Classifying with Fuzzy Chi-Square Test: The case of Invasive Species

Vardis-Dimitris Anezakis[1, a)] Konstantinos Demertzis[1, b)] Lazaros Iliadis[2, c)]

[1]*Democritus University of Thrace, Lab of Forest Informatics, 193 Pandazidou st., 68200 N Orestiada, Greece*
[2]*Democritus University of Thrace, School of Engineering, Department of Civil Engineering, University Campus, Kimmeria, 67100, Xanthi, Greece*

[a)] danezaki@fmenr.duth.gr
[b)] kdemertz@fmenr.duth.gr
[c)] liliadis@civil.duth.gr

Abstract. Given that the Chi-Square Test imparts a binary correlation between the variables examined and it does not offer the exact degree of dependence or independence which is always a major issue, this research proposes an innovative method of yielding these values with precision and accuracy. More specifically, this paper introduces and applies the Fuzzy Chi-Square Test, which fuzzifies the Chi-Square Test's p-value by employing linguistics like Low, Medium, High in order to incorporate the level of dependence or independence of the variables examined. In this way it renders non-rigid inference mechanisms, easier understanding and ability to model functions of arbitrary complexity.

Keywords: Chi-Square Test, Fuzzy Chi-Square Test, Fuzzy Theory, Fuzzy Set, Depended Variables, Climate Change, Invasive Species

1. INTRODUCTION

The Chi-Squared hypothesis-testing is a non-parametric statistical test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true. The null hypothesis H_0 usually refers to a general statement or default position that there is no relationship between two measured phenomena, or no difference among groups. The H_0 is assumed to be true until evidence suggest otherwise [1], [2]. For the H_0 hypothesis, the critical values for the test statistic X^2 are estimated by the X^2 distribution after considering the degrees of freedom. If the result of the test statistic is less than the value of the Chi-Square distribution, then we accept H_0 otherwise we reject it. This paper proposes a new method of decision making regarding the result of the statistical chi-square test.

The result of the fuzzy chi-square test no longer resides in the decision that the variables examined are dependent or independent of each other but determines the degree of dependence of the two variables with membership degrees using fuzzy sets and fuzzification procedures.

1.1. The Problem of Invasive Species

The potential impacts of climate change are evident at various levels of the biological organization and particularly in the disturbances observed in biodiversity, organism extinction and appearance of invasive species [3]. Invasive species enter new foreign habitats and they can stifle natural flora or fauna harming the environmental balance. The existence of invasive species in different countries is directly related to climate change and in particular, to major changes in meteorological conditions in the areas under consideration.

For example, with the Fuzzy Chi-Square Test, it is possible to identify the exact degree of membership of a statistically significant relationship between an invasive species (IN_SP) and the climate conditions of its habitat. The species selected as for the development and presentation of the proposed model are presented in Table 1 below.

TABLE1. Number of recorded instances of four invasive species in four countries

	Aedes albopictus	**Anguillicola crassus**	**Carassius auratus**	**Biden spilosa**
Italy	14	2	20	11
Greece	1	29	11	8
France	2	41	18	7
Malaysia	22	38	21	13

1.2. Literature Review

Gil et al., [4] mentioned that if the hypothetical distribution involves unknown parameters the extension of the chi-square goodness of fit test requires the estimation of those parameters from fuzzy data. Taheri et al., [5] extended the classical methods of analysis of a two-way contingency table to the fuzzy environment. The α-cuts approach was used to extend the usual concepts of the test statistic, resulting in the use of fuzzy test statistic and fuzzy p-values. Huang [6] used fuzzy statistical analysis and chi-square test for the significant differences between ethnic majority and ethnic minority students. Taheri et al., [7] tested the hypothesis of independence using a novel method of decision making based on a concept of fuzzy p-value. Grzegorzewski and Szymanowski [8] proposed a method for constructing a generalized version of the chi-square test of homogeneity which allows fuzzy data. The above review clearly shows that there are efforts in the literature towards the fuzzification of the statistical Chi-square test. However to the best of our knowledge, such hybrid approaches have not been used in environmental risk modeling and more specifically for the case of invasive species.

2. ALGORITHM OF THE PROPOSED FUZZY CHI-SQUARE TEST

The algorithmic process of the proposed Fuzzy Chi-Square Test includes the following eight stages:

Step 1: Fuzzification of the values of the examined variables and assignment to four Risk Linguistics namely: *Low, Medium, High, Extreme* by using membership degrees based on a properly designed Mamdani Fuzzy Inference System (FIS). The number of Linguistics employed is usually determined by the range of variables as well as by statistical analysis using dispersion measures.

