IJAMEC
e-ISSN: 2147-8228

INTERNATIONAL JOURNAL OF APPLIED MATHEMATICS ELECTRONICS AND COMPUTERS

www.dergipark.org.tr/ijamec

International Open Access

Volume 09
Issue 04

December, 2021

*Research Article*

# A classification study for re-determination of the geographical regions: the case of Turkey

***Burcu Durmuş*** [a,*], ***Öznur İşçi Güneri*** [a]

[a] *Mugla Sitki Kocman University, Department of Statistics, 48000, Mugla, Turkey*

ARTICLE INFO

*Article history:*
Received 29 August 2021
Accepted 13 December 2021

*Keywords:*
Classification algorithms
Geographical regions
Regional border
Urban areas

ABSTRACT

The rapid urbanization of the residential areas in an unplanned and irregular manner causes ecological and environmental changes. Understanding this growth is an important criterion for developing environmentally friendly policies and creating new limits. Although determining the boundaries of the region cannot be connected to a single basic principle, a planning can be made by making statistical inferences. In this regard, in this study, Turkey's geographic boundaries are discussed again. A classification analysis was conducted to discuss how the change of environmental and human criteria affects the limits. The study was carried out with two data sets. The first is the provinces that exist in the present, and the second is the selected developing districts. The algorithm that gave the best results for the data was determined. The reason for the provinces classified incorrectly according to this algorithm was discussed and a new border map was proposed.



## 1. Introduction

Geographical regions are the borders drawn in relation to social, cultural and economic structures. These are the areas that are formed by the mutual interaction of all-natural factors on the earth and construct a whole through the characteristics they have [25]. The separation of regions provides institutional responses to the problems by creating regional administrations. These regions are useful at different scales and for various purposes such as economic, cultural and political [14-16]. For example, metropolitan regions of Turkey are considered as a combination of districts or political subdivisions, and region concepts are used for different statistical, management, and planning purposes. Recent studies show that the growing districts and sub-district settlements have started to become cities [8]. This growth rate indicates the need to keep the situation under control in order to better plan future growth. It is very important to follow these constantly growing cities both for determining the distribution of labour force and for administrative planning goals. When a growing city becomes a metropolis, it affects the surrounding cities in social and economic terms and also changes the balances between the cities. Understanding growth that is generally unplanned is an important criterion for developing environmentally sound policies and creating new limits. In the last century, many developing countries have rapidly entered the process of urbanization [5-6]. Between 1950 and 2014, the share of the population living in the urban area increased from 30% to 54%, and 2.5 billion people are expected to be added to this number by 2050, especially in Asia and Africa [8].

Developing technology, living standards, climate change and other conditions are changing day by day. Correspondingly, many changes have also been occurring in human settlements. For example, a city famous for its history can be turned into a city of sea tourism and the nature of industrial, agricultural and animal breeding activities is changing with the evolving technology. Urbanization process affects economic [30] and social development deeply [5]. This has direct consequences for biodiversity, resource conservation and environmental degradation [21-23]. It is known that an indicator of urbanization is groundcover [12]. It is observed that the land cover also changes in the rapidly developing and changing cities. Considering that the world values are always in a dynamic

* Corresponding author. E-mail address: burcudurmus@mu.edu.tr
DOI: 10.18100/ijamec.988273

structure, it is a fact that this situation is inevitable. It is difficult to make decisions on the regional borders of cities for many reasons such as economic, social, cultural, physical and political. In addition, when the state sets some conditions statistically, especially in the opportunity regions, it is a problem to determine the regional borders. Therefore, it is impossible to determine the geographical regions according to a fixed principle [32]. All these and similar reasons made this research necessary.

Turkey is a huge country and geographical research carried out so far has not reached the level to create sufficient knowledge yet [4]. For this reason, a new data set was presented for the geographical region boundaries in the current study. By examining the boundaries of the region with new methodologies, it was investigated how especially the developing cities affected the border changes. Statistical and geographical implications were made using the algorithms in the field of machine learning for the determination of the properties related to the borders of Turkey, which is in an urban transitional period. The methodology in the study is such that it can be easily applied to other countries and regions. In addition, this study differs from previous reviews in several respects. With the developing technology, all kinds of information can be produced, collected and recorded today. These data contain many information and correlations that will benefit the institutions. The increase of storage areas prevents the loss of information while at the same time enables the storage of data in many areas [3]. These data can be used to conduct many analyzes and valuable information can be obtained from these analyzes. However, this information can be complex and meaningless. Therefore, it is necessary to reveal meaningful and useful information from these stacks [2]. In this regard, the current study is important because it, for the first time, presents a comprehensive set of geographic data for Turkey. Therefore, the firstly, a comprehensive geographic data set was given available for Turkey. Some data sets in the literature are not approved, their basic accuracy is insufficient, and most of them are small in size and spatial level. This data set is a referenced data set for the use and verification of existing information. Secondly, a second data set was presented for the cities that can gain province status based on the factors determining Turkey's borders. Third, the findings were supported by several different methodologies throughout the study and more meaningful results were tried to be obtained. Classifiers were evaluated by feature selection, data normalization, k-fold cross verification. Finally, the boundaries of the region are handled from a different perspective. Machine learning techniques were used to determine the variables underlying the decisions regarding the borders.

