



The Success of Restricted Ordination Methods in Data Analysis with Variables at Different Scale Levels

Mehmet Tahir HUYUT^{1*} , Sıddık KESKİN² 

^{1*}: Erzincan Binali Yıldırım University, Faculty of Medicine, Biostatistics, Erzincan, Turkey.

²: Yüzüncü Yıl University, Faculty of Medicine, Biostatistics, Van, Turkey.

Geliş / Received: 22/10/2020, Kabul / Accepted: 27/01/2021

Abstract

Events in nature occur with the effect of many interrelated variables, either separately or together. It is important to introduce and use the methods used for the analysis of data sets at different scale levels with linear and nonlinear relationship structure between variables. Redundancy Analysis and Canonical Correspondence Analysis are among the methods used in the analysis of such data. Aforementioned techniques are generally carried out by ecologists and there are limited studies in the field of health. In the study, the application of the methods was performed with a data set in the field of Cardiology including variables at different scale levels and their performances were compared. Determination Coefficient (R^2) and MAPE (Mean Absolute Percentage Error) value were calculated as performance criteria. According to the results, it was seen that CCA and RDA, which analyze the relationship structures between variables in different scale types (cardiological data set), explain the variation sufficiently. Also, it was emphasized that both methods classify well with low MAPE value (less than 10%) and perform ordination diagram. In addition, it has been observed that restricted ordination diagram models give satisfactory results in determining the relationships between coronary heart disease data and so that they can be used in the field of health too.

Keywords: Canonical form, Constrained ordination methods, Ordination diagram, Dimension reduction

Farklı Ölçek Seviyelerindeki Değişkenlerin Analizinde Kısıtlı Ordınasyon Yöntemlerinin Başarısı

Öz

Doğadaki olaylar, birbiriyle ilişkili birçok değişkenin ayrı ayrı veya birlikte etkisiyle meydana gelir. Değişkenler arası doğrusal ve doğrusal olmayan ilişki yapısı ile farklı ölçek düzeylerindeki veri setlerinin analizi için kullanılan yöntemlerin tanıtılması ve kullanılması önemlidir. Bu tür verilerin analizinde kullanılan yöntemler arasında Gereksizlik Analizi ve Kanonik Uyum Analizi yer almaktadır. Yukarıda belirtilen tekniklerle ilgili çalışmaların genellikle ekolojistler tarafından yapıldığı ve sağlık alanında sınırlı sayıda çalışma olduğu görülmüştür. Bu nedenle çalışmada, belirtilen yöntemlerin uygulamaları farklı ölçek düzeylerindeki değişkenleri içeren bir Kardiyoloji veri seti üzerinde gerçekleştirilmiş ve performansları karşılaştırılmıştır. Performans kriteri olarak Belirleme Katsayısı (R^2) ve MAPE (Ortalama Mutlak Yüzde Hata) değeri hesaplandı. Elde edilen sonuçlara göre, farklı ölçek türlerindeki (kardiyolojik veri seti) değişkenler arasındaki ilişki yapılarını inceleyen CCA ve RDA'nın, varyasyonu yeterince açıkladığı görülmüştür. Ayrıca her iki yöntemin de düşük MAPE değeri (% 10'dan az) ile iyi sınıflandırıldığı ve koordinasyon diyagramı gerçekleştirdiği vurgulanmıştır. Ayrıca, kısıtlı koordinasyon diyagram modellerinin koroner kalp hastalığı verileri arasındaki ilişkilerin belirlenmesinde tatmin edici sonuçlar verdiği ve benzeri sağlık alanlarında da kullanılabilirliği görülmüştür.

Anahtar Kelimeler: Kanonik form, Kısıtlı ordınasyon metodu, Ordınasyon diyagramı, Boyut indirgeme

1. INTRODUCTION

Events in nature occur with the effects of many variables that are related to each other, separately or together. Scientific studies generally examine the relationship structure between the variable of interest and other variables thought to affect this variable under two headings: linear or non-linear relations.

One of the common problems in research is to determine the relationship structure between a large number of response variables and explanatory (environmental) variables with high accuracy. Any answer variable (existence of objects, type composition, etc.) can vary depending on the explanatory variables. Therefore, it is very important to determine the relationship between response variables (frequency of the presence of diseases or species) and explanatory (environmental) variables in different locations (settlements or regions).

In analyzing the data of this structure; Previously, it was generally assumed that the relationships between variables were linear or it was preferred to make separate multiple regression analyzes for the response variables. However, instead of these approaches, new methods that deal with the relationships together were proposed. Regression and ordination methods were added to multivariate direct gradient analysis algorithms and these were called canonical (or constrained) ordination methods. The most common canonical ordination methods are Redundancy Analysis (RDA) and Canonical Correspondence Analysis (CCA). Ordination methods other than these methods are known as indirect gradient analysis.

RDA is an analysis method used to explain and summarize variation in more than one response variable with more than one explanatory variable. In other words, RDA is a direct gradient analysis method that summarizes the linear relationships between a set of explanatory variables and the components of the response variables.

CCA is one of the multivariate analysis methods developed to determine the relationships between the existence of diseases, objects or species taken as response variables and the explanatory variables thought to be effective on them (Ter Braak, 1986).

It has been observed that the studies on Direct Gradient Analysis (DGA) are generally carried out by ecologists and environmental scientists, and there are limited studies on the applications of these analyzes in the field of health. The performance of the methods was evaluated by applying the methods with a real data set in the field of cardiology, which includes the data of variables at different scale levels. Thus, in the study, CCA and RDA were examined and their usability in the field of health was shown.

