to evaluate the river pollution based on its water quality. The model for river pollution prediction built under the Decision Tree can obtain a better forecast result.

## 2. Research method

### *2.1 Study area*

Dezikou River is on the northern side of Yilan County, meanders for 19.3 km through a basin area of 113.17 square km, Toucheng Town. Dezikou River is a typical urban river. The importance of the area is derived from the presence of different habitat types and the numerous flora and fauna species. However, human activities in recent years, such as land use change and urbanization, deteriorated the water quality and the ecosystem of Dezikou River. At present, the species that found are most tolerant species, but its number is still not large enough.

### *2.2 The Decision Tree's algorithm*

Decision tree learning uses a decision tree (as a predictive model). It is one of the predictive modelling approaches used in statistics, data mining and machine learning. At present, academia and practitioners use Decision Tree's algorithm including CART, CHAID, and C5.0 etc, due to its research target's variation to be that of ordering type of variation, it cannot use QUEST to undergo analysis.

### *2.3 Data Resource*

*2.3.1 Water quality data information of investigation* This research collects relevant monitoring data information on water quality monthly from the Yuan EPA National Environmental Quality Monitoring Information Network providing information which it has collected in a total five year from 2009 to 2013, in which the information was obtained from Yilan County Water Pollution Control Inspection and basin management plan. Every monthly water quality monitoring data builds relevant data, and builds a water quality database [5] [6].

*2.3.2 Ecological survey* Survey collected fish ecology information between 2009 and 2013 from Yilan County Council EPA It records fish' category, species, individual number, feeding characteristic, movement characteristic, pollution resistance characteristic, habitat characteristic and pollution resistance factor to build fish ecological database by its actual field investigation. Furthermore, its ecology factor is calculated through Shannon-Weaver diversity, Margalef's Species Richness Index, Fish Tolerant Index, and Simpson's Index of Diversity.

## 3. Result & Discussion

The research is to adopt sample database to build a river pollution forecast model, and to evaluate the model.

### *3.1 Classification and Regression Analysis CART Algorithm*

Through CART algorithm, we can find that if the species diversity is lower or extremely low and if there is less or lesser abundant species, then the RPI has 75% chance to be Severely-polluted; if species diversity is lower or extremely low and if the species abundance is low, higher, and extremely high, then the RPI has 96.88% chance to be Moderately-polluted; if there is a higher level of species

diversity, then the RPI has 57.14% chance to be lightly-polluted. If there is an extremely high level of species diversity, then the RPI has 100% chance to be Non (Slightly)-polluted (see Figure 1).