In summary, performance results were compared by taking classification algorithms for determining the geographical regions of Turkey according to today's features. To this end, various features of the 81 cities in Turkey were identified. Data for each feature was compiled from sources in the respective fields. Classification analysis was performed after data pre-processing and data cleaning. Algorithms suitable for the data set which are among the decision trees such as Bayes classifier, k-nearest neighbor classification methods that are used in the literature were selected for analysis. After applying each algorithm to the data set, the algorithm that gave the best results for the region classification was chosen by comparing the results. The study generally consists of two stages. In addition to the province's dataset, a dataset was also arranged for some districts selected. At the end of the study, based on the results of the algorithm, a new map was proposed for the region borders of Turkey and how the results affect the borders was discussed.

## 2. Material and Method

### *2.1. Study Area*

Turkey is considered as one of the world's popular countries with its location, as well as its economic, cultural, and political activities. In line with the decisions taken as a result of the 1st Geography Congress, the country is divided into 7 geographical regions [1]. Figure 1 shows the limits of Turkey's seven geographical regions. The residential areas of the country are rapidly urbanizing. According to the data of the Ministry of Environment and Urbanization, the population according to the first census carried out in 1927 was 13.648.270 with 75.8% of the population living in villages, while 24.2% living in the provinces and districts. The population started to gather in urban areas after 1950. While the population is rapidly increasing from one side, on the other hand, migration from rural to urban areas is experienced at the same rate in Turkey, which demonstrates a rapidly changing dynamic demographic structure [10]. When we came from 1965 to 2009, the table, which previously showed low and moderate urbanization, was replaced by high and very high urbanization [11]. According to World Bank data, population living in urban areas in 2017 is 74.4% of the whole population in Turkey, while it is 76.4% in the EU-28 countries [9]. According to the analysis done by the UN for 2050 for Turkey, it is estimated that the urbanization rate will reach 87% and will capture the OECD average [29].

Urbanization in Turkey is also reflected in the rapid expansion of settlements. Region conditions and properties are affected by many profiles, including land, land cover and, accordingly, weather and climate changes. For example, there are population movements towards industry to Çorlu, Çerkezköy, Muratlı, Ergene and Kapaklı districts in the Thrace region [20]. This situation is important for the region in two aspects. First, population mobility has increased the rate of urbanization. Secondly, there are changes in the vegetation and climate of the region due to reasons such as air and water pollution and the conversion of agricultural

lands to the industrial zone. In another study carried out in the Naip Plain, as a result of the spatial and volumetric growths experienced in the mines, it was emphasized that environmental problems such as pollution and natural landscape degradation, land use change, mass movements and erosion occurred [26].


**Figure 1.** Turkey's Geographic Region Boundaries

While determining the boundaries of the region at the 1st Geography Congress, climate, vegetation, physical values, country borders, land structure and human factors were taken into consideration. Given that there are 7 regions in Turkey, the borders of these regions should be revised as a result of the changing conditions. With this study, an updated review of the data related to the above-mentioned characteristic of Turkey was made.

### *2.2. Dataset Construction*

It is known that the methods or variables used in determining the boundaries of the region change day by day. Limits can be determined by taking many physical or human criteria into consideration according to the needs. In this study, analyzes were made on the variables taken as basis in the 1st Geography Congress. The data used were compiled from official or referenced institutions on the internet. Turkey's border regions were evaluated over the existing 7 regions (Marmara, Aegean, Mediterranean, Black Sea, Central Anatolia, Eastern Anatolia, and South East Anatolia) in terms of 18 variables. Information on these variables is given in Table 1.

**Table 1.** Summary of variables and source of information

| | Variable | Summary Information | Source |
|---|---|---|---|
| 1 | Latitude | Mathematical position (degrees) | Municipality, Governorate or Ministry |
| 2 | Longitude | Mathematical position (degrees) | Municipality, Governorate or Ministry |
| 3 | Population | Population number (numerical) | Turkish Statistical Institute |
| 4 | Area | $km^2$ (numerical) | Municipality, Governorate or Ministry |
| 5 | Climate | Natural structure (categorical) | General Directorate of Meteorology |
| 6 | Flora | Natural structure (categorical) | Environment and Urban Ministry |
| 7 | Soil type | Natural structure (categorical) | Municipality, Governorate or Ministry |
| 8 | Temperature | Annual average (degrees) | tr.climate-data.org |
| 9 | Precipitation | Annual average (mm) | General Directorate of Meteorology |
| 10 | Altitude | Natural structure (numerical) | Municipality, Governorate or Ministry |
| 11 | Mountains | Natural structure (parallel, upright, scattered) | Municipality, Governorate or Ministry |
| 12 | Lakes | Number of lakes (numerical) | Municipality, Governorate or Ministry |
| 13 | Seashore | Natural structure (yes, no) | Municipality, Governorate or Ministry |
| 14 | Transportation | Number of transportation (numerical) | Municipality, Governorate or Ministry |
| 15 | Industry | Dominant industry type (categorical) | Municipality, Governorate or Ministry |
| 16 | Tourism | Dominant tourism type (categorical) | Culture and Tourism Ministry |
| 17 | Trade | Trade intensity (categorical) | Municipality, Governorate or Ministry |
| 18 | Region | Current geographic region | 1st Geography Congress |

### *2.3. Data Preprocessing and Data Cleaning*

The first step in the data set creation is data preprocessing and data cleaning. In this step, the data is cleared of noisy data. Noisy data negatively affect the results of the analysis, reducing the success rate.