2. ORDINATION METHODS

In scientific studies, the characteristics considered are generally under the influence of many factors. For this reason, it is important to examine the variables together in order to obtain valid and reliable results. There are appropriate analysis methods according to whether the variables considered in the studies are continuous, categorical or sequential variables and

according to the models that will be directly associated with the explanatory variables. RDA and CCA, which are called restricted ordination methods, can be counted among these methods.

Ordination methods are used to show or represent a lower dimensional variation of a data matrix (Palmer, 2017). These methods are not affected by the exogenous variables included in the model after the calculations, and also allow to explain the data matrix without limiting itself. Ordination techniques based on linear (PCA-principal component analysis and RDA-Redundancy analysis) and weighted average (UA-correspondence analysis and CCA-canonic correspondence analysis) represent ensemble data in drastically different ways.

Ter Braak (1986) stated that a multivariate statistical analysis method (based on a common response model) is necessary to determine how the composition of the community varies with environmental variables (especially when the number of environmental variables exceeds two or three).

Ter Braak and Prentice (1988) stated that unlike regression and calibration, ordination problem requires simultaneous estimation of many parameters and has some limitations in practice. They stated that the number of ordination axes to be removed should be small and the response model should be restricted to allow a solution. They also emphasized that ordination techniques are robust against deviations from simple models.

2. 1. Triplot

Triplot adapts the presence of each response variable in the area defined by the coordination axes to the plane. The scheme in which points and response variables are marked with arrows is called a "triplot" (Gabriel, 1971). The arrows show the direction of the maximum variation in the presence of the response variable, and their length varies in proportion to the maximum rate of change. Ter Braak and Prentice (1988) stated that variables moving away from the origin in the diagram have an increased contribution to variation, while variables located close to the origin make less contribution to variation.

In the resulting diagram, three items are used to describe the relationships between data. These; sample points are points of response variables and environmental vectors. The position of a sample point in the diagram is determined by both the total composition of the species (ie the number of each species collected in the sample) and the environmental conditions at the time of sampling. If two objects differ in species composition or environmental conditions, they will appear in different places on the ordination diagram, but if they are similar, they will appear close to each other (O'Connell et al., 2004).

2. 2. Direct Gradient Analysis (Restricted Ordination)

The constrained ordination is directly related to the matrix containing response variables and explanatory variables. Shows the stages of ordination in both matrices. In this approach, ordination and multiple regression methods are combined (Borcard, 2006). The constrained ordination method includes all

response variables in the analysis at the same time (Legendre and Legendre, 1998).

Ter Braak and Prentice (1988) stated that constrained ordination axes can be thought of as hidden variables or hypothetical environmental variables constructed to optimize the fit of a particular statistical model (linear or unimodal) of how the abundance of response variables changes along gradients.

2. 3. Redundancy Analysis (RDA)

Israels (1992) accepted the RDA algorithm as the process of calculating a series of simultaneous regression equations for variables in the Y set. He also stated that the RDA uses only linear combination functions of explanatory variables (X), and therefore pointed out that the method can also be considered as reduced order regression.

RDA is a method used to explain and summarize the variation in the response (response) variable set using the explanatory variable set.

In other words, RDA is a direct gradient analysis method that summarizes the linear relationships between the explanatory variable set and the components of the "redundant" response variables. RDA extends Multiple linear regression by allowing regression of multiple response variables with multiple explanatory variables. A matrix of the regulated values of all response variables produced through multiple linear regression is then subjected to Principal Component Analysis.

2. 3. 1. Relationship between redundancy analysis (RDA) and canonical correlation analysis (CCorA)

Canonical axes in CCorA maximize the correlation between linear combinations of two sets of Y and X variables. For this, group covariance (or correlation) between variables is maximized instead of within-variables group covariance (or correlation). However, in RDA, the variation in Y variable set is tried to be explained at maximum level with linear functions obtained from X variable set. In the canonical correlation analysis, the correlation coefficients obtained do not provide information about the variation of the variable sets. This situation creates a disadvantage for Canonical correlation analysis. In this case, redundancy analysis is proposed.

2. 3. 2. Relationship between redundancy analysis (RDA) and principal component analysis (PCA)

In Principal Components Analysis (PCA), the components obtained according to their contribution to the variation in the variable set show the number of dimensions. Thus, with the help of independent components, the number of variables in the set of variables can be reduced. In the Redundancy Analysis, since the redundancy determination index is arranged with the help of the created components, the components for which this index has the maximum value show the number of dimensions.

2. 4. Canonical Correspondence Analysis (CCA)

Relationships between sets of variables are generally asymmetric, and two main goals are pursued in analyzing these variable sets:

The first is to determine which dependent variable/s or their combinations are the most. The second is to predict the outcome variables that most affect explanatory variables and their combinations. Multivariate Linear Methods (methods such as multivariate regression or multivariate analysis of variance (MANOVA)) can only achieve the second purpose. On the other hand, multivariate methods (such as canonical correlation analysis) that do not take into account the asymmetric relationship between explanatory and outcome variable groups cannot achieve the second goal (Mardia et al., 1979). However, one of the methods that achieve the two goals described above is Canonical Correspondence Analysis. (Ter Braak, 1986; Saporta, 1990). CCA combines Multivariate Regression, MANOVA and PCA methods.

CCA, developed as an extension of correspondence analysis, predicts the unimodal responses of response variables to environmental gradients (Makarenkov and Legendre, 2002). It has a very explanatory power about the relationship between the answer variable and environmental or explanatory variables (Oksanen, 2004).

