Analysis of Chi-square Automatic Interaction Detection (CHAID) and Classification and Regression Tree (CRT) for Classification of Corn Production

Yuliana Susanti 1, a), Etik Zukhronah2-b), Hasih Pratiwi3-c), Respatiwulan4-d), and Sri Sulistijowati H.5,e)

1,2,3,4,5Study Program of Statistics, Universitas Sebelas Maret, Surakarta 57126, Indonesia

a)Corresponding author: yulianasusanti@staff.uns.ac.id
b)etikzukhronah@staff.uns.ac.id
c)hasihpratiwi@staff.uns.ac.id
d)respatiwulan@staff.uns.ac.id
e)rr.ssh@staff.uns.ac.id

Abstract. To achieve food resilience in Indonesia, food diversification by exploring potentials of local food is required. Corn is one of alternating staple food of Javanese society. For that reason, corn production needs to be improved by considering the influencing factors. CHAID and CRT are methods of data mining which can be used to classify the influencing variables. The present study seeks to dig up information on the potentials of local food availability of corn in regencies and cities in Java Island. CHAID analysis yields four classifications with accuracy of 78.8%, while CRT analysis yields seven classifications with accuracy of 79.6%.

1. Introduction
Corn is primary staple food widely produced in Indonesia. In reference to the 2015 data of Central Agency of Statistics (Badan Pusat Statistik—BPS) [3], national corn production of 19.61 million tons increased 3.18% compared to the 2014 data. The increase can be seen from the rise of corn production in regencies/ cities in Java of 0.46 million tons. The regencies or cities producing corn in Java are spread over such provinces as Banten, West Java, Central Java, Yogyakarta, and East Java. If corn production in Java cannot meet the needs of corn, it will depend on the corn production in regions outside Java.

Food availability is one of important instruments to create national food safety. Realizing the importance of attempts to optimize the potential of local food, we require an appropriate method to classify, predict, create production map, and formulate distribution patterns of food in Indonesia (Susanti et al., [11]). With regard to the importance of attempts to optimize corn production in regencies/ cities in Java, methods of exploring data to classify data by building such classification trees as chi Square Automatic Interaction Detection (CHAID) and Classification and regression trees (CRT).

The present study applies CHAID and CRT methods to classify corn production in regencies/ cities in Java in the year of 2015. The classification will result in tree diagrams and variables influencing corn production in regencies/ cities in Java. The classification of corn production in Java will provide positive inputs for policy strategies related to the optimization of corn production in regencies/ cities in Java.
2. Research Method

2.1. CHAID Method

There are three important things in CHAID methods (Sharp, et al.[10], (Bagozzi [2]):

1. The chi-Square test of independence is performed to identify significant independent variables in the data.
2. The Bonferroni correction is carried out when several statistical tests for independence are being performed simultaneously.
3. The CHAID algorithms are used to combine the influencing variable categories.

According to Gallagher [6], CHAID is a method which uses the chi-Square test of independence to examine categorical independent variables in classification individually towards the categorical dependent variables. The statistic for the chi-Square test of independence is

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where $i = 1, 2, ..., r$, $j = 1, 2, ..., c$. $O_{ij}$ represents the observed frequency of the cell and $E_{ij}$ represents the expected frequency of the cell. Bonferroni correction is a method used when several statistical tests for independence are being performed simultaneously (Sharp, et al.[10]). Galanger [6] defined it as the number of the ways $c$ categories can be merged into $r$ categories if independent variables having $c$ categories exist and can be alleviated into $r$ categories in procedure of merging. Therefore, the obtained $p$-value of the chi-Square test of independence is multiplication result of $p$-value and Bonferroni correction in accordance with the variable type.

The procedures in CHAID method is divided into three, i.e. merging, splitting, and stopping (Bagozzi [2], Alkhasawneh [1]).

1. Merging
   
   This stage involves a significance test of each categorical independent variable towards the dependent variable. The procedures include:
   a. Draw a contingency table for each categorical independent variable and dependent variable.
   b. Compute the chi-Square test of independence for each pair of categorical independent variables and dependent variables by creating $2\times J$ table contingency, where $J$ is the number of categorical dependent variables.