*Feature Selection*

Feature selection algorithms aim to find and extract a subset of features that best represent the set [7]. Some features in the dataset may have noise that will adversely affect processing performance. Removing these features from the dataset can be effective in increasing the accuracy of the result. In addition, reducing the size of the data to be used in algorithms saves time on issues that affect the processing process such as processing power, memory requirement and storage space [31]. In this study, Turkey's border regions are examined with machine learning methods. Feature selection is used to remove noise from data. Finally, how this method affects the operation of algorithms is discussed.

*Normalization*

Normalization process is to handle the data in a single order and facilitate the comparability of the data in cases where there is a great difference between the data in the data set [28]. Considering the min-max values of the data, too much of this range may affect the success of other values. Wide range of data is normalized to reduce the effect of the variable on other variables. Thus, it can be observed that the result is more successful [18]. Data normalization can be done with various algorithms such as z-score, min-max normalization.

The data for our study were collected with different scales. How the data behaved after the normalization process was controlled to increase the consistency between the data and obtain more successful results.

### *2.4. Classification Analysis*

Classification is the assignment of a variable to one of the predefined classes by determining its properties. The important point here is that the classes are known in advance. Classification consists of two stages: training and testing. In the training phase, a classification model is created using training data and learning algorithms. In the testing phase, classes are estimated by applying the test data to the model. In the absence of separate test data, test data are obtained from the current data set.

*Cross Validation*

Cross validation is a technique used in model selection. In this technique, the data set is divided into k subgroups. One group is used as a test set and the remaining groups are used as a training set. This calculation repeats k times. In studies, generally 5-fold or 10-fold cross validity is used depending on the size of the data set.

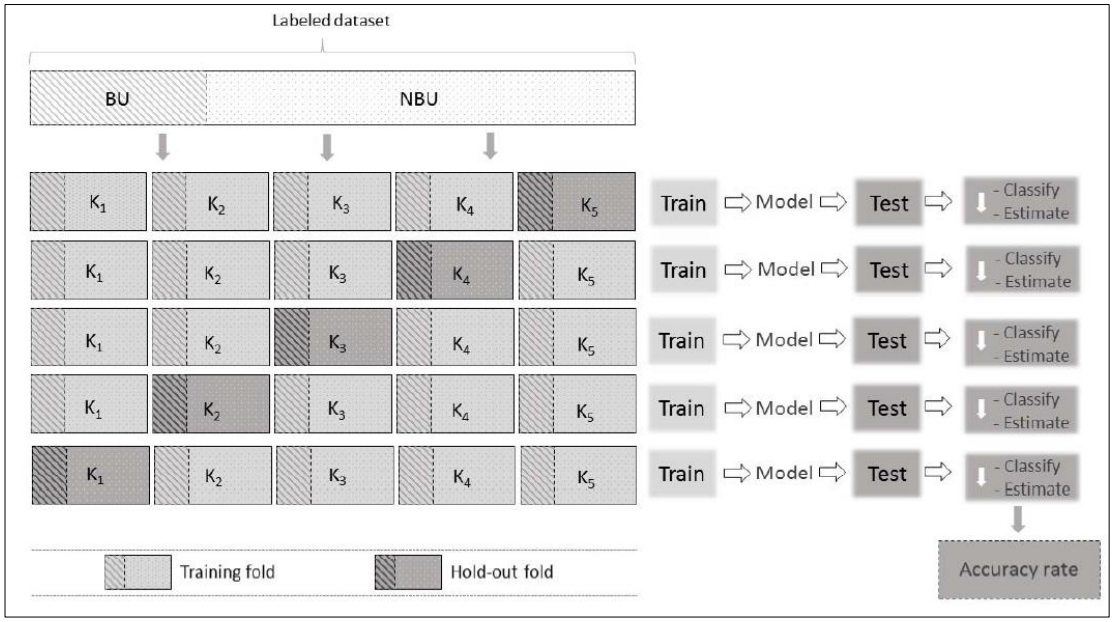


**Figure 2.** k-fold (5-fold) cross validation scheme (This figure is adapted from [27])

*Accuracy*

In evaluating the performance of the selected algorithms, the criteria such as correct classification rate and incorrect classification rate are used by calculating over the cross table (Table-2).

**Table 2.** 2x2 cross table

| | | Predicted Class | |
|---|---|---|---|
| | | Class=1 | Class=2 |
| Actual Class | Class=1 | TP | FN |
| | Class=2 | FP | TN |

TP: True Positive, FP: False Positive,
TN: True Negative, FN: False Negative

The correct classification rate was used to evaluate algorithm success in the study. However, it was still checked whether other criteria give parallel results. The correct classification rate is a measurement obtained by proportioning the number of correctly classified observations to the total number of samples (Durmus et al., 2019). It is an indicator of how many of the variables subject to classification are classified correctly.