While deciding whether the CCA application is appropriate or not; It is necessary to question the relationship structure between the measured environmental variables and the variables that emerge with only the response variable, or whether the reasonable responses (responses) of the species are unimodal under the reasonable conditions of the environment (Figure 1) (Mc Cune and Grace, 2002).

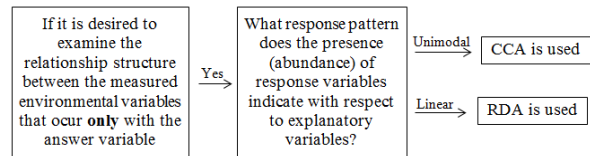


Figure 1. Decision tree for the use of CCA and RDA in community data (variables emerging together with the answer variable and data belonging to explanatory variables) (Mc Cune and Grace, 2002).

3. MATERIAL AND METHOD

3. 1. Material

In the study, real data obtained from the free access

"<https://archive.ics.uci.edu/ml/datasets.html>" (Access date: 05/09/2020) was used as the application material.

The data set related to the field of cardiology includes 270 individuals and 13 characteristics (variables) in different variable types (categorical, numerical and ordinal). The properties of the original variables in the data set are summarized below.

- 1- Age, 2- Gender, 3- Type of chest pain (a: Typical angina, b: Atypical angina, c: Non-anginal pain, d: Asymptomatic), 4- Resting blood pressure, 5- Serum cholesterol amount (mg / dl), 6- Fasting blood glucose level > 120 mg / dl, 7- Electrocardiographic (EKG) results during rest (a: ST-T segment abnormality, b: Left ventricular hypertrophy sign, c: EKG result is normal), 8- Maximum heart rate, 9- Pain formation during exercise, 10- ST segment according to rest whether there is collapse during exercise, 11- Shape of exercise induced ST segment (a: Upward, b: Straight, c: Downward), 12- Number of open main vessels in the heart (a: One vessel closed, b: Two vessels closed c: Three vessels closed d: Three vessels open), 13-

Thallium scintigraphy result (a: Normal b: Fixed defect c: Recyclable)

Variables 1, 4, 5 and 8 in the study are continuous; The variable number 11 is ordinal; Variables 2, 6, 9 and 10 are binary; Variables numbered 3, 7, 12 and 13 have a categorical scale structure.

3. 2. Method

A) Redundancy Analysis (RDA)

The redundancy determination index was defined as a measure of how much of the average variance value in a variable set is explained with other sets of variables using the Canonical Analysis of Variance Method. In other words, the variation in the Y variable set is tried to be explained with the linear functions obtained from the X variable set. (Lambert et al., 1988).

Calculation steps of redundancy analysis

The following algebraic steps explain how to achieve Redundancy Analysis equation (Equation 8) through Multiple Regression and Principal Components Analysis.

- i) Regression of each variable of Y on all variables in X is taken, their values are arranged and residual values are calculated.
- ii) Principal Component Analysis is performed on the matrix of predicted (fitted) values for calculating eigenvalues and eigenvectors.
- iii) As a result, the following two ordinations take place.

In the first ordination, the ordination of the explanatory variables in two-dimensional space is obtained. With this ordination, canonical axes called "linear combinations of environmental variables" are obtained and site scores are calculated based on these axes.

In the second ordination, the ordination of dependent variables at sites is obtained. This ordination gives "sample scores" or "answer variable scores". (Legendre and Legendre, 1998; Borcard, 2006).

First step: For each answer variable in Y, multiple linear regression is calculated with all variables in X. Equality of regression coefficients for each Y variable:

$$b = [X'X]^{-1}X'Y$$

is calculated. This equality is expressed by matrix notation for all Y variables as

$$B = [X'X]^{-1}X' \quad (1)$$

In Equation (1), B is the matrix of regression coefficients for the explanatory variables for all answer (Y) variables. In multiple regression, the predicted (fitted) \hat{Y} values are obtained with the equation

$$\hat{Y} = XB \quad (2)$$

All estimated (fitted) values (\hat{Y}) set can be calculated by a single matrix operation as shown in Equation (2). When Equation (1) is written in place in Equation (2), the estimated (fitted) set of values (\hat{Y}) is calculated as

$$\hat{Y} = X[X'X]^{-1}X'Y \quad (3)$$

(since the variables X and Y are centralized, B vectors have no regression constant parameters). As in standart multiple regression, \hat{Y} vectors are usually centralized and when $m = n$, X becomes a squared matrix. In this case, multiple regressions fully explain the variables in the Y matrix and $\hat{Y} = Y$ (Legendre and Legendre, 1998).

Second step: A suitable covariance matrix corresponding to the predicted (fitted) \hat{Y} values; It is obtained by the equation

$$S_{\hat{Y}'\hat{Y}} = [1/(n - 1)]\hat{Y}'\hat{Y} \quad (4)$$

When the value of in Equation (3) is replaced in Equation (4), covariance-matrix; It is calculated by the equation

$$S_{\hat{Y}'\hat{Y}} = \left[\frac{1}{n - 1} \right] Y'X[X'X]^{-1}X'X[X'X]^{-1}X'Y \quad (5)$$

By arranging Equation (5), the covariance matrix is obtained as

$$S_{\hat{Y}'\hat{Y}} = S_{YX}S_{XX}^{-1}S'_{XX} \quad (6)$$

In Equation (6), S_{YY} shows (pxp) dimensional covariance matrix between response variables, S_{XX} shows mxm dimensional covariance matrix between multivariate regressions, S_{YX} shows a covariance matrix with pxm size between two sets of variables ($S'_{YX} = S_{XY}$ (m xp) is a dimensional matrix). If the Y variables are standardized, the equation $R_{YX}R_{XX}^{-1}R'_{YX}$, becomes the multiple determination coefficient for Equation (6) (Legendre and Legendre, 1998).