2. Splitting
   
   Splitting is carried out by selecting independent variables having the smallest or the most significant $p$-value which will be used as a node separator. The node separator is selected by comparing $p$-values obtained in previous stage in each independent variable. In case that there is no independent variable having the most significant $p$-value, splitting is not carried out and node is determined as the last node due to the absence of branch.

3. Stopping
   
   It is performed by repeating the stage of merging to analyze the following subgroups. If all of the subgroups have been analyzed, the process is stopped.

The results of segment formation in CHAID method are displayed in a classification tree diagram (Lehmann and Eberler [8]). In the diagram, a dependent variable is represented by a root. The independent variable directly associated with the root is that having significant smallest significant corrected $p$-value. The classification tree diagram follows top-down rule; the tree diagram is constructed from parents to subgroups split from the parents based on criterion of the number of independent variable categories after merging procedure (Myers [9]). Bagozzi [2] mentioned that a classification tree diagram in CHAID can illustrate:

1. Categorical independent variables which exert influence on merged categorical dependent variables.
2. Data size in a group which is represented by $N$
2.2. **CRT Method**

CRT is one of methods of decision tree applied to yield data set as attribute of a classification and to illustrate the relationship between response variables and one or more predictor variables. According to Breiman, et al. [4], if response variables are continuous, regression tree method is employed. Meanwhile, if response variables are categorical, classification tree method is used. The present study involves continuous response variables and therefore regression tree method was applied.

The regression tree method is technically known as binary recursive partitioning. It involves binary splitting since data set called node is always split into multiple partitions called child nodes. Breiman, et al. [4] pointed out that the main elements of the regression tree cover rules for splitting of each node, determining of each node, and determining of value with response estimation for each last node. The CRT method procedures include:

1. **Rules of splitting**

According to Breiman et al. [4], a regression tree is constructed through splitting of data groups and a series of binary splitting until the last node is yielded. To split a node into two child notes, the following rules should be followed:

   a. Each of the splitting depends on the value of one predictor variable.
   b. For continuous variables, median of two consecutive observed values of variables is used in the splitting.
   c. For categorical predictor variables, the splitting is derived from splitting probability based on the formation of two child nodes and of two disjoint sets.

2. **Process of splitting**

According to Breiman et al., [4] each node’s process of splitting is conducted by:

   a. Determining all splitting probabilities in each predictor variable.
   b. Selecting the best splitting in each predictor variable and in the best splitting sets.

The best splitting is determined with regard to the sum of squared deviation difference of each child node and parent node. The biggest difference will be selected as the best splitting. The error sum of squares on the t-th node is determined as a criterion of homogeneity in each node. Mean response in the node is predicted by the mean response in the node and is formulated by

   \[ \hat{Y}(t) = \frac{1}{n(t)} \sum_{x \in t} Y_i \]

The sum squares of error on node t is expressed as

   \[ SSE(t) = \sum_{x \in t} [Y_i - \hat{Y}(t)]^2 \]

For instance, the splitting s splits t into left child tL and right child tR. The function of the splitter is

   \[ \varphi(s, t) = SSE(t) - [SSE(t_L) + SSE(t_R)] \]

where \( \Phi(s, t) \) is the function of splitter in regression tree, SSE(t) is the sum squares of error of parent nodes, SSE(tL) the sum squares of error of left child, and SSE(tR) is the sum squares of error of right child. The best splitting s* is

   \[ \varphi(s^*, t) = \max_{s \in S} \Phi(s, t) \]

where S denotes a set of splitting probability.

3. **Rules of stopping**

The splitting is determined by splitting data into two parts. The parts have both left child and right child. The process of splitting is repeated towards both children until it is impossible to do and finally it is stopped. Breiman, et al. [4] stated that the recursive process will discontinue regarding the minimum number of observations. In addition, the regression tree is discontinued by maximizing the homogeneity of the variance of each node, and it is influenced by variables affecting responses. Unsplittable nodes are termed the last nodes.

