Original Research Paper Agricultural Science

Examination of Factors Influencing the Population of Turkey Using Chaid Analysis

Senol Celik | Bingol University, Faculty of Agriculture, Department of Animal Science, Biometry and Genetics, Bingol, Turkey

ABSTRACT

In this study, affect of variables such as export, housing sell, health, cinema, sport, number of vehicle, number of animal, environment, number of agriculture machine and tourism were investigated on population of Turkey. This effect was examined by CHAID algorithm. In the study, population was the most significant agent for number and vehicle, followed by health, culture and tourism. Most populous average population was observed for number of vehicle > 527796. As the number of motor vehicles increases, as population have been increased. Besides, as health, culture and tourism services are growing, as population have been increased. Consequently, CHAID method is not effected by outlier and multicollinearity. Thus, it appears beneficial results for such research

KEYWORDS CHAID, regression tree, population

INTRODUCTION

Population is an important indicator for the country. In the world, there are many factors affecting the population. These factors can be industry, trade, tourism, health, transport, education, culture, agriculture etc. Plans and policies can be created for countries or cities identified factors affecting the population. Good planning is made using effective and the best statistical methods. Data mining methods are done in the best way, in this kind of research. CHAID algorithm is one of the data mining algorithms. There are studies in various fields by CHAID algorithm.

The CHAID prediction model of student performance was constructed with seven class predictor variable. Nguyen Thai-Nghe, Andre Busche, and Lars Schmidt-Thieme (Cortez and Silva, 2008) used machine learning techniques to improve the prediction results of academic performances in real case studies.

Yu et al. (2013), established a four-component framework consisting of Data Mining techniques. They proposed a step-by-step data analysis process that starts from problem definition to knowledge discovery.

Ali et al. (2015), were examined use of Exhaustive CHAID in the prediction of body weight presented the best fit in decision tree diagrams, which may provide some advantageous in exhibition of some breed standards of the Harnai sheep.

Eyduran et al. (2016), were reported to predict the fleece weight from some wool characteristics of the sheep reared at Gözlü State Farm located in the central Anatolia region of Turkey with CHAID algorithm.

The target of this study is to examine influence of the some socio-economic variables on population using CHAID algorithm in Turkey.

MATERIAL AND METHOD

The variables used in this study are given in Table 1. Examining the data are the data of 2014, organized by provinces in Turkey. Namely, all data covers 81 provinces in Turkey. Health data from the Ministry of Health, Tourism data Ministry of Culture and Tourism and sports data are taken from the Youth and Sports Ministry. Other data are collected from the TUIK Statistics. Data analysis has been performed using the SPSS 22.0.

Table 1. The data is analyzed

Variables	Explanation
Export	Export Value: Thousand US ($)
Housing	Residential sales statistics
Health	Number of beds
Cinema	The number of cinema attendances
Sport	The number of sports clubs
Vehicle	The number of road motor vehicles
Animal	Beef (culture, hybrid, native), buffalo, sheep (domestic merino), goats (hair, lint)
Environment	Number of subscribers
Culture	Number of public libraries
Agriculture machine	Combines and other equipment and machinery
Tourism	Number of establishment tourism investment licenced and tourism opertaion licenced

Note: Contains information on public and private pre-school education institutions, primary and secondary education institutions and on libraries of private motor vehicle driving courses, private teaching centers and various private courses (Culture: Number of public libraries).

Chi-square Automatic Interaction Detection (CHAID): CHAID as a methodological approach in the literature under different names. CHAID algorithm was introduced by Kass in 1980. But, it has been little used in the segmentation of markets specifically: It has tended to have been applied more to general consumer research (Haughton and Oulabi, 1997; Levin and Zahav, 2001;Riquier et al., 1997).

CHAID is an analysis based on a criterion variable with two or more categories. Researchers to determine the segmentation with respect to that variable and in accordance with the combination of a range of independent variables (predictors) (Chen, 2003; Díaz-Pérez et al., 2005; Legohérel et al., 2015). It was used to explore hotel preferences based on demographic variables of tourists (Chung et al., 2004), and shopping preferences among Japanese tourists to revisit Korea (Kim et al., 2011).

CHAID allows very useful segmentation variables for tourism markets to be included such as gender, age, household income, nationality, season and category of the establishment. Some of these variables are categorical or nominal, others are ordinal or interval-based. Under such circumstances, a tech-

nique that is not subject to the rigidity of the normal distribution and the requirement of ordinal variables will generally be the most appropriate: hence, Chi-square is the ideal statistical method for these cases (Diepen and Franses, 2006).

