Ecological Footprint of Consumption in Europe

Jingqi Duan

December 16, 2018

Abstract

The dataset includes 10 variables and 38 observations. Each observation is a country in Europe. 9 numerical variables are investigated in detail using principal component analysis, multiple linear regression, and canonical correlation analysis. 6 variables measure the number of global hectares of certain land type required to support consumption. To understand the meaning of land type variables, two phrases need to be defined. Ecological footprint measures human demand on nature, i.e. the quantity of nature it takes to support people. Global hectare is a measurement unit for the ecological footprint of people or activities. The overall finding is that the country that demands more global hectares of all land types, especially land for carbon emission and food consumption, tends to have higher standard of living. Countries in southeastern Europe, most of which are less developed, require small global hectares of all land types. Most countries in northwestern Europe are high developed and consume large global hectares of all land types, especially forest for carbon emission.

Introduction

The relationship between the environment and development of a country is always debatable. Environmental pressure grows more and more intense due to the overconsumption and overpopulation. The global hectares greatly measure the amount of biological production for human use and human waste assimilation. In other words, a country that has large global hectares of certain land type means that country consume that certain land type more.

The dataset is related to consumption of ecological footprint per capita in Europe in 2014. The dataset has 10 variables with 38 observations. 9 variables are numerical variables and 1 variable categorizes the countries into eastern, southern, western, and northern Europe. Land type variables include *crop*, *grazing*, *forest*, *fishing*, *built-up*, *carbon*, and *total*. All these seven variables measure the ecological footprint consumption per capita.

A. Description of the variables (with *variable names*)

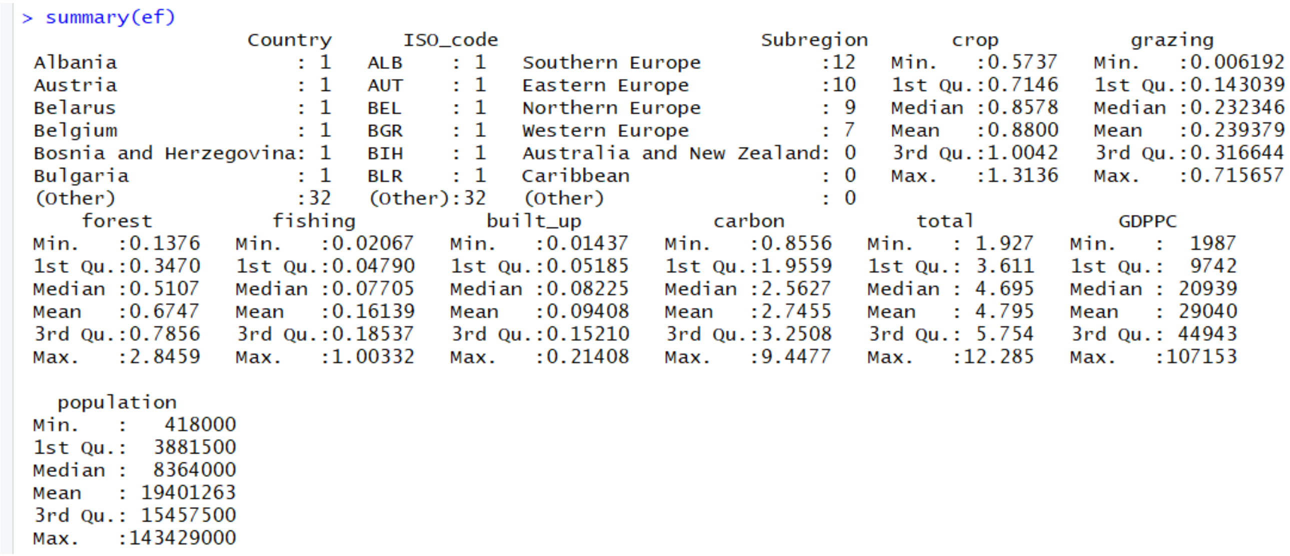
Variables	Descriptions
country	Country names, 38 countries
ISO_code	Iso alpha-3 codes, 38 codes
Subregion	Eastern Europe, Southern Europe, Western Europe, Northern Europe
crop	Global hectares of available or demanded cropland for crops and crops-derived products.
grazing	Global hectares of available or demanded grazing land for meat, diary, leather, etc. It not includes cropland used to produce feed for animals.
forest	Global hectares of available or demanded forest land for sequestration and timber, pulp, or timber products.
fishing	Global hectares of available or demanded marine and inland fishing grounds for fish and fish products.
built_up	Global hectares of available or demanded built-up land for human infrastructure.
carbon	Global hectares of world-average forest required to sequester carbon emissions.
total	The sum of all land types for this country in 2014
GDPPC	Per capita GDP in constant 2010 USD.
population	Population rounded to thousands