Step 2: Record the number of observations assigned to each Linguistic and enter them in a table of rXc dimensions

TABLE2. Record the number of Linguistic observations

Variables A	**Low Variable B (LVB)**	**Medium Variable B (MVB)**	**High Variable B (HVB)**	**Extreme Variable B (EVB)**	**Total**
Low Variable (LVA)	LVA-LVB	LVA-MVB	LVA-HVB	LVA-EVB	Sum LVA
Medium Variable (MVA)	MVA-LVB	MVA-MVB	MVA-HVB	MVA-EVB	Sum MVA
High Variable (HVA)	HVA-LVB	HVA-MVB	HVA-HVB	HVA-EVB	Sum HVA
Extreme Variable (EVA)	EVA-LVB	EVA-MVB	EVA-HVB	EVA-EVB	Sum EVA
Total	Sum LVB	Sum MVB	Sum HVB	Sum EVB	Sum

Step 3: Calculate the expected values in all cells of the table to find the test statistic using the mathematical formulae of the chi-square test.

Step 4: Calculate degrees of freedom based on the number of Linguistics of the variables examined, or alternatively on the dimensions rXc of the table. The degrees of freedom are calculated based in equation 1.

$$df = (r-1)\times(c-1) \tag{1}$$

The value of the degrees of freedom for the selected confidence interval (C_INT) is calculated from the statistical Chi-Square Test.

Step 5: We select a C_INT equal to 0.95 and a significance level $\alpha = 0.05$ for the fuzzification of the p-value.

Step 6: Comparison of the critical p-value in the confidence interval 0.95 (for the calculated degrees of freedom) with the selected significance level a=0.05. The p-value a=0.05 value is considered as critical for the dependence or independence of the variables. Variable dependence is determined with a $p - value < a$ and the independence with $p - value > a$. The estimated p-values contain an error probability in the interval $[0-1]$.

Step 7: Construction of FIS-Mamdani membership functions for p-values<a in the interval [0, 0.049999] to represent the dependence Linguistics of the two variables. Strong dependence with membership value (MV) equal to 1 is assigned to values more distant from the 0.049999 value approaching the value 0, having the smallest possible error. On the other hand, p-values close to 0.049999 exhibit Lower dependence and their MV will be close to 0. Totally 50,000 values were fuzzified all included in the interval $[0.000000 \; - 49999 \;] \times 10^{-6}$. Every error value was multiplied by 10 and it is raised to the power of -6 $\left(p - value \times 10^{-6}\right)$. In table 3 we present the Dependency MV of chosen p-values and their classification in dependency Linguistics with the corresponding MV. P-values in the interval $[0 - 0.012499 \;]$ belong to the High Dependency Linguistic, whereas p-values in the interval $[0.012501 \; - 0.037499 \;]$ belong to the Moderate Dependency level. P-Values in the interval $[0.0375 \; - 0.049999 \;]$belong to the Low Dependency Linguistic.

TABLE3. Presentation of the MV of the Dependency Linguistics in the interval [0-0.049999]

P-Value	**Linguistics**	**High Dependence Degree of Membership**	**Medium Dependence Degree of Membership**	**Low Dependence Degree of Membership**
0	High	1	0	0
0.00001	High	0.99945	0	0
0.012499	High	0.37505	0.37495	0
0.012501	Medium	0.37495	0.37505	0
0.03	Medium	0	0.75	0
0.035	Medium	0	0.5	0.2500125
0.037499	Medium	0	0.37505	0.374968
0.0375	Low	0	0.375	0.3750187
0.049999	Low	0	0	1

Step 8: Defuzzifying the $p - value > a$ in the interval $[0.050001 \; - 1]$ by constructing membership functions to calculate the MV of the independent variables and their inclusion in classes of independence Linguistics. P-values close to 1, (the largest possible error) indicate that the variables will be independent with a MV equal to 1 (strongly independent) whereas for p-values close to 0.050001 MV will be close to zero and the independence of the variables will decrease (Low independence). We have fuzzified 950,000 $p - values$ (the values that are independent in the interval $[50001 \; - 1000000 \;] \times 10^{-6}$. P-values in the interval $[0.050001 \; - 0.287550 \;]$ belong to the Low Independence Linguistic, whereas p-values in the interval $[0.287551 \; - 0.762518 \;]$ are classified as Moderate Independent. Finally, p-values in the interval $[0.762519 \; - 1]$belong to the High Independence class.