RESULTS

As shown in Figure 1, the most influential factor on population was number of vehicle (Adj. P=0.000), followed byhealth (Adj. P=0.000), culture (Adj. P=0.000) and tourism (Adj. P=0.000). Node 0, root node, was divided into 4 new child nodes (Nodes 1- 4) according to number of vehicle factor, respectively (Adj. P=0.000, F=21.767, df1=3, and df2=77). The result showed that population increased as number of vehicle increased from Node 1 to Node 4.In the regression tree diagram, terminal nodes were Nodes numbered 4, 6,7, 8, 9, 10, 11, 12, 13, and 14, respectively. Shortly, the nodes reached to adequatelyhomogenous in regression tree diagram.

Node 1, the subgroup of vehicle ≤ 59833, was splittinginto3 child nodes, Nodes5, 6 and 7 in terms of Health factor, respectively (139342,217165 and 368105).Nodes 6 and 7 are terminal nodes in which splitting were choked back at next steps. Defined as a subgroup of 355 ≤ Health < 618, Node 6, gave the average population of 217165 (S=33686.373), approximative. Average population for the subgroup of provinces together with Health > 618, which was also indicated as Node 7, was 368105 (126387.210). Node 5, the subgroup of provinces together with Health ≤ 355 by Node 5 (139342 and S=48646.629). Node 5, the subgroup of provinces vehicle ≤ 59833 and Health ≤ 355 was disunited into new child nodes, Nodes 13 and 14, in terms of Tourism, respectively (Adj. P=0.007, F=21.138, df1=1, and df2=6).

Node 2, the subgroup of 59833<vehicle ≤ 197297, was splitting into 3 child nodes, Nodes 8, 9 and 10 in terms of Culture factor, respectively (Adj. P=0.001, F=15.491, df1=2, df=30). Nodes 8, 9 and 10 are terminal nodes in which splitting were choked back at next steps. Defined as a subgroup of Culture ≤ 193, Node 8, gave the average population of 341274 (S=73467.877), approximative. Average population for the subgroup of provinces with 59833 < Vehicle ≤ 197297 and193 < Culture ≤ 237, which was also indicated as Node 9, was 495774 (S=108188.121). Average population for the subgroup of provinces with 59833 < Vehicle ≤ 197297 and Culture >237, which was also indicated as Node 10, was 791976 (S=327232.200).

Node 3, the subgroup of 197297 < Vehicle ≤ 527796, was splitting into 2 child nodes, Nodes 11 and 12 in terms of Culture factor, respectively (Adj. P=0.003, F=18.650, df1=2, df=14). Nodes 11and 12 are terminal nodes in which splitting were choked back at next steps. Defined as a subgroup of Culture ≤ 406, Node 11, gave the average population of 911951 (S=129883.788), approximative. Average population for the subgroup of provinces with 197297 <Vehicle ≤ 527796 and Culture >406, which was also indicated as Node 12, was 1456199 (S=311102.363). Node 4 which with Vehicle > 527796 is terminal Node (4231490 and

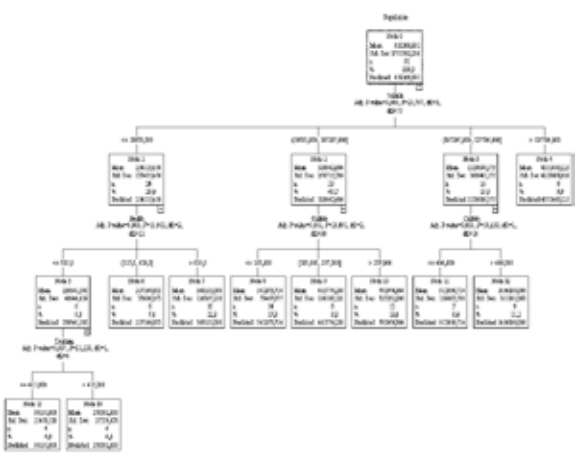

Figure 1. Regression tree diagram for population in Turkey

CONCLUSION

In present research, results of CHAID algorithm reflected that statistically significant effects of vehicle, health, culture and tourism on population of Turkey were determined. The obtained results were expressed as follows:

- The biggest significant factor which statistically affected population of provinces in Turkey were vehicle, other important factors are health, culture and tourism.
- The most populous average population of province was from group among Vehicle > 527996 at Node 4.
- Population of Node 1, which was the group of Vehicle ≤ 59833, was statistical effected by Health factor.
- Population of Node 2, which was the group of 59833<Vehicle ≤ 197297, was statistical effected by Culture factor.
- Population of Node 3, which was the group of 197297 < Vehicle ≤ 527796, was statistical effected by Culture factor.
- Population of Node 5, which was the group of Vehicle ≤ 59833 and Health ≤ 355, was statistical effected by Tourism factor.