Goals

- One important goal is to regress per capita GDP variable on land type variables and population to detect the features of developed or undeveloped countries. In other words, it is to find which land type variables are the best predictor of the per capita GDP.
- Countries in Europe can be divided into four groups by their locations. Many developed countries are in northern or western Europe, such as Switzerland and Norway. It would be meaningful to analyze whether the population or the demanded global hectares of certain land type is highly associated with the economic development of a country.
- Another interesting idea is to investigate the relationships between the production of different land types. For example, whether a country that consumes greatly from cropland also consumes much from fishing grounds.
- Commonly, a country that relies on crop or grazing probably has less carbon emissions that the country that relies on industrial factories. An analysis about the correlation between the demand of land for food consumption and the demand of land for shelter or production will be explored.

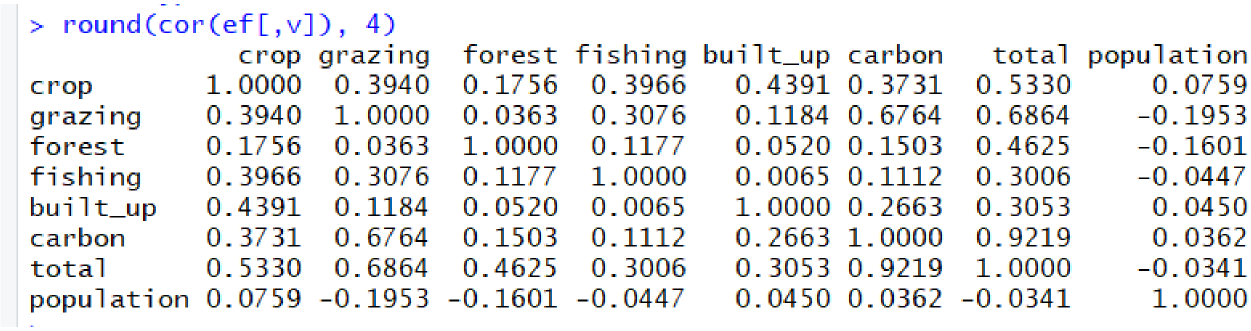
Main results

Prior to undertaking the further analysis on the ecological footprint data, an initial assessment of the data is constructed. It will be helpful to gain some insights into the data by *Table 1*. The data are collected from 38 nations in Europe: 12 in southern Europe, 10 in eastern Europe, 9 in northern Europe, 7 in western Europe. All land type variables have the same unit of measurement, which is the number of global hectares required to support assumption. Overall, land for carbon emissions and forest land for wood production and sequestration are required more than the other land type. Compared among the land type for food support, cropland is needed more than the grazing land fishing grounds. The correlation between numerical variables will be discussed.



```
> summary(ef)
                     Country      ISO_code                       Subregion       crop            grazing        
 Albania                 : 1   ALB    : 1   Southern Europe          :12   Min.   :0.5737   Min.   :0.006192  
 Austria                 : 1   AUT    : 1   Eastern Europe           :10   1st Qu.:0.7146   1st Qu.:0.143039  
 Belarus                 : 1   BEL    : 1   Northern Europe          : 9   Median :0.8578   Median :0.232346  
 Belgium                 : 1   BGR    : 1   Western Europe           : 7   Mean   :0.8800   Mean   :0.239379  
 Bosnia and Herzegovina: 1   BIH    : 1   Australia and New Zealand: 0   3rd Qu.:1.0042   3rd Qu.:0.316644  
 Bulgaria                : 1   BLR    : 1   Caribbean                : 0   Max.   :1.3136   Max.   :0.715657  
 (Other)                 :32   (Other):32   (Other)                  : 0                                      
     forest          fishing           built_up           carbon           total            GDPPC       