TABLE4. Membership Values of Independence Linguistics in the interval [0.050001-1]

P-Value	**Linguistics**	**Low Independence Degree of Membership**	**Medium Independence Degree of Membership**	**High Independence Degree of Membership**
0.050001	Low	1	0	0
0.05001	Low	0.99997	0	0
0.287550	Low	0.374871	0.374868	0
0.287551	Medium	0.374868	0.374871	0
0.4301	Medium	0	0.75	0
0.71505	Medium	0	0.5	0.25013157
0.762518	Medium	0	0.375051	0.375047
0.762519	High	0	0.375048	0.375050
1	High	0	0	1

The same algorithmic process can be applied to other confidence intervals by modifying the boundaries of the triangular functions based on the value of the chosen significance level a.

3. RESULTS AND COMPARATIVE ANALYSIS

The above approach is applied in the case of four invasive species in four countries as shown in Table 1. The species registration number consists of crisp values which at this stage have not been fuzzified by proper Linguistics. However, considering variables of different nature such as meteorological or atmospheric pollution, the values of the variables could be fuzzified in risk Linguistics (low, medium, high, extreme) as recorded in Table 2. The test statistic is estimated by examining each row r and each column c of Table 1. Initially, the expected values are estimated for every element of Table 1.

$$P\left(R_{i,}C_{j}\right)=\left(\frac{c_{i}totals}{Rtotals}\right)\times\left(\frac{R_{j}totals}{Rtotals}\right)\times\left(Rtotals\right) \tag{2}$$

After having estimated the expected values of each element of Table 1, we calculate the fraction of the differences between observed and expected values, divided by the expected values raised to the second power. The sum of these fractions (of all elements of the table) is the statistical test and in our case it is equal to 56.29.

$$P\left(R_{i,}C_{j}\right)=\left(\frac{Observed\ \ - Expected}{Expected}\right)^{2} \tag{3}$$

The Degrees of Freedom (DOF) are estimated based on the table's dimensions e.g. $df\ =(4-1)\times(4-1)=9$. From the Chi-Square Test distribution, it is estimated that the DOF df_9 for the confidence interval 0.95 with significance level 0.05 equals 16.92. The test statistic equals 56.29 and it is higher than the value that emerged from the chi-square test whereas the $p-value$ is <0.00001. Based on the results a small $p-value$ and a high test-statistic, declare a high dependence between the recorded instances in the four involved countries.

The algorithmic process of calculating the test-statistic and the DOF is the same in both methods. On the other hand, the fuzzy chi-square test is differentiated in the interpretation of the $p-value$. The $p-value\langle 0.00001$ is interpreted differently in the fuzzy chi-square test as it has been fuzzified in the interval $[0-0.049999\ ]$ defining the Dependence Linguistic and the MV of the Linguistic. Thus, $p-value\langle 0.00001$ in the fuzzy chi-square test belong to the High Dependence Linguistic with MV equal to 1 proving the absolute high Dependence between the appearance of the four invasive species in the countries under consideration. This dependence is due to the meteorological topographical and vegetative conditions of the countries which are identical to the habitat of the species. The p-value Fuzzification process differentiates the decision-making as every calculated value is classified to a proper Dependence Linguistic with an estimated MV which shows the strength of this relation.

4. CONCLUSIONS

This paper proposes a new modeling approach, based on the fuzzy chi-square test in which the degree of dependence or independence is precisely defined, using fuzzy logic and fuzzy intelligence. In addition, the p-value Fuzzification process provides Linguistic categorization of the degree of Dependence or Independence, contributing to the extraction of useful conclusions during the modeling process. This method is very effective and flexible and it can be globally applied in real world problems.

REFERENCES

1. G.W. Corder and D.I. Foreman, Nonparametric Statistics: A Step-by-Step Approach 2nd Edition, edited by Wiley& Sons Inc., (New York, 2014) p.288 ISBN 978-1-118-84031-3.
2. P.E. Greenwood and M.S. Nikulin, A guide to chi-squared testing, edited by Wiley & Sons Inc., (New York, 1996) p.304 ISBN 978-0-471-55779-1.
3. F. Rahel, J.D. Olden, Conservation Biology 22(3), 521–533 (2008). doi:10.1111/j.1523-1739.2008.00950.x
4. M.A. Gil, N. Corral, P. Gil, Journal of Statistical Planning and Inference 19, 95-115 (1988).
5. S.M. Taheri, G. Hesamian, R. Viertl, Communications in Statistics-Theory and Methods 45(20), 5906-5917 (2016). doi:10.1080/03610926.2014.953688
6. H.M. Huang, International Journal of Social Science and Humanity 2(2), 151-155(2012).
   doi: 10.7763/IJSSH.2012.V2.86
7. S.M. Taheri, G. Hesamian, Kybernetika 47(1), 110-122 (2011).
8. P. Grzegorzewski, H. Szymanowski, Advances in Intelligent Systems and Computing 315, 151-158 (2015). doi:10.1007/978-3-319-10765-3_18