*Classification Algorithms*

Different results can be obtained by selecting the algorithm in the classification model. These algorithms have been developed based on the mathematical and statistical methods. There are many algorithms in the literature. In this study, decision trees, Bayes classifier and nearest neighbor algorithms were used. While selecting the algorithm, algorithms that are proven in the literature and accepted by the researchers were preferred.

*Decision Trees:* In order to create a decision tree, there is a variable that best expresses the examples in the training set. A new variable is found from the samples above the selected branch, thus creating new branches. If there is no other variable to distinguish the samples, the branching is finished. Decision tree can be defined as a flow chart with root at the top, branches and leaves underneath. In order to classify a previously unknown element, the feature of the element is tested in the decision tree structure and an attempt is made to reach the prediction of which class the element will be included in [19].

*Bayes Algorithm:* This model is a classification method based on Bayes decision theory. Qualifications are evaluated independently and equally effectively. Bayes algorithm works with categorical data. Therefore, the continuous values in the dataset need to be categorized. Data that are not categorized may decrease model success more or less depending on the degree of impact to the model. The algorithm calculates how many times each output in the learning set occurs. In addition, it calculates the frequency of occurrence of the relationship between independent or dependent variables. Both calculations are used to make predictions [17].

*Nearest Neighbor:* In this algorithm, it is decided to classify by looking at how many k values are in the classification process. In the classification of the new value, the k value is compared to the closest k value. The one with the most classes becomes the class of the new value [24]. When calculating the nearest point k, various distance calculation functions such as Manhatten calculation, Minkowski calculation, Euclidian calculation are used.

### *2.5. Weka Program*

Weka is an open-source data mining software developed in Java language. It is a useful program for basic data mining operations such as classification, clustering, association rules analysis, data preprocessing. In order to be able to operate with Weka, the data should be edited with the *.arff file type, a simple text file format specific to the Weka program. It also supports text-based formats such as *.csv, *.libsvm. In the study, Weka software was used for both data preprocessing and classification analysis.

## 3. Results

Data set comprising the features of Turkey's cities passed through several preprocessing stages. Thus, it is aimed to create a more successful model free from noisy data. First, variable reduction operation was conducted on the 18 factors that could affect the geographical regional borders of Turkey by using the feature selection method. With the full training set test option, longitude, latitude, vegetation, temperature, precipitation, mountains, soil type, and industry variables were found to contribute significantly to the model. In addition, with the 10-fold and 5-fold cross-validation test option, it was found that the climate (folds: 10) and altitude (folds: 5) features are more effective than the other 8 factors.

Based on these results, a total of 10 factors obtained from both test options were taken into account in the analysis. Secondly, the data were standardized by the normalization process. Although it was observed that success increased with normalization most of the time, it was observed that this process did not contribute to the model in determining the limit. There may be two reasons for this. First, the data is already suitable for classification in its original form. Second, the data obtained by feature selection gives appropriate results without the need for normalization. Considering the effect of feature selection on classification success, it is obvious that the second situation will be more feasible.

Classification Analysis for City Boundaries

In this study, where the appropriate algorithm and model success were investigated, the classification analysis was carried out by using the 10-fold and 5-fold cross validity test with both the original data and the noiseless data obtained by feature selection. According to the results in Table 3, Bayes Net algorithm was found to be more successful than the other algorithms for the original data. The algorithm made better classification in 10-fold test selection with a 3.70% difference compared to the 5-fold test selection. Other algorithms yielded results much better Bayes Net algorithm. When the noiseless data were examined, it was seen that Bayes Net algorithm came to the fore again. Bayes Net exhibited a success with the same accuracy rate in the selection of 10-fold test. The algorithm lagged behind the 10-fold test selection with a 2.47% difference in the 5-fold test selection for noiseless data. To evaluate the adequacy of the classification test, the ROC analysis of the receiver operating characteristics was performed [13]. As the ROC score approaches 1, positives are better separated from negatives. Accordingly, when the algorithm results are examined, Bayes Net algorithm is said to be more sufficient in both original data and noiseless data with 1.00 ROC score. Almost all algorithms are economical in terms of model calculation time.