4. **Rules of determining the estimate value of response**
In a regression tree method, if a node is determined as the last node, the response estimation of the observation of the last node is the mean response. Therefore, the mean response is used as the response estimator.

3. Results and Discussion
The data used in the present study involve corn production in regencies and cities in Java Island in the year of 2015 which were obtained from the BPS [3] and Department of Agriculture [5]. The dependent variables include corn production in regencies and cities in Java, while the independent variables include the harvested area, temperature, rain fall, and area altitude of each regency and city.

The areas with corn production in Java in 2015 included 113 regencies and cities. Dependent variable (corn production) is categorized based on median of corn production in all regencies/cities in Java. The categories are: low (1) (< 6,067 tons) and high (2) (≥ 60,670 tons). For independent variables in CHAID method, the width of harvested area is classified as narrow (1) (< 35,000 ha), moderate (2) (35 ha – 100,000 ha) and wide (3) (> 100,000 ha), rainfall is classified as low (1) (< 85 mm/month), moderate (2) (85 – 200 mm/month and high (3) (> 200 mm/month), while temperature as low (1) (< 230C), moderate (2) (230C - 230C) and high (3) (> 230C) and area altitude as low (1) (<100 m), moderate (2) (100-600 m), and high (3) (>600 m). In CRT method, independent variables are continuous.

A tree diagram of CHAID method is illustrated in Figure 1 and the classification, along with the characteristics (Table 1), as well as the percentage of each classification are demonstrated (Table 2). Table 2 indicated that the biggest percentage of total production of corn in regencies/cities in Java of less than 6,067 tons (low corn production) is on the 1st classification (the total production of corn with low and high rainfall is 84.8%). Meanwhile, the total production of corn in regencies/ cities in Java which is greater than or equal to 6,067 tons (high corn production) is on the 4th classification (corn production with moderate rainfall, moderate and wide harvested area is 79.1%).

![Figure 1. The Classification of Corn Production in Java Using CHAID method](image-url)
Table 1. Classification of Corn Production Based on Tree Diagram of CHAID

| Classification | Node | Characteristics |
|----------------|------|-----------------|
| 1st            | 1    | areas with low or high rainfall will yield low corn production |
| 2nd            | 2,3,6| areas with moderate rain fall, narrow harvested area, and high or low temperature will yield low corn production |
| 3rd            | 2,3,5| areas with moderate rain fall, narrow harvested area, and moderate temperature will yield high corn production |
| 4th            | 2,4  | areas with moderate rain fall, wide harvested area will also yield high corn production |

Table 2. Percentage of Classification of Corn Production Using CHAID Method

| Classification | Low | High |
|----------------|-----|------|
|                | Number of regencies/cities | Percentage | Number of regencies/cities | Percentage |
| 1              | 28  | 84.8%| 5     | 15.2% |
| 2              | 13  | 30.0%| 14    | 70.0% |
| 3              | 6   | 76.5%| 4     | 23.5% |
| 4              | 9   | 20.9%| 34    | 79.1% |

Figure 2. The Classification of Corn Production in Java Using CRT method
Table 3. The Classification based on CRT Tree Diagram

| Classification | Node | Characteristics |
|----------------|------|-----------------|
| 1st            | 1,3,7| harvested area (< 72,852 ha), rainfall (< 156,693 mm/month) and area altitude of < 28,595 m will yield low corn production. |
| 2nd            | 1,3,8| low harvested area, low rainfall and area altitude > 28,595 m will result in lower maize production |
| 3rd            | 1,4,9| narrow harvested area, high rainfall, and area altitude < 151,645 m will yield high corn production |
| 4th            | 1,4,10| narrow harvested area and area altitude > 151,645 m will yield low corn production |
| 5th            | 2,5  | wide harvested area and area altitude < 131,105 m will yield high corn production |
| 6th            | 2,6,11| wide harvested area, area altitude of between 131,105 m and 350,265 m will yield low corn production |
| 7th            | 2,6,12| wide harvested area and area altitude of > 350,265 m will yield high corn production |