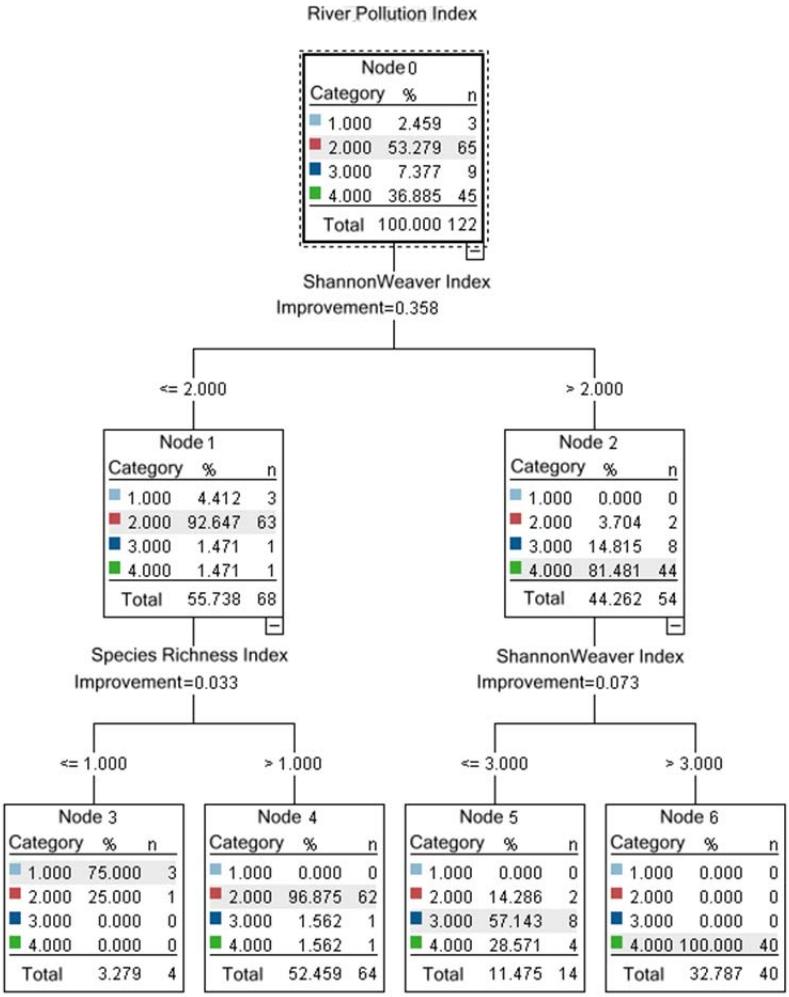


**Figure 1**. Classification and Regression Analysis (CART) Algorithm

### *3.2 Chi-Square Automatic Interaction Detection , (CHAID Algorithm*

Via CHAID, we can find that if it there are more fish species with severely-polluted indicator, then the RPI has 50% chance to be Severely-polluted; if there are more fish species with moderately polluted indicators and if there is a low or lower level of low fish diversity, then the RPI has 97.62% chance to be Moderately-polluted; if there are more fish species with lightly-moderated indicators, then RPI has 54.5% chance to be lightly-polluted. If there are more fish species with non- (slightly-) polluted indicators, then then the RPI has 98.44% chance to be non- (slightly)-polluted (see Figure 2).

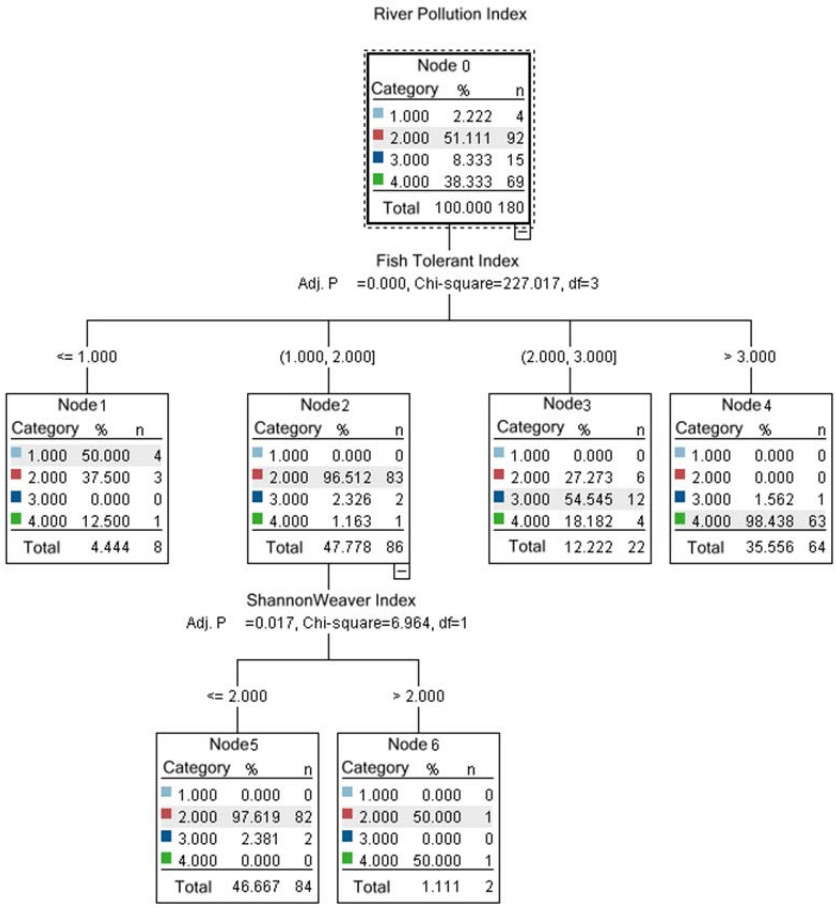


**Figure 2**. Chi-Square Automatic Interaction Detection , CHAID Algorithm

### *3.3 Decision Tree C5.0 Algorithm*

By C5.0 algorithm, we find that if it is severely, moderately and lightly polluted has more fish species and the level of diversity among fish species is low / lower and not abundant fish species then RPI has 66.67% rate to be Severely-polluted; if severely, moderately, and lightly polluted has more fish species and low / lower level of fish diversity among fish species and not abundant fish species , then RPI has 95.56% rate to be Moderately-polluted; if severely, moderately and lightly polluted has more fish species and high / higher level of diversity among fish, then RPI has 60% rate to be lightly-polluted; if non(slightly) polluted has more fish species then RPI has 98.44% rate to be non(slightly)-polluted (see Figure 3).