**Table 3.** Analysis results of classification algorithms for 10-fold and 5-fold test selection

| Algorithm | Correct Classification Rate | | | | ROC Analysis | | | | Calculation Time (sec) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Original | | Noiseless | | Original | | Noiseless | | Original | | Noiseless | |
| | 10-fold | 5-fold | 10-fold | 5-fold | 10-fold | 5-fold | 10-fold | 5-fold | 10-fold | 5-fold | 10-fold | 5-fold |
| J48 | 70.37 | 69.13 | 72.84 | 71.60 | 0.89 | 0.91 | 0.99 | 0.97 | 0.01 | 0.00 | 0.00 | 0.00 |
| LMT | 74.07 | 69.14 | 74.07 | 74.07 | 0.97 | 0.98 | 0.98 | 0.98 | 0.02 | 0.06 | 0.07 | 0.06 |
| R. Forest | 79.01 | 72.84 | 77.78 | 75.30 | 0.99 | 0.98 | 0.98 | 0.98 | 0.05 | 0.01 | 0.03 | 0.01 |
| R. Tree | 60.49 | 60.49 | 58.02 | 67.90 | 0.79 | 0.69 | 0.77 | 0.84 | 0.00 | 0.00 | 0.00 | 0.00 |
| **Bayes Net** | **83.95** | 80.25 | **83.95** | 81.48 | 1.00 | 1.00 | 1.00 | 1.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| N. Bayes | 71.60 | 74.07 | 79.01 | 77.78 | 0.99 | 0.99 | 0.99 | 0.98 | 0.00 | 0.00 | 0.00 | 0.00 |
| SMO | 67.90 | 69.14 | 74.07 | 72.84 | 0.94 | 0.95 | 0.96 | 0.96 | 0.17 | 0.11 | 0.16 | 0.09 |
| Logistic | 64.20 | 69.14 | 70.37 | 69.14 | 0.98 | 0.96 | 0.98 | 0.98 | 0.35 | 0.11 | 0.25 | 0.08 |
| S. Logistic | 72.84 | 69.14 | 74.07 | 74.07 | 0.97 | 0.98 | 0.98 | 0.98 | 0.17 | 0.03 | 0.03 | 0.02 |
| KStar | 23.45 | 23.45 | 79.01 | 77.78 | 0.70 | 0.70 | 0.99 | 0.99 | 0.00 | 0.00 | 0.00 | 0.00 |
| IBk | 67.90 | 66.67 | 72.84 | 72.84 | 0.81 | 0.84 | 0.94 | 0.93 | 0.00 | 0.00 | 0.00 | 0.00 |

In summary, the regions with existing geographical borders of Turkey's provinces can be determined by 83.95% accuracy with both original and noiseless data. All variables included in the analysis in both data sets contributed to the classification. In order to determine which data to continue with, Bayes Net algorithm results were examined with detailed classification statistics in Table 4. According to the results, it was seen that close results were obtained for both classifications.

**Table 4.** Detailed classification statistics for Bayes Net algorithm

| Summary for Original Data | | Summary for Smoothed Data | |
|---|---|---|---|
| Kappa statistic | 0.81 | Kappa statistic | 0.81 |
| Mean absolute | 0.06 | Mean absolute | 0.05 |
| Root mean | 0.20 | Root mean | 0.19 |
| Relative absolute | 22.76 % | Relative absolute | 21.79 |
| Root relative | 56.40 % | Root relative | 54.88 |
| TP Rate | 0.84 | TP Rate | 0.84 |
| FP Rate | 0.06 | FP Rate | 0.03 |
| Precision | 0.75 | Precision | 0.86 |
| Recall | 0.92 | Recall | 0.84 |
| F-measure | 0.83 | F-measure | 0.84 |
| MCC | 0.80 | MCC | 0.82 |
| PRC | .096 | PRC | .092 |

The detailed statistics of Bayes Net algorithm did not give a clear result about which data (original or noiseless) to use for border analysis. In addition, wrongly classified provinces were also examined. The results are presented in Table 5. According to these results, 10 cities were classified incorrectly in both cases. Apart from this, Tokat, Denizli and Çorum were misclassified in the original data, while Çankırı, Hakkâri, Kilis were misclassified in the noiseless data. When the information in Table 5 is analyzed, it can be seen that the reason why vast majority of the provinces were classified incorrectly is transition climates and, accordingly, the characteristics of soil type and vegetation. It is thought that the wrong classification of Burdur and Amasya provinces is due to the inability to estimate the longitude. This explanation shows that the Bayes Net classifier is actually doing the right thing.

*Classification Analysis for District Boundaries*

In this part of the study, we measured how well the new sample set was classified with the Supplied Test Set option.

For this test; characteristics of the districts projected to gain provincial status in Turkey's verbal and written press were considered as a new test kit. These districts whose classification will be tested are given in Table 6.

This new data set, which covers 37 districts, was classified in the first stage with the best performing Bayes Net algorithm. Classification analysis was done with both original and noiseless data. The classification result can be seen in Table 7. When the table is examined, it is seen that Bayes Net algorithm classifies the provinces test set with 86.4865% accuracy rate and 0.98 ROC value in both cases in line with the model obtained.