 Min.   :0.1376   Min.   :0.02067   Min.   :0.01437   Min.   :0.8556   Min.   : 1.927   Min.   :  1987  
 1st Qu.:0.3470   1st Qu.:0.04790   1st Qu.:0.05185   1st Qu.:1.9559   1st Qu.: 3.611   1st Qu.:  9742  
 Median :0.5107   Median :0.07705   Median :0.08225   Median :2.5627   Median : 4.695   Median : 20939  
 Mean   :0.6747   Mean   :0.16139   Mean   :0.09408   Mean   :2.7455   Mean   : 4.795   Mean   : 29040  
 3rd Qu.:0.7856   3rd Qu.:0.18537   3rd Qu.:0.15210   3rd Qu.:3.2508   3rd Qu.: 5.754   3rd Qu.: 44943  
 Max.   :2.8459   Max.   :1.00332   Max.   :0.21408   Max.   :9.4477   Max.   :12.285   Max.   :107153  
                                                                                                          
   population        
 Min.   :   418000  
 1st Qu.:  3881500  
 Median :  8364000  
 Mean   : 19401263  
 3rd Qu.: 15457500  
 Max.   :143429000  
```

***Table 1.** Descriptive statistics of all variables*



```
> round(cor(ef[,v]), 4)
             crop grazing  forest fishing built_up carbon   total population
crop       1.0000  0.3940  0.1756  0.3966   0.4391 0.3731  0.5330     0.0759
grazing    0.3940  1.0000  0.0363  0.3076   0.1184 0.6764  0.6864    -0.1953
forest     0.1756  0.0363  1.0000  0.1177   0.0520 0.1503  0.4625    -0.1601
fishing    0.3966  0.3076  0.1177  1.0000   0.0065 0.1112  0.3006    -0.0447
built_up   0.4391  0.1184  0.0520  0.0065   1.0000 0.2663  0.3053     0.0450
carbon     0.3731  0.6764  0.1503  0.1112   0.2663 1.0000  0.9219     0.0362
total      0.5330  0.6864  0.4625  0.3006   0.3053 0.9219  1.0000    -0.0341
population 0.0759 -0.1953 -0.1601 -0.0447   0.0450 0.0362 -0.0341     1.0000
```

***Table 2.** Correlation matrix of all numerical variables except GDPPC*

Examination of the correlation matrix shown in *Table 2* shows that all pairs of variables regarding to global hectares of land types are positively correlated, some moderately and others slightly. *Total* is positively correlated with *crop, grazing, forest, fishing, built-up,* and *carbon* because it measures the sum of all land types. Interestingly, the correlation between *total* and *carbon* is 0.922, which is extremely high while *total* is moderately correlated with the rest five land type variables – range from 0.30 to 0.69. The high correlation seems to make sense because the value of *carbon* is much higher than that of other land type variable and *total* accounts for the sum of all land type variables. *Forest* is slightly correlated with the other five land type variables – range from 0.03 to 0.18. One possible explanation is forest land is required for sequestration and timber products while the other five land types are demanded for basic consumption, such as food and shelter. The correlations between *population* and all seven land type variables are slightly weak; some correlations are positive, and some correlations are negative.