Third Step: The predicted (fitted) \hat{Y} values are subjected to Principal Components Analysis to reduce the size of the solution. In this analysis, equality to solve the eigenvalue problem; It is expressed as

$$(S_{\hat{Y}'\hat{Y}} - \lambda_k I)u_k = 0 \quad (7)$$

Equation (7) is converted to equation

$$(S_{YX}S_{XX}^{-1}S'_{YX} - \lambda_k I)u_k = 0 \quad (8)$$

by using Equation (6). The obtained Equation (8) gives the Redundancy Analysis equation

B) Canonical Correspondence Analysis (CCA)

CCA aims to visualize a model of community variation and the basic properties of the distribution of response variables across environmental variables (Ter Braak, 1987). The Maximum Likelihood (ML) Method is used to obtain estimates in the analysis. Although the solution of CCA is usually obtained with the weighted average algorithm, the method is basically an eigenvalue analysis and therefore the solution can be provided with any eigenvalue algorithm (Ter Braak 1986, 1987). The statistical model in the basic approach of CCA expresses the presence or frequency of the response variable along the circumferential gradients as a unimodal position function. CCA; it is a Gaussian regression approach under certain assumptions and it is strong against failure of these assumptions (Ter Braak and Prentice, 1988).

In calculating the site or location scores obtained from the weighted average of the answer variables; site scores show dependent variables (response variables), while environmental variables show independent variables (explanatory variables). The new site scores correspond to the values predicted by the regression equation (Palmer, 1993).

Multiple regressions or correlation coefficients of site scores on environmental variables are calculated as follows (Ter Braak, 1986).

$$x_i = b_0 + \sum_{j=1}^q b_j z_{ij} \quad (9)$$

With Equation (9), the ordination axis is graphically associated with the explanatory (environmental) variables, and in Equation

(9) b_0 is the constant term and b_j is the regression coefficient for the j th environmental variable. Keeping x_i and u_k constant, the regression coefficient b_j is estimated. Therefore, the response variable data are indirectly related to environmental variables through the coordination axis. The transition equations from Gaussian Ordination to Correspondence Analysis can be summarized as follows.

$$\lambda u_k = \sum_{i=1}^n \frac{y_{ik} x_i}{y_{+k}} \quad (10)$$

$$x_i^* = \sum_{k=1}^m \frac{y_{ik} u_k}{y_{i+}} \quad (11)$$

$$\mathbf{b} = (\mathbf{Z}'\mathbf{R}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{R}\mathbf{x}^* \quad (12)$$

$$\mathbf{x} = \mathbf{Z}\mathbf{b} \quad (13)$$

In Equation (10 and 11); u_k is the weighted average of the k th (except m) answer variables; x_i is the value of the i th (except n) site (eg humidity value); y_{ik} represents the abundance (frequency) of response variables k in site i ; y_{+k} indicates the total abundance of the answer variable k and y_{i+} indicates the total sites.

\mathbf{R} is $n \times n$ diagonal matrix consisting of y_{i+} . $\mathbf{Z} = \{Z_{ij}\}$ is an $n \times (q + 1)$ dimensional matrix containing environmental data. \mathbf{b} , \mathbf{x} and \mathbf{x}^* are column vectors. $\mathbf{b} = (b_0, b_1, \dots, b_q)'$, $\mathbf{x} = (x_0, x_1, \dots, x_n)'$ and $\mathbf{x}^* = (x_1^*, \dots, x_n^*)'$.

The CCA algorithm is solved by applying the following steps (iteration algorithm) related to weighted average and multiple regression.

1) It starts with random but unequal initial site scores (x_i values in Equation (9)).

2) Answer variable scores are calculated according to the weighted averages of the site

scores (u_k values are calculated by taking $\lambda = 1$ in Equation (10)).

3) New site scores are calculated by taking the weighted average of the answer variable scores (x_i^* in Equation (11): new site scores are calculated with the weighted average).

4) Regression coefficients are obtained by weighted multiple regression of site scores on environmental variables (\mathbf{b} values in Equation (12) are obtained).

5) New site scores are calculated by Equation (13) or Equation (9) which are equal to each other. The new site scores are actually the corrected regression coefficients of the previous step.

6) Site values are centralized and standardized with the equation

$$\sum_{i=1}^n y_{i+} x_i = 0 \text{ and } \sum_{i=1}^n y_{i+} x_i^2 = 1 \quad (14)$$

7) The iteration process stops when the new site scores converge sufficiently to the previous iteration site scores. Otherwise, the iteration repeats from the second step (Ter Braak, 1986).

The significance level of the axes (eigenvalues) in RDA and CCA is realized in two stages:

Stage 1: The first stage is the test of independence (H_0) between answer and explanation variable sets.

For this, first the sum of all canonical eigenvalues is tested using the pseudo F ratio. Alternative hypothesis (H_1), which is the sum of all canonical eigenvalues; It is tested for the explanation of the variation in

the response variable using permutation of raw data or permutation of residuals. The sum of all canonical eigenvalues is divided by the total variation in the answer data matrix. The obtained R^2 result indicates the part of the variation in Y explained by X (similar to R^2 in multiple regression) (Legendre and Legendre, 1998).

Stage 2: In this stage, the significance of canonical eigenvalues is tested one by one.