**Table 5.** Provinces classified incorrectly with Bayes Net algorithm

| Original Data | | | Smoothed Data | | |
|---|---|---|---|---|---|
| Provinces | Observed Region | Real Region | Provinces | Observed Region | Real Region |
| Sakarya | Black Sea | Marmara | Sakarya | Black Sea | Marmara |
| İzmir | Mediterranean | Aegean | İzmir | Mediterranean | Aegean |
| Muğla | Mediterranean | Aegean | Muğla | Mediterranean | Aegean |
| Kütahya | Central Anatolia | Aegean | Kütahya | Central Anatolia | Aegean |
| Afyon | Central Anatolia | Aegean | Afyon | Central Anatolia | Aegean |
| Burdur | Southeastern Anatolia | Mediterranean | Burdur | Southeastern Anatolia | Mediterranean |
| Bayburt | Eastern Anatolia | Black Sea | Bayburt | Eastern Anatolia | Black Sea |
| Amasya | Southeastern Anatolia | Black Sea | Amasya | Southeastern Anatolia | Black Sea |
| Sivas | Eastern Anatolia | Central Anatolia | Sivas | Eastern Anatolia | Central Anatolia |
| Siirt | Eastern Anatolia | Southeastern Anatolia | Siirt | Eastern Anatolia | Southeastern Anatolia |
| Tokat | Central Anatolia | Black Sea | Çankırı | Black Sea | Central Anatolia |
| Çorum | Central Anatolia | Black Sea | Hakkâri | Southeastern Anatolia | Eastern Anatolia |
| Denizli | Mediterranean | Aegean | Kilis | Mediterranean | Southeastern Anatolia |

**Table 6.** Districts selected for the test set

| | | |
|---|---|---|
| Çorlu (Tekirdağ) | Erbaa (Tokat) | Kahta |
| Alanya (Antalya) | İnegöl (Bursa) | Kozan (Adana) |
| Şebinkarahisar | Ergani | Suşehri (Sivas) |
| Beypazarı | İskenderun | Sungurlu |
| Manavgat | Ereğli (Konya) | Yüksekova |
| Şereflikoçhisar | Midyat | Cizre (Şırnak) |
| Bandırma | Fethiye | Zile (Tokat) |
| Anamur (Mersin) | Malazgirt | Niksar (Tokat) |
| Akhisar (Manisa) | Nazilli (Aydın) | Polatlı |
| Develi (Kayseri) | Siverek | Ereğli |
| Elbistan | Lüleburgaz | Erciş (Van) |
| Edremit | Tarsus | Gebze |
| Turhal (Tokat) | | |

**Table 7.** Bayes Net classification results for test set

| Summary for Original Data | | Summary for Smoothed | |
|---|---|---|---|
| Correctly | 86.49 | Correctly | 86.49 |
| Kappa statistic | 0.84 | Kappa statistic | 0.84 |
| Mean absolute | 0.04 | Mean absolute | 0.04 |
| Root mean | 0.18 | Root mean | 0.18 |
| Relative absolute | 16.55 | Relative absolute | 17.99 |
| Root relative | 51.52 | Root relative | 52.52 |
| TP Rate | 0.87 | TP Rate | 0.87 |
| FP Rate | 0.02 | FP Rate | 0.02 |
| Precision | 0.89 | Precision | 0.89 |
| Recall | 0.87 | Recall | 0.87 |
| F-measure | 0.87 | F-measure | 0.87 |
| MCC | 0.85 | MCC | 0.85 |
| ROC | 0.98 | ROC | 0.98 |
| PRC | 0.94 | PRC | 0.94 |
| Time | 0.03 | Time | 0.05 |

5 wrongly classified districts with Bayes Net algorithm are given in Table-8. Considering the mathematical positions of the districts, it can be said that this is due to the fact that the cities are located in the transition climate regions. These results also support the results obtained in Table-5.

**Table 8.** Districts improperly classified by Bayes Net algorithm

| Districts | Observed | Real Region |
|---|---|---|
| Edremit | Aegean | Marmara |
| Fethiye (Muğla) | Mediterranean | Aegean |
| Elbistan | Eastern | Mediterranean |
| Sungurlu | Central | Black Sea |
| Suşehri (Sivas) | Eastern | Central |

## 4. Conclusions

During the last century, many developing countries have been urbanizing rapidly and land cover and land use have been changing. Various ecological, environmental, social and economic effects of these processes emerge. Many ideas put forward to determine the boundaries of cities in geographical and economic terms lose their meaning by being affected by these changes. These dynamically changing and developing dynamics, combined with information technologies, have shifted the use of research in regional and urban areas towards new generation methods. In addition to these, interpretation of and giving meaning to many recorded data with the development of technology has increased the importance of machine learning.

When the studies focused on Turkey are considered, studies conducted to determine its geographic boundaries are limited or insufficient for the analysis of urbanization processes. Moreover, resources for classification analysis

are completely insufficient. The current study contributes to the field by presenting the classification results obtained from the attempt to determine the borders of the regions in Turkey. The proposed dataset within the context of the study was obtained from reliable sources and explains the current state of Turkey.

In this study, which was prepared considering 18 different features, the performances of classification algorithms for the region classification were measured. When the results were evaluated, Bayes Net algorithm with the correct classification rate of 83.9506%, Kappa statistics of 0.81 and calculation time of ~ 0.00 seconds was found to have the highest performance. Parallel to this analysis, the current situation analysis of some rapidly urbanized provinces was made with the Bayes Net algorithm. With an accuracy rate of 86.49%, the districts in the urbanization process have remained connected to the existing regions which they are already connected to. The districts classified incorrectly were located in a different region due to the transition climate and due to changing vegetation and soil type Turkey Regions Map', a visual interpretation for Turkey, was presented in Figure-3. While constructing the map, the results obtained from both provinces and districts data were taken into consideration. Moreover, it is clearly seen in Table-5 that the maritime attribute causes an error in the classification of the provinces. For this reason, provinces such as Denizli, İzmir and Muğla are shown in their own regions. A second reason for this situation is that when the district to gain provincial status leaves the province, the remaining territory boundaries of the province adapt to the existing regional characteristics. When Fethiye, which is a district of Muğla, is removed, the general features of the Muğla region bear the Mediterranean transition climate characteristics more clearly. In this case, Muğla can better reflect the characteristics of the Aegean region. This interpretation actually highlights the interaction of urbanization and geographical boundaries.