Table 4. The Percentage of Classification of Corn Production Using CRT Method

| Classification | Low (< 6,067 tons) | High (≥ 6,067 tons) |
|----------------|--------------------|--------------------|
|                | Number of regencies/cities | Percentage | Number of regencies/cities | Percentage |
| 1              | 5                   | 55.6%              | 4                   | 44.4%     |
| 2              | 22                  | 91.7%              | 2                   | 8.3%      |
| 3              | 4                   | 25%                | 12                  | 75%       |
| 4              | 13                  | 65%                | 7                   | 35%       |
| 5              | 3                   | 12%                | 22                  | 88%       |
| 6              | 8                   | 80%                | 2                   | 20%       |
| 7              | 1                   | 11.1%              | 8                   | 88.9%     |

The classification of corn production in Java using CRT method is shown by Figure 2. Figure 2 explains the classification and their characteristics while Table 3 and Table 4 demonstrates the percentage. It is clear from Table 4 that the biggest percentage of total production of corn in regencies/cities in Java of less than 6,067 tons is on the 2st classification. The classification involves corn production in regencies/cities in Java having width of harvested area less than or equal to 72,852 hectares, rainfall less than 156,693 mm and area altitude more than 28,595 m with percentage of 81.8%. The total production of corn in regencies/cities in Java with greater than or equal to 6,067 tons is on the 7th classification (corn production in regencies/cities in Java with width of harvested area greater than 72,852 hectares and area altitude greater than 350.265 m) with percentage of 88.9%. By using CRT method, the classification of data of corn production in regencies/cities in Java is found out.
CHAID analysis results in four classifications: 1) the first classification indicates that areas with low or high rainfall will yield low corn production, 2) the second classification shows that areas with moderate rainfall, narrow harvested area, and high or low temperature will yield low corn production, 3) the third classification demonstrates that areas with moderate rainfall, narrow harvested area, and moderate temperature will also yield high corn production, and 4) the fourth classification shows that areas with moderate rainfall and wide harvested area will yield high corn production. Variables influencing corn production are rainfall, harvested area, and temperature.

Meanwhile, CRT analysis identifies seven classifications: 1) the first classification shows that areas with narrow harvested area less than 72,852 ha, low rainfall less than 156,693 mm/month and area altitude less than 28,595 m will yield low corn production, 2) the second classification demonstrates that areas with narrow harvested area, low rainfall and area altitude greater than 28,595 m will yield low corn production, 3) the third classification indicates that areas with narrow harvested area, high rainfall and area altitude less than 151,645 m will yield high corn production, 4) the forth classification shows that areas with narrow harvested area, high rainfall and area altitude greater than 151,645 m will yield low corn production, 5) the fifth classification points out that areas with wide harvested area and area altitude less than 131,105 m will yield high corn production, 6) the sixth classification shows that areas with wide harvested area, area altitude between 131,105 m and 350,265 m will yield low corn production, and 7) the seventh classification indicates that areas with wide harvested area and area altitude greater than 350,265 m will yield high corn production.

**Table 5. The risk of classifying corn production by CHAID and CRT methods**

| Method | Accuracy | Standard error |
|--------|----------|----------------|
| CHAID  | 0.788    | 0.038          |
| CRT    | 0.796    | 0.038          |

**Table 6. Prediction of classification of corn production by CHAID method**

| Observed | Predicted | Low (< 6067 Ton) | High (>= 6067 Ton) | Percent Correct |
|----------|-----------|------------------|--------------------|-----------------|
| Low (< 6067 tons) | 41 | 15 | 73.2% |
| High(>= 6067 tons) | 9 | 48 | 84.2% |
| Overall Percentage | 44.2% | 55.8% | 78.8% |

**Table 7. Prediction of classification of corn production by CRT method**

| Observed | Predicted | Low (< 6067 Ton) | High (>= 6067 Ton) | Percent Correct |
|----------|-----------|------------------|--------------------|-----------------|
| Low (< 6067 tons) | 48 | 8 | 85.7% |
| High(>= 6067 tons) | 15 | 42 | 73.7% |
| Overall Percentage | 55.8% | 44.2% | 79.6% |