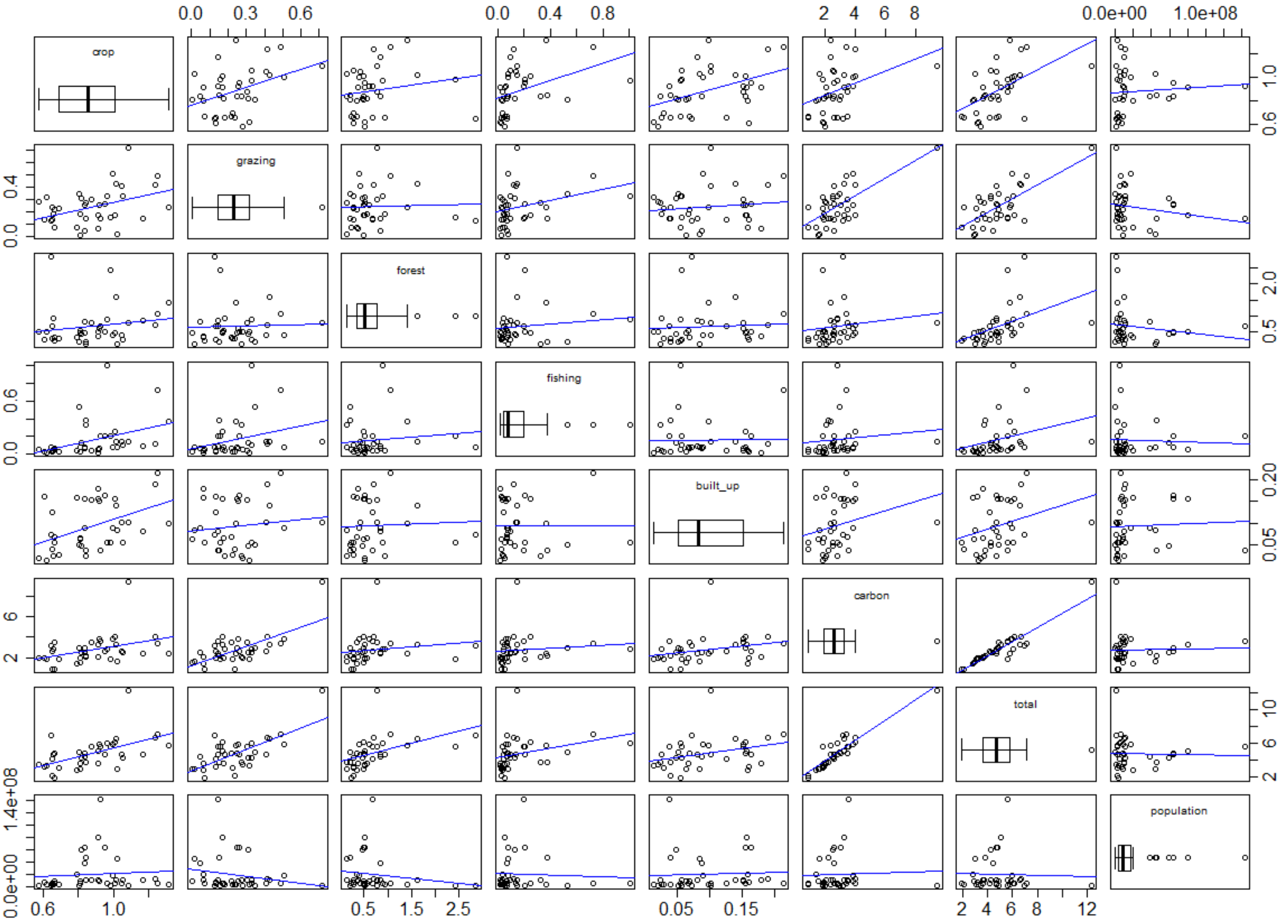


Fig. 1. *Scatterplot matrix showing the linear fit of each pair of variables*

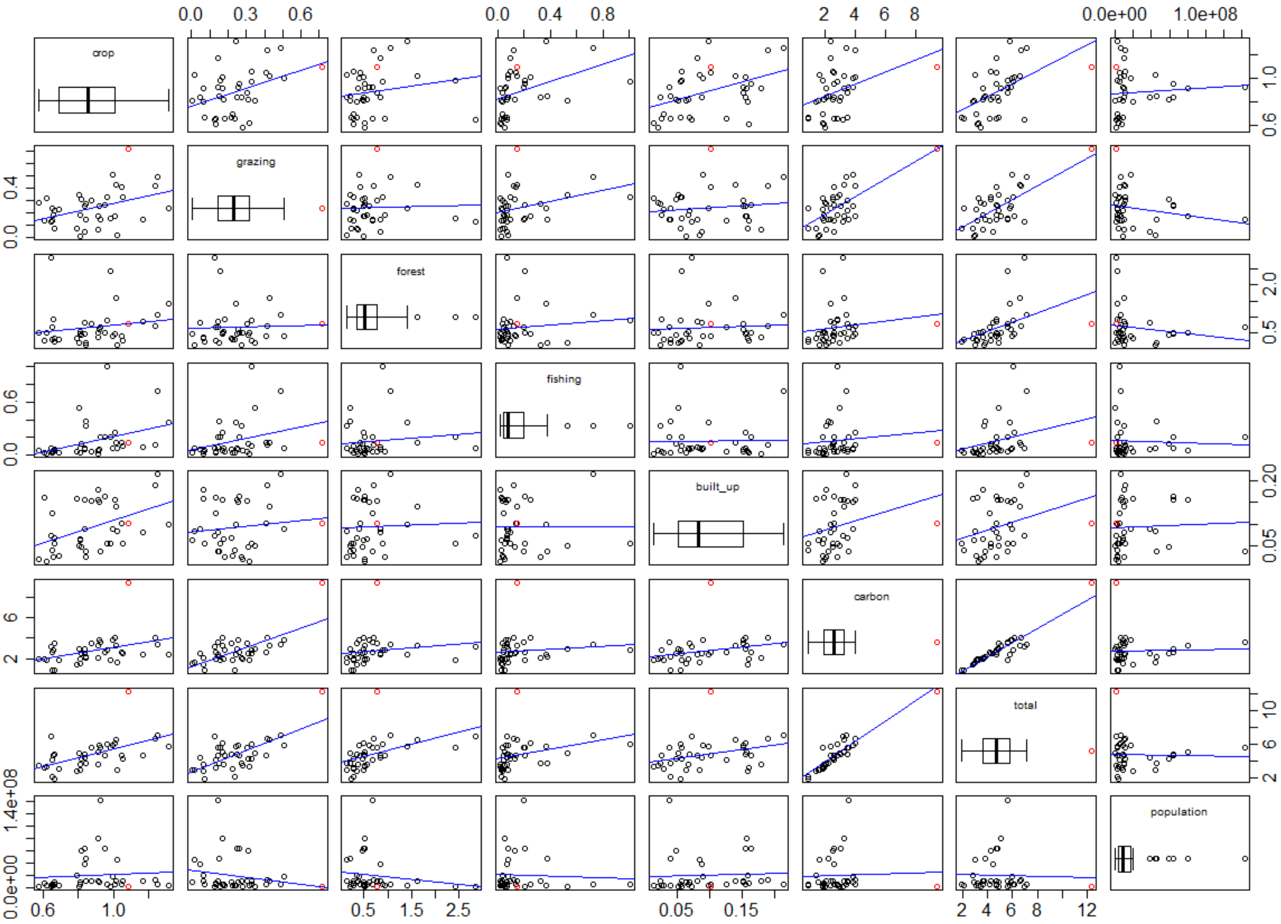


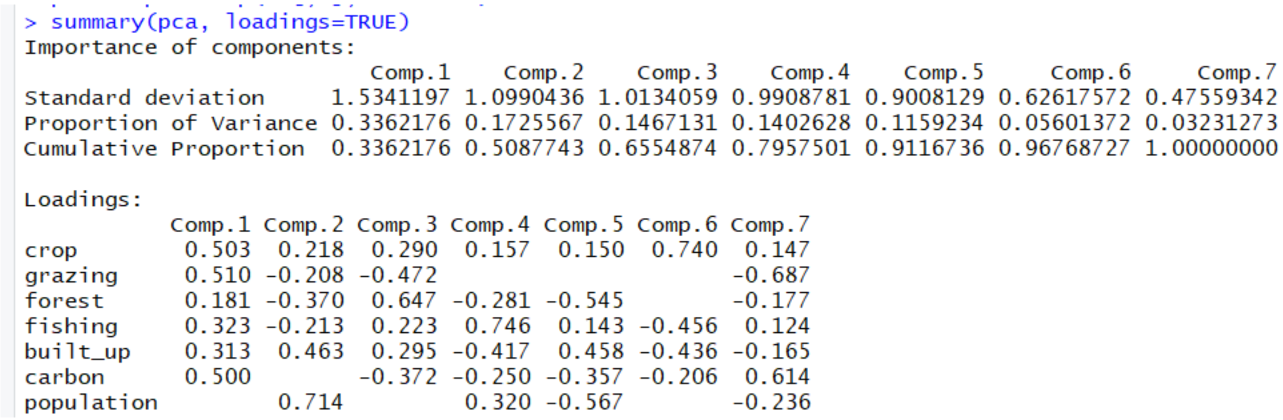
Fig. 2. *Scatterplot matrix coloring Luxembourg red with the linear fit of each variables*

A scatterplot matrix of the eight variables with boxplot for each variable on the main diagonal is shown in *Figure 1*. One very clear observation in this plot is that for all variables there is one country which needs quite more hectares of *carbon* land and *total* land than the other countries in Europe. That country is Luxembourg. Investigating some background about environment of Luxembourg, there is one possible explanation. Mentioned previously, *total* is highly correlated with *carbon*. The 2007 carbon dioxide emissions for Luxembourg are the highest in Europe and substantially higher than that of Europe; therefore, the global hectares of total land and land required for carbon emissions by Luxemburg are much larger than that of the other countries. A second scatterplot matrix of these variables with Luxembourg colored red is shown in *Figure 2*. Luxembourg is also an outlier