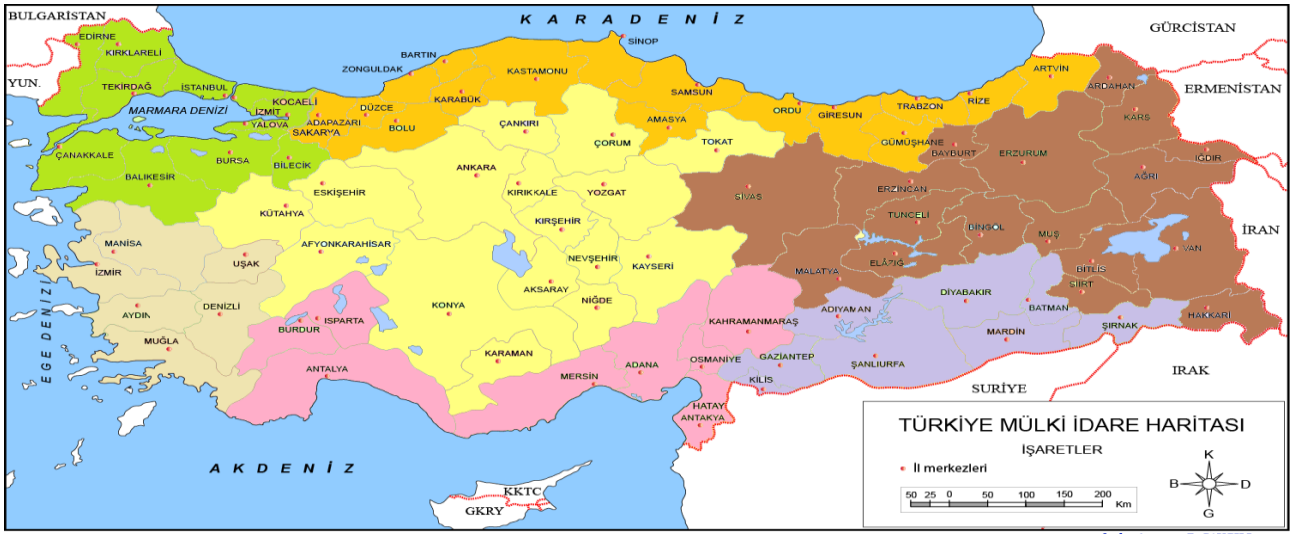


**Figure 3.** Turkey's drawn geographic boundaries according to Bayes Net algorithm results

In this study, especially the use of machine learning method in urban research was emphasized. The applicability of our data set was demonstrated for the determination of urban areas in Turkey, which has large population and a wide variety of land covers. The overall accuracy rate of the classifiers for our data set was found to be approximately 83.95% with Bayes Net algorithm. It measures the success of the methodology we use for spatial assessment in determining the geographical boundaries of urban areas. This assessment is suitable for studies analyzing different regions. Data mining has many applications in different fields. It is thought that this study will contribute to the literature in terms of being an application example in the field of geography and examining the performance of different algorithms. In addition, expanding the approach by adding additional features to classifiers such as socio-economic variables, accuracy rate of classifiers can be increased. The approach can be reconsidered by a new study with other learning algorithms or with various setting parameters of the classifiers.

## References

[1] 1st Geography Congress, "Raporlar, müzakereler, kararlar", *T.C. Maarif Vekilliği*, 1941.

[2] M. Alan, S. Dündar, "Yatırım teşvik verilerinin veri madenciliği ile analizi", *Kırıkkale Üniversitesi Sosyal Bilimler Dergisi*, vol. 7, no. 2, pp. 119-130, 2017.

[3] M. Albayrak, "The use of data mining in scientific research", *International Journal of Social Sciences and Education Research*, vol. 3, no. 3, pp. 751-760, 2017.

[4] K. Arınç, "Coğrafi metodoloji açısından bölgesel coğrafya, bölge bilimi ve coğrafi bölgeler", *Türk Coğrafya Dergisi*, vol. 6, pp. 13-24, 2013.

[5] H. Buhaug, H. Urdal, "An urbanization bomb? Population growth and social disorder in cities", *Glob. Environ. Chang.*, vol. 23, no. 1, pp. 1-10, 2013.

[6] E.L.A. Glaeser, "A world of cities: The causes and consequences of urbanization in poorer countries", *J. Eur. Econ. Assoc*, vol. 12, pp. 1154-1199, 2019.