Based on Table 5, the CHAID and CRT methods provide the same standard error of 0.038. The CHAID method can predict the classification of corn production in Java with a precise maximum of 78.8%, while the CRT method can predict the classification of corn production in Java with a precise maximum of 79.6%. Table 6 shows the percentage for predicting a lower corn production or less than 6067 tons correctly equal to 73.2%, meanwhile for predicting higher corn production or more than 6067 tons correctly equal to 84.2%. This shows that the CHAID model is the best to predict higher corn production or more than 6067 tons with a percentage of 84.2%. Table 7 shows the percentage for predicting lower corn production or less than 6067 tons in Java correctly equal to 85.7%, meanwhile
for predicting higher corn production or more than 6067 tons in Java correctly equal to 73.7%. This suggests that the CRT model is the best to predict lower corn production or less than 6067 tons with a percentage of 85.7%.

4. Conclusion
CHIAID analysis yields four classifications and variables that affect corn production are rainfall, harvested area, and temperature. CRT analysis yields seven classifications and influencing variables are harvested area, temperature, and area altitude. The CHIAID and CRT methods can predict the classification of corn production in Java with a precise maximum of 78.8% and 79.6% respectively.

5. Acknowledgments
The author would like to thank the Indonesian Ministry of Research, Technology and Higher Education as well as Universitas Sebelas Maret for financial support through Grant of Leading Research of University 2017.

6. References
[1] M.Sh. Alkhasawneh, Umi, K.N., Lea, T.T., Nor Ashidi, M.I., and Mohammad, S.A., Modeling and Testing Landslide Hazard Using Decision Tree, Journal of Applied Mathematics, Hindawi Pub. Corp., 2014
[2] R.P. Bagozzi, Advanced Methods of Marketing Research, Blackwell Publishers Ltd, Oxford, 1994
[3] Badan Pusat Statistik, Production of Paddy Maize and Soybeans, www.bps.go.id/release/Production of Paddy Maize and Soybeans, 2015
[4] L. Breiman, Friedman, J.H, Olshen, R.A and Stone, C.J. Classification And Regression Tree, Chapman And Hall, New York, 1993
[5] Dinas Pertanian, Suhu dan Curah Hujan dalam Wilayah Regional Pulau Jawa, www.distan.go.id/download/outlookshudancurahhujan.pdf, 2015
[6] C.A. Gallagher, An Iterative Approach to Classification Analysis, Journal of Applied Statistics, 29(2000), pp 256-266
[7] G.V. Kass, An Exploratory Technique for Investigating Large Quantities of Categorical Data, Journal of Applied Statistics 29 (1980), pp 119-127
[8] T. Lehmann, and Eherler, D., Responder Profiling with CHIAID and Dependency Analysis, Jena University, Jena, 2001
[9] J.H. Myers, Segmentation and Positioning for Strategic Marketing Decisions, American Marketing Association, Chicago, 1996
[10] A. Sharp and J. Romaniuk, The Performance of Segmentation Variables: A Comparative Study, www.anzmac1998/Sharp222.pdf, 2002
[11] Y. Susanti, Pratiwi, H., and Handayani, S.S., Paddy Availability Modeling in Indonesia Using Spatial Regression, International Journal of Applied Mathematics, 45 (2015) pp 349-360
[12] A. Widiastuti, Penelusuran Karakteristik Kepemilikan Kartu Kredit dengan Metode CHAID dan Exhaustive CHAID, Skripsi, Institut Pertanian Bogor, 2001
[13] W. Yuniar, Penentuan Penciri Keragaman Konsumsi Pangan Rumah Tangga Rawan Pangan dengan Metode CHAID dan Exhaustive CHAID, Skripsi, Institut Pertanian Bogor, 2014
[14] J. Zhang and Chikaraishi, M., Interdependences Household Residential and Car Ownership Behavior, Journal of Transport Geography 34 (2014), pp 165-174