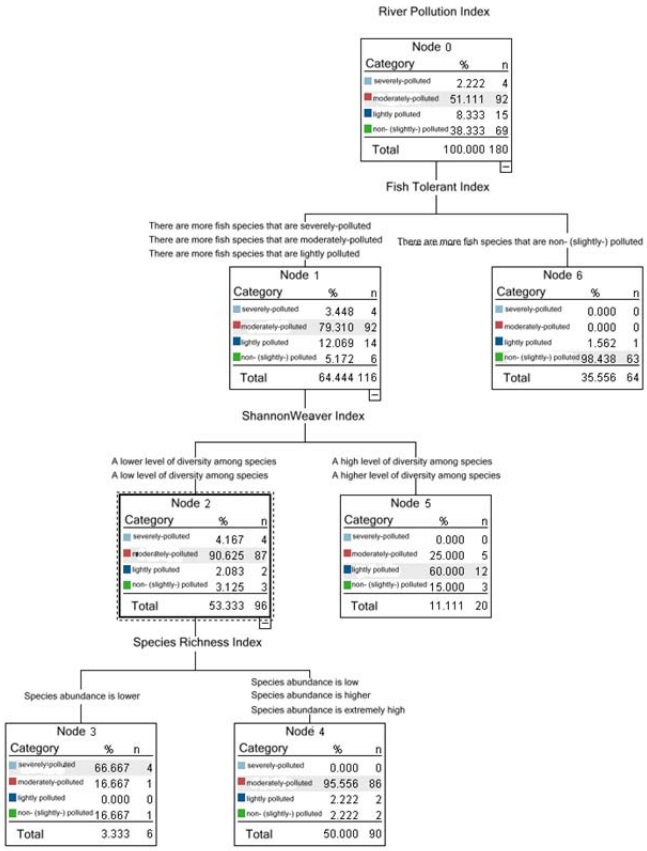


**Figure 3**. Decision tree C5.0 diagram

### *3.4 Multiple logistic regression analysis – building river pollution forecast model*

First test to see how match the variables are within the multivariate logistic regression models. Its test result will show model's P = 0.000, so the variables (independent variable) overall can explain its dependent variables to have significance of the statistics (see Table 1). Their classification and actual categories the accurate rate is 86.11%.

**Table 1.** Multiple logistic regression models all samples goodness-of-fit test analysis

| Method | -2 Log Likelihood | Chi-Square | Sig. |
|---|---|---|---|
| Forward Stepwise | 768.579 | 755.763 | 0.000*** |

*P < 0.05; **P < 0.01; ***P < 0.001

## 4 Conclusion

This research has tried to conduct RPI based on Decision Tree (CART, C5.0, CHAID ), Shannon-Weaver diversity, Pielou's Evenness Index, Margalef's Species Richness Index, Fish Tolerant Index and Simpson's Index of Diversity to categorize and reasoning/evaluate, to obtain the river pollution data of Dezikou River basin and build an ecology forecast model for obtaining a better forecast effect. On evaluation of model side, the accurate rate of forecast is that CART is 88.33%, CHAID is 90%, C5.0 is 91.67% and multiple logistic regression analysis is 86.11%.

Although Decision Tree C5.0 algorithm has the most accurate rate, CHAID algorithm can use only two sets of variables of Fish Tolerant Index and Shannon-Weaver Index to achieve the same result with that of Decision Tree C5.0, which includes three sets of variables; Fish Tolerant Index, Shannon–Weaver Index, Margalef's Species Richness Index. Moreover, Decision Tree model is simpler. Therefore, a simple and easy to understand the Decision Tree's algorithm has better forecast outcome than traditional multiple logistic regression analysis (see Table 2).

**Table 2.** Decision Tree **a**nd Multiple logistic regression analysis to predict the accuracy

| | | | |
|---|---|---|---|
| Classification and Regression Analysis | Correct | 159 | 88.33% |
| | Error | 21 | 11.67% |
| | Total | 180 | |
| Chi-Square Automatic Interaction Detection | Correct | 162 | 90% |
| | Error | 18 | 10% |
| | Total | 180 | |
| Decision Tree C5.0 | Correct | 165 | 91.67% |
| | Error | 15 | 8.33% |
| | Total | 180 | |
| Multiple logistic regression analysis | Correct | 155 | 86.11 |
| | Error | 25 | 13.89 |
| | Total | 180 | |

This research's purpose is to provide a simple method to forecast water quality of rivers, based on the survey of the current fish species to substitute complex water quality examination. Due to sample and subject condition limitation, we cannot conduct sample division testing and verify further. One can apply this method to proceed to more specific evaluation when all the data has been provided.

## References

[1] Naftulin IS, Rebrova OY 2010 Application of C&RT, CHAID, C4.5 and WizWhy algorithms for Omar F. Althuwaynee, Biswajeet Pradhan, Hyuck-Jin Park & Jung Hyun Lee, Landslides, 2014, 11(6) 1063–1078..stroke type diagnosis, Artificial Intelligence and Soft Computing. Springer, 651–656.

[2] Plafkin, JL, MT. Barbour, KD. Porter, SK. Gross, & R M. Hughes. 1989. Rapid bioassessmentprotocols for use in streams and rivers: Benthic macroinvertebrates and Fish EPA/440/4-89/001. Washington, DC: U.S. Environmental Protection Agency.

[3] De'Ath G and Fabricius E 2000 Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81(11): 3178-3192.

[4] Omar F. Althuwaynee, Biswajeet Pradhan, Hyuck-Jin Park & Jung Hyun Lee, Landslides, 2014, 11(6) 1063–1078..

[5] Environmental Protection Administration Executive Yuan, R.O.C. (Taiwan) 2009∽ 2013 River Water Quality History Data, Environmental Water Quality Information,

[6] Environmental Protection Bureau Yilan County, River ecology species survey, http://works.ilepb.gov.tw/01001_W_01/p3_2_1.html#river1