[7] R. Dash, R. Dash, D. Mishra, "A hybridized rough-PCA approach of attribute reduction for high dimensional data set", *European Journal of Scientific Research*, vol. 44, no. 1, pp. 29-

38, 2010.
[8] Department of Economic and Social Affairs, Population Division, United Nations. *World urbanization prospects: the 2014 revision*, United Nations: New York, NY, USA, 2015.
[9] Eurostat, "The EU in the world – population", 2020.
[10] N. Garipağaoğlu, "Türkiye'de hava kirliliği sorununun coğrafi bölgelere göre dağılımı", *Doğu Coğrafya Dergisi*, vol. 8, pp. 55-77, 2003.
[11] N. Garipağaoğlu, "Türkiye'de kentleşmenin, kent sayısı, kentli nüfus kriterlerine göre incelenmesi ve coğrafi dağılışı", *Marmara Coğrafya Dergisi*, vol. 22, pp. 1-42, 2010.
[12] R. Goldblatt, W. You, G. Hanson, A.K. Khandelwal, "Detecting the boundaries of urban areas in India: A dataset for pixel-based image classification in google earth engine", *Remote Sens*., vol. 8, pp. 634, 2016.
[13] M. Gribskov, N.L. Robinson, "Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching", *Comput Chem*., vol. 20, no. 1, pp. 25-33, 1996.
[14] M. Jones, A. Paasi, "Regional world(s): Advancing the geography of regions", *Regional Studies*, vol. 47, no. 1, pp. 1-5, 2013.
[15] A. Paasi, "Regional planning and the mobilization of `regional identity': from bounded spaces to relational complexity", *Regional Studies*, vol. 47, no. 8, pp. 1206-1219, 2013.
[16] A. Pike, "Whither regional studies?", *Regional Studies*, vol. 41, no. 9, pp. 1143-1148, 2010.
[17] K. Kayaalp, "Asenkron motorlarda veri madenciliği ile hata tespiti", *Süleyman Demirel Üniversitesi*, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Isparta, 2007.
[18] Y. Kırelli, "E-ticaret siteleri için sahtekârlık tespit sistemleri", *İstanbul Ticaret Üniversitesi*, Yüksek Lisans Tezi, Bilgisayar Mühendisliği Ana Bilim Dalı, İstanbul, 2016.
[19] E. Kirkos, C. Spathis, Y. Manolopoulos, "Data Mining techniques for the detection of fraudulent financial statements", *Expert Systems with Applications*, vol. 32, no. 4, pp. 995-1003, 2007.
[20] A. Kubaş, "Tekirdağ ilinde sanayileşme ve çevre yönetimi", *Social Research Journal*, vol. 6, no. 4, pp. 109-114, 2017.
[21] K.C. Seto, B. Güneralp, L.R. Hutyra, "Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools", *Proc. Natl. Acad. Sci. USA*, vol. 109, pp. 16083-16088, 2012.
[22] M.L. McKinney, "Urbanization, Biodiversity, and Conservation: The impacts of urbanization on native species are poorly studied, but educating a highly urbanized human population about these impacts can greatly improve species conservation in all ecosystems", *Bioscience*, vol. 52, pp. 883-890, 2002.
[23] C. Pugh, "Sustainability the environment and urbanisation", Earthscan: New York, NY, USA, 1996.
[24] N. Özalp, S. Aykanat, Z. Erdem, U. Ayan, "Yazar tespiti için iyileştirilmiş Naive Bayesian algoritması", 2013.
[25] A. Özçağlar, "Türkiye'de yapılan bölge ayrımları ve bölge planlama üzerindeki etkileri", *Coğrafî Bilimler Dergisi*, vol. 1, no. 1, pp. 3-18, 2003.
[26] E. Özşahin, H. Sarı, İ. Eroğlu, "Naip Ovası ve yakın çevresindeki taş ocaklarında zamansal ve mekânsal değişimlerin çevresel etkileri (Tekirdağ)", *Yüzüncü Yıl Üniversitesi Tarım Bilimleri Dergisi*, vol. 28, no. 3, pp. 331-344, 2018.
[27] P. Refaeilzadeh, L. Tang, H. Liu, "Encyclopedia of Database Systems", Springer US, 2009.
[28] Ş.E. Şeker, "İstatistiksel Normalleştirme (Statistical Normalisation)", 2012.
[29] United Nations, "Population division publications", 2020.
[30] K.Y., Wu, X.Y. Ye, Z.F. Qi, H. Zhang, "Impacts of land use/land cover change and socioeconomic development on regional ecosystem services: The case of fast-growing Hangzhou metropolitan area, China", *Cities*, vol. 31, pp. 276-284, 2013.
[31] B. Yazıcı, F. Yaslı, H.Y. Gürleyik, U.O. Turgut, M.S. Aktas, O. Kalıpsız, "Veri madenciliğinde özellik seçim tekniklerinin bankacılık verisine uygulanması üzerine araştırma ve karşılaştırmalı uygulama". *9. Ulusal Yazılım Mühendisliği Sempozyumu (UYMS-15)*, Türkiye, 15-17 September, 2015.
[32] A. Yiğit, "Türkiye'nin doğu bölgelerinin yeniden belirlenmesi hakkında düşünceler", *Fırat Üniversitesi Sosyal Bilimler Dergisi*, vol. 8, no. 1, pp. 357-376, 1996.