in *grazing* because Luxemburg consumes the most meat per person of all the countries in the world. However, in the scatterplots involving the rest variables, it does not stand out. Since there are only 38 countries and Luxembourg is not an extreme outlier in all aspects, the further analysis will use the original dataset without omitting any observations.

I. Principal Components Analysis

Since most pairs of variables are moderately correlated and there are only 38 observations with 7 variables, the principal component analysis is applied to the ecological footprint data. The principal components of the data are extracted from the correlation matrix instead of the covariance, because the variables are of completely different types – global hectares of land type and population. To investigate the relationship *crop, grazing*, *forest, fishing*, *built-up, carbon,* and *population* and *GDPPC*, the multiple linear regression will be applied to principal components.

A. Principal components analysis



```
> summary(pca, loadings=TRUE)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5     Comp.6     Comp.7
Standard deviation     1.5341197 1.0990436 1.0134059 0.9908781 0.9008129 0.62617572 0.47559342
Proportion of Variance 0.3362176 0.1725567 0.1467131 0.1402628 0.1159234 0.05601372 0.03231273
Cumulative Proportion  0.3362176 0.5087743 0.6554874 0.7957501 0.9116736 0.96768727 1.00000000

Loadings:
           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
crop        0.503  0.218  0.290  0.157  0.150  0.740  0.147
grazing     0.510 -0.208 -0.472                      -0.687
forest      0.181 -0.370  0.647 -0.281 -0.545        -0.177
fishing     0.323 -0.213  0.223  0.746  0.143 -0.456  0.124
built_up    0.313  0.463  0.295 -0.417  0.458 -0.436 -0.165
carbon      0.500        -0.372 -0.250 -0.357 -0.206  0.614
population         0.714         0.320 -0.567        -0.236
```

Table 3. *Summary of the principal components*

The summary of principal components is shown in *Table 3*. The first four components have variances 1.534, 1.100, 1.013, and 0.991 and together account for almost 80% of the variance of the original variables. Also, a scree plot of variances of these seven principal components is constructed in *Figure 4* to better visualize. Scores on these four component are used later to visualize the data in matrix of two-dimensional plots.