REFERENCES

1. Ali, M., E., Eyduran, M. M., Tariq, C., Tirink, F., Abbas, M. A. Bajwa, M. H. Baloch,
2. A. H.Nizamani, A. Waheed, M. A. Awan, S. H. Shah, Z. Ahmad, and S. Jan (2015). Comparison of artificial neural network and decision tree algorithms used for predicting live weight at post weaning period from some biometrical characteristics in Harnai sheep. Pakistan J. Zool. 47:1579-1585.
3. Chen, J. S. 2003. Market segmentation by tourists' sentiments. Annals of Tourism
4. Research, 30(1): 178–193.
5. Chung, K. Y., Oh, S. Y., Kim, S. S., Han, S. Y. 2004. Three Representative Market Segmentation Methodologies for Hotel Guest Room Customers. Tourism Management, 25(4): 429-441.
6. Cortez, P., Silva, A. 2008. Using Data Mining to Predict Secondary School Student Performance, In EURO SIS, A. Brito and J. Teixeira (Eds.), 5-12.
7. Díaz-Pérez, F. M., Bethencourt-Cejas, M., Álvarez-González, J. A. 2005. Thesegmentation of Canary island tourism markets by expenditure: Implication
8. for tourism policy. Tourism Management, 26(6): 961–964.
9. Diepen, M. van, Franses, P. H. 2006. Evaluating chi-squared automatic interaction detection. Information Systems, 31: 814–831.
10. Eyduran, E., Keskin, I., Erturk, Y. E., Dag, B., Tatliyer, A., Tirink, C., Aksahan, R.,
11. Tarıq, M. M. 2016. Prediction of Fleece Weight from Wool Characteristics of
12. Sheep Using Regression Tree Method (Chaid Algorithm). Pakistan Journal of Zoology, 48 (4): 957-960.
13. Haughton, D., Oulabi, S. 1997. Direct marketing modeling with CART and CHAID.
14. Journal of Interactive Marketing, 11(4): 42–53.
15. Kass, G. V. 1980. An Exploratory Technique for Investigating Large Quantities of
16. Categorical Data. Journal of Applied Statistics, 29(2): 119-127.
17. Kim, S. S., Timothy, D. J., Hwang, J. 2011. Understanding Japanese Tourists'
18. Shopping Preferences Using the Decision Tree Analysis Method. Tourism Management, 32(3): 544-554.
19. Legohérel, P., Hsu, C. H. C., Daucé, B. 2015. Variety-seeking: Using the CHAIDsegmentation approach in analyzing the international traveler market. Tourism Management, 46: 359–366.
20. Levin, N., Zahav, J. 2001. Predictive modeling using segmentation. Journal of Interactive Marketing, 20(2): 3–22.
21. Ministry of Culture and Tourism, 2014. Certified Facility Management and
22. Investment Statistics 2014. The Number of Tourism Licenced Establishment by the Classification of Statistical Region Units, Room and Beds. http://yigm.kulturturizm.gov.tr/ TR,9860/turizm-belgeli-tesisler.html (Accessed to: 26.07.2016).
23. Ministry of Health, 2014. Distribution of hospitals and beds by provinces, 2014.
24. Ministry of Youth and Sports, 2014. Statistical Classification of Territorial Units for
25. the year and the number of sports clubs, 2007 - 2014.
26. Riquier, C., Luxton, S., Sharp, B. 1997. Probabilistic segmentation modelling. Journal
27. of the Market Research Society, 39(4): 571–588.
28. TurkStat, Foreign Trade Statistics, 2016. Exports by province, 2002-2016.

TurkStat,
29. Municipal Water Statistics, 2014. Amount of water distributed by municipalities via water supply network, 2014.
30. Turkish Statistical Institute, 2014. Culture statistics. https://biruni.tuik.gov.tr/medas/?kn=106&locale=tr (Accessed to: 22.07.2016).
31. TurkStat, Library Statistics 2014 Press Release, 2014. Formal and non-formal Education institutions libraries by Classification of Statistical Region and number of materials.
32. TurkStat, Road Motor Vehicles, December 2014. The number of road motor vehicles
33. by province, 2014.
34. Turkish Statistical Institute, 2014. Livestock Statistics. Cattle. https://biruni.tuik.gov.tr
35. /hayvancilikapp/hayvancilik.zul (Accessed to: 13.07.2016).
36. Turkish Statistical Institute, 2014. Livestock Statistics. Small ruminants. https://biruni.
37. tuik.gov.tr/hayvancilikapp/hayvancilik.zul (Accessed to: 13.07.2016).
38. Turkish Statistical Institute, 2014. Residential sales statistics.
39. https://biruni.tuik.gov.tr/medas/?kn=73&locale=tr (Accessed to: 21.07.2016).
40. Turkish Statistical Institute, 2014. Number of Agricultural Equipment and Machinery. https://biruni.tuik.gov.tr/bitkiselapp/tarimalet.zul (Accessed to: 06.07.2016).
41. Yu, Z., Fung, B. C. M., Haghighat, F. 2013. Extracting knowledge from building-related data –A data mining framework. Building Simulation, 6: 207–222.