At this stage, an attempt is made to identify more axes that explain (separately) explain the canonical variation. After this test, the important axes to be used in the coordination diagram are decided. While the null hypothesis (H_0) expresses the independence between Y and X; The alternative hypothesis (H_1) states that the variation in Y is explained using permutation of raw data or permutation of residues (Legendre and Legendre, 1998).

C) Performance Evaluation of Methods

In the study, R^2 (determination coefficient) and MAPE (Mean Absolute Percent Error) were used to determine the success and performance of the methods in prediction (Anonymous 1, 2019). MAPE is calculated using the equation

$$MAPE = \frac{100 (\%) }{n} \quad (15)$$

In Equation (15); y is the real value, \hat{y} is the predicted value and n is the number of observations. MAPE is a statistical measure of how accurate a prediction system is and measures this accuracy as a percentage. This criterion is resistant to the effects of outliers thanks to the use of absolute values. It is stated that MAPE is an error criterion that gives good results when there are no extreme

values in the data (also does not contain zero value) (Anonymous 2, 2019).

Witt and Witt (1992) classified the prediction models with MAPE values below 10% as "high accuracy" and models with 10% to 20% correct estimates.

Similarly, Lewis (1982) classified the models with a MAPE value of less than 10% as "very good", between 10% and 20% as "good", between 20% and 50% as "acceptable" and above 50% as "wrong and erroneous".

The determination coefficient (R^2) of the methods was calculated with the following equation (Anonymous 2, 2019)

$$R^2 = 1 - \frac{SS \text{ Error}}{SS \text{ Total}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (16)$$

In Equation (16) SS Error, the sum of error squares; SS Total, the general sum of squares; y_i , the real value of the i th answer variable; \hat{y}_i , the predicted value of the i th answer variable and \bar{y} represents the mean of the actual data.

With the CCA method an application was made according to the answer variable (location, number of sites), object (categorical variables) and explanatory variables (continuous variable) (Table 1).

Table 1. CCA implementation according to the scale types of response variable (site), object and explanatory variables

Answer Variable (Number of Locations)	Objects (Categorical variables)	Explanatory variables (Continuous variables)
OMNV (4 categories)	8 Categorical Variables (TPC, SR, R-EKG, Sex, PDE, Diabetes, ST-C, ST-İ)	2 Continuous Variables (MNH and Cholesterol)

ST-C: Collapse of the ST segment, TPC: Type of pain in the chest, ST-İ: the inclination of the ST segment, PDE: Presence of angina during exercise, MNH: Maximum number of heartbeat, R-EKG: Elektrokardiyogram result at rest, OMNV: Open main number of vessels, SR: Scintigraphy result.

With the RDA method, an application was made according to the number of sites (categorical variable), the number of response variables (continuous variables) and the scale types of explanatory variables (Table 2).

Table 2. RDA application according to the number of sites and response variables and the scale types of explanatory variables

Number of sites (regions where the answer variable occurs)	Explanatory variables	Response variables (continuous variables)
OMNV (4 categories)	8 Categorical Variables (TPC, SR, EKG-R, Sex, PDE, Diabetes, ST-C, ST-İ) and 2 Continuous Variables (BP-R, Age)	2 Continuous Variables (MNH, Cholesterol)

ST-C: Collapse of the ST segment, TPC: Type of pain in the chest, ST-İ: the inclination of the ST segment, PDE: Presence of angina during exercise, MNH:

Maximum heartbeat, EKG-R: Elektrokardiyogram result at rest, OMNV: Open main number of vessels, SR: Scintigraphy result, BP-R: Blood pressure at rest

XLSTAT (version: 2018. 5) package program was used for the applications of CCA and RDA methods used in the analysis of the data obtained.

4. RESULTS

The eigenvalue and variance explanation rates of the basic coordinates obtained as a result of CCA and RDA application are summarized in Table 3.

Table 3. Eigenvalues of axes and explanation rates of total variation as a result of CCA and RDA application

Applied Analysis	Eigenvalue of the 1st axis / explained variance rate (%)	Eigenvalue of the 2nd axis / explained variance rate (%)	Total variance explained by the constrained axes (%)
CCA	0.036 (86.934)	0.005 (13.066)	100.00
RDA	0.355 (77.493)	0.34 (22.507)	100.00

When Table 3 is examined, it is seen that the first axis of the diagram obtained as a result of CCA application explains a large part (86.934%) of the variation in variables. Together with the second axis, the CCA diagram can explain 100% of the total variation. Likewise, the first axis of the diagram obtained as a result of RDA application explained 77.493% of the variance of the answer variable. With the second axis, 100% of the total variance was explained. In CCA and RDA methods, there was no unexplained variance in total, since the axes were constrained as a linear combination of explanatory variables.

According to the rate of explanation of the total variation of the axes, it can be said that two-dimensional CCA and RDA triplots (diagram) may be sufficient in analyzing the relationships between the response variable, locations (regions where variables occur with the response variable), objects and explanatory variables.

The relationship model between the answer variable and the explanatory variables was examined by permutation test. First, the (pseudo) F value was obtained for the validity of the model and the hypothesis was checked according to the critical value obtained from the 500-repeat permutation test. The result obtained accordingly is given in Table 4.

Table 4. Permutation test result of raw data in CCA and RDA application

Permutation test result	CCA	RDA
Permutation number	500	500
pseudo F	1168.00	9.805
p-value	0.748	0.0001
α -value	0.050	0.050

H_0 hypothesis is accepted according to the p value ($0.748 > 0.05$) calculated as a result of the Permutation test applied for CCA. Accordingly, as a result of the CCA application, it is understood that there is no linear relationship between the answer variable and the explanatory variables, and that there are independent variables (it is assumed that there is a unimodal relationship between the response variable and the explanatory variables). When looking at the p value ($0.0001 < 0.05$) obtained for RDA as a result of the permutation test, the H_0 hypothesis is rejected. Accordingly, as a result of RDA application, it is understood that there is a linear relationship between the

answer variable and the explanatory variables.

According to the permutation test results, the restricted ordination axes of CCA and RDA were found to be significant. Accordingly, constrained ordination axes were constructed as linear combinations of explanatory variables.

The triplot representation of response variable, objects and explanatory variables in two-dimensional space obtained by CCA ordination is given in Figure 2. CCA ordination shows the relationships between locations (sites), objects (species or individuals) and explanatory (environmental) variables on the same diagram.

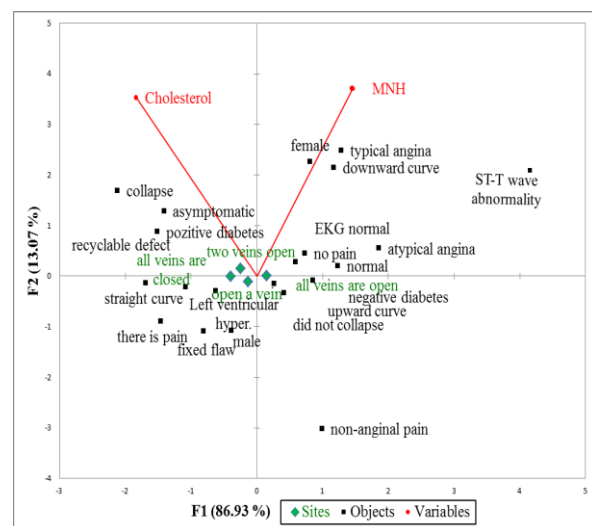


Figure 2. Triplot diagram of CCA application

When Figure 2 is examined; By looking at the length of the explanatory variables, it can be said that the cholesterol variable is more important than the maximum heart rate explanatory variable. It is seen that individuals with "open three vessels" have low cholesterol levels, and individuals with "two and one open vessels" have higher cholesterol levels. Typical chest pain was associated with a high maximum heart rate, and asymptomatic chest pain with moderate

and high cholesterol levels. In addition, the variable of "atypical angina" chest pain was found to be highly correlated with the variables of "upward" ST segment curve and "normal" EKG results. In addition, it can be said that individuals who have "pain" during exercise have moderate cholesterol levels. Individuals with an "upward" ST segment curve and those with "no pain" during exercise were found to have moderate heart rate.

Also; It can be said that there is a high correlation between the variables of "left ventricular hypertrophy sign", "flat slope" of the ST segment, and "fixed defect" in the scintigraphy result. The "collapse" state of the ST segment was more common in medium and high cholesterol values. In addition, the category of "abnormality" in the ST segment was more intense at the high maximum heart rate level. It was determined that the category of "no pain" appeared more intensely at low values of the explanatory variables of cholesterol and maximum heart rate. A high correlation was found between the "reversible defect" category and the "fixed defect" category of the scintigraphy result.

It was found that individuals with "downward slope" of the ST segment generally had "left ventricular hypertrophy symptom", "asymptomatic" chest pain, "abnormal" ST segments, but did not experience "pain" during exercise. In addition, it can be said that the ST segment slope is generally "downward" in "female" individuals and these individuals often experience "typical angina" chest pain. In addition, individuals with moderate and low cholesterol levels were found to be scintigraphy results as "fixed defect" and "reversible defect" and generally experienced

"asymptomatic" chest pain. In addition, diabetes was less common in individuals with "all vessels open" and there was no ST segment depression in these individuals.

Individuals with moderate and low maximum heartbeat and cholesterol levels had "left ventricular hypertrophy sign" and "diabetes" was more common at these levels. Besides, the cholesterol explanatory variable; A high positive association was found with the categories of "collapse" of the ST segment, "positive diabetes" and "asymptomatic" chest pain. It was observed that individuals whose ST segment did not "collapse" had a low heart rate. In addition, the ST segments of these individuals were found to be "upward" and scintigraphy results to be "normal".

The EKG-R results of individuals with 'downward' ST segment were found to be highly correlated with the maximum heart rate explanatory variable. "Typical-angina" and "atypical angina" types of chest pain are generally found to be more associated with maximum heart rate.

The triplot representation of the locations, responses and explanatory variables in two-dimensional space obtained by RDA ordination is given in Figure 3.

When Figure 3 is examined; By looking at the length of the explanatory variables, it can be said that the "resting blood pressure" variable is more important than the "age" explanatory variable. The "cholesterol" response variable was found to have a higher positive relationship with the "resting blood pressure" explanatory variable than the "age" explanatory variable.

"Non-angina" and "asymptomatic" chest pain categories were found to be positively associated with moderate and low cholesterol

and heartbeat categories. In addition, it can be said that the “blood pressure at rest” explanatory variable is positively correlated with the categorical explanatory variables in the first region. A low correlation was found between the "maximum heart rate" response variable and the "cholesterol" response variable. In addition, it can be said that the "cholesterol" response variable is highly correlated with the explanatory variables in the first region. A high correlation was observed between "age" and "blood pressure at rest" explanatory variables and other explanatory variables. It was found that the “asymptotic” chest pain category was more common in middle and low-age individuals and was highly correlated with categorical explanatory variables in the third region. It has been determined that the categorical explanatory variables in the third region are more common in individuals with “open veins”.

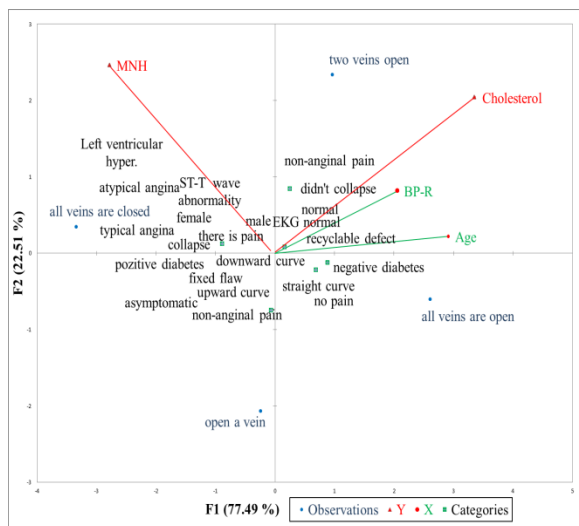


Figure 3. Triplot diagram of RDA application

In addition, it was found that "individuals with open two vessels" were more intense at medium and high "cholesterol" and "blood pressure" levels. In addition, individuals with "closed all vessels" had a high correlation with the "maximum heart rate" response

variable and categorical explanatory variables in the second region. “

A positive correlation was found between the individuals with "open all vessels" and the categorical explanatory variables in the fourth region and it can be said that these individuals generally have low heart rates. Individuals with "left ventricular hypertrophy" symptoms and "ST wave abnormalities" were found to be more concentrated in the middle levels of the "maximum heart rate" response variable. In addition, it can be said that individuals whose ST segment slope is "downward" has a high rate of positive correlation with the "reversible defect" variable of the scintigraphy result.

In addition, it was observed that individuals with “open vein” were highly correlated with categorical explanatory variables in the fourth region. However, this category was found to be the variable that contributed the least to variation.

MAPE (Mean Absolute Percent Error) and R^2 values for the methods are given in Table 5.

Table 5. Assessment criteria for restricted ordination methods

Method	R^2 (Coefficient of Determination)	MAPE (Mean Absolute Percent Error)	n (Observation Number)
CCA	1	0.04	270
RDA	0.94	0.06	270

As seen in Table 5, it can be said that the variance explanation rates of the methods are quite high. While the explanation rate of the explanatory variables for the total variation in the CCA ordination diagram was found to be 100%; In RDA ordination, this rate was found to be 94%. According to the

coefficient of determination, it can be said that constrained ordination methods are sufficient to explain the variation in order to analyze the different relationship structures between the response variable, objects, and explanatory variables.

In the study, while the MAPE value of CCA was calculated as 0.04 (4%), the MAPE value of RDA was calculated as 0.06 (6%). Accordingly, in the constrained ordination diagram obtained by the CCA method, the error in determining the different relationship structures between variables was 4%, while this error was 6% in the RDA method. Accordingly, it was observed that the error was close to each other and less than 10% in both restricted ordination methods.

5. DISCUSSION AND CONCLUSION

In the study, the pseudo F values of the methods were calculated and the critical value was obtained by performing 500-repeat permutation test. Hypotheses were tested according to pseudo F values. Accordingly, in CCA application, it was found that there was no linear relationship between response and explanatory variables (unimodal relationship was assumed), whereas in RDA application, a linear (linear) relationship was found between response and explanatory variables. Similarly, Muller (1981); Ramette (2007); Buttigieg and Ramette (2014) stated that the significance value for RDA / CCA axes can be determined by the permutation test. They stated that when the overall solution is significant, the significance check of axes or explanatory variables can be done by permutation test.

It has been observed that the restricted ordination methods explain the variation adequately (at least 94%). Accordingly, it can be said that the relationship estimates

made with ordination diagram models produced by restricted ordination methods are reliable, consistent and largely explain the total variation.

Muller (1981), Ter Braak (1994) and Ramette (2007) have stated that RDA produces an ordination that summarizes the basic change patterns in the response matrix that can be explained by a matrix of explanatory variables. In addition, they emphasized that the total variation was calculated by dividing the data set into constrained and unrestricted variances, and stated that this process showed how much of the variation in the explanatory variables was unnecessary compared to the variation in the response variables. Besides, Muller (1981); Ramette (2007); Buttigieg and Ramette (2014) stated that examining the non-canonical (unconstrained) vectors of an RDA solution by ordination and correlation provides information about the behavior of unnecessary variables.

The MAPE value for CCA and RDA methods was found below 10%. According to the MAPE criteria, it can be said that the aforementioned methods classify well and perform the ordination diagram with a low error value.

In the study, although there was a high correlation between the original variables, a meaningful second axis was obtained in practice. This result supports Ter Braak's (1987) statement that "CCA ordination diagram is not affected by high correlation between species or between environmental variables". Similarly, Palmer (1993) stated that if highly correlated variables in CCA are included in the analysis, it will not prevent obtaining a meaningful second axis. However, Ter Braak and Prentice (1988)

pointed out that when the explanatory variables are strongly correlated with each other (in practice, when the number of explanatory variables converges to the number of response variables), it is not possible to decompose the effects of different explanatory variables on total variance, and in this case, the canonical coefficients become unstable.

Ter Braak (1987) stated that when the response variable shows unimodal response function, CCorA and RDA are less successful than CCA. Palmer (1993) stated that since the significance tests in CCA are not dependent on parametric distribution assumptions, it is not necessary to transform the original data to make the variables suitable for any distribution.

In addition, Palmer (1993) stated that CCA performs well in distorted species distributions, so the method will be affected minimally by the transformation of environmental data.

Tso (1981) and Jongman et al. (1987) pointed out that although there are similarities between CCorA and RDA, the assumptions regarding the error component may differ. They emphasized that the errors in RDA are uncorrelated and have equal variance, but the errors in CCorA (Canonical Correlation Analysis) show correlated and normal distribution.

In RDA, it can be said that continuous explanatory variables contribute more to the variation, but when the explanatory variables consist of different scale levels, a decrease in the variance rate explained by the first axis was observed. However, in RDA, it can be said that the numerical increase in the answer variable generally decreases the variance rate explained for the first constrained axis. On

the other hand, it is thought that the number of explanatory variables more than 1 in RDA will make the restriction process more meaningful. In addition, in order to obtain more accurate RDA results, the number of sites should be at least 2 more than the number of answer variables, and the number of answer variables should be at least 2.

In CCA, it has been observed that the more the number of categories of the answer variable is greater than the number of explanatory variables, the higher the amount of variance explained by the first restricted axis. In RDA, it has been observed that the more the number of explanatory variables is higher than the number of answer variables, the higher the amount of variance explained by the first restricted axis.

As a result, it was seen that restricted ordination methods (CCA and RDA) that analyze the relationship structures between variables (response variable, objects and explanatory variables) in different scale types at the same time classify well with low MAPE values and explain the variation sufficiently. Thus, it was determined that restricted ordination diagrams can also be used in the field of health.

REFERENCES

- Anonymous 1 (2019). Access address: https://www.researchgate.net/publication/262980567_Root_mean_square_error_RMS_E_or_mean_absolute_error_MAE. Date of access: 12. 04. 2019
- Anonymous 2 (2019). Access address: <https://www.statisticshowto.datasciencecentral.com/mean-absolute-percentage-error-mape/>. Date of access: 12. 04. 2019

- Borcard, D. Université Laval Multivariate analysis-February (2006). Access address: http://ubio.oinfo.cnio.es/Cursos/CEU_MDA07_practicals/Further%20reading/Multivariate%20analysis%20Borcard%202006/Chap_4b.pdf. Date of access: 18/07/2018.
- Buttigieg, P. L. and Ramette, A. (2014). A Guide to Statistical Analysis in Microbial Ecology: a community-focused. living review of multivariate data analyses. *FEMS Microbiol Ecol.* 90(1), 543–50.
- Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58(1), 453–67.
- Israels, AZ. Redundancy Analysis for Various Types of Variables. *Statistica Applicata*, 4(4): 531-42, (1992).
- Jongman, R. H. G., Ter Braak, C. J. F. and Van Tongeren, O. R. (1987). *Data Analysis in Community and Landscape Ecology*. Pudoc. Wageningen.
- Lambert, Z., Wildt, A. R. and Durand, R.M. (1988). Redundancy analysis: An alternative to canonical correlation and multivariate multiple regression in exploring interspecies associations. *Psychological Bull.* 104(2), 282-9.
- Legendre, P. and Legendre, L. (1988) *Numerical Ecology*. Second English Edition. Elsevier Science B.V. Amsterdam. Netherlands. ISBN 978-0444892508.
- Lewis, C.D. (1982). *Industrial and Business Forecasting Methods*. Londra: Butterworths Publishing.
- Mardia, K.V. Kent, J.T. and Bibby, J. M. (1979). *Multivariate analysis*. Academic Press. London. England.
- McCune, B. and Grace, J. B. (2002). *Analysis of Ecological Communities*. MjM Software Design. Gleneden Beach. Oregon.
- Muller, K. E. (1981). Relationships Between Redundancy Analysis, Canonical Correlation and Multivariate Regression. *Psychometrika* 46(2), 139-42.
- O'Connell, M. T., Robert, C. C. and Christopher, S. S. (2004). Fish Assemblage Stability Over Fifty Years in the Lake Pontchartrain Estuary; Comparisons Among Habitats Using Canonical Correspondence Analysis. *Estuaries* 5(27), 807–17.
- Oksanen, J. (2004). *Multivariate Analysis in Ecology (Lecture notes)*. Department of Biology. University of Oulu.
- Palmer, M. W. (1993). Putting things in even better order: the advantages of canonical correspondence analysis. *Ecol.* 8(74), 2215-30.
- Palmer, M. W. (2017). Access address: http://ordination.okstate.edu/overview.htm#Redundancy_Analysis. Date of access: 18/05/2018.
- Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiol Ecol.* 62(2), 142–60.
- Saporta, G. (1990). *Probabilites. analyses de données et statistiques*. Editions Technip. Paris. 492-93.
- Ter Braak, C. J. F. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *ESA.* 5(67), 1167-79.

Ter Braak, C. J. F. (1987). The Analysis of Vegetation-Environment Relationships by Canonical Correspondence Analysis. *Vegetatio* 64, 69-77.

Ter Braak, C. J. F. and Prentice, I.C. (1988). A theory of gradient analysis. *Adv Ecol Res.* 18, 271–317.

Ter Braak, C. J. F. (1994). Canonical community ordination. Part I: basic theory and linear methods. *Ecoscience* 1, 127–40.

Tso, M. K. S. (1981). Reduced-rank regression and canonical analysis. *J Roy Statist Soc.* 43(B), 183–89.

Witt, S. F. and Witt, C. A. (1992). *Modelling and forecasting demand in tourism*. London: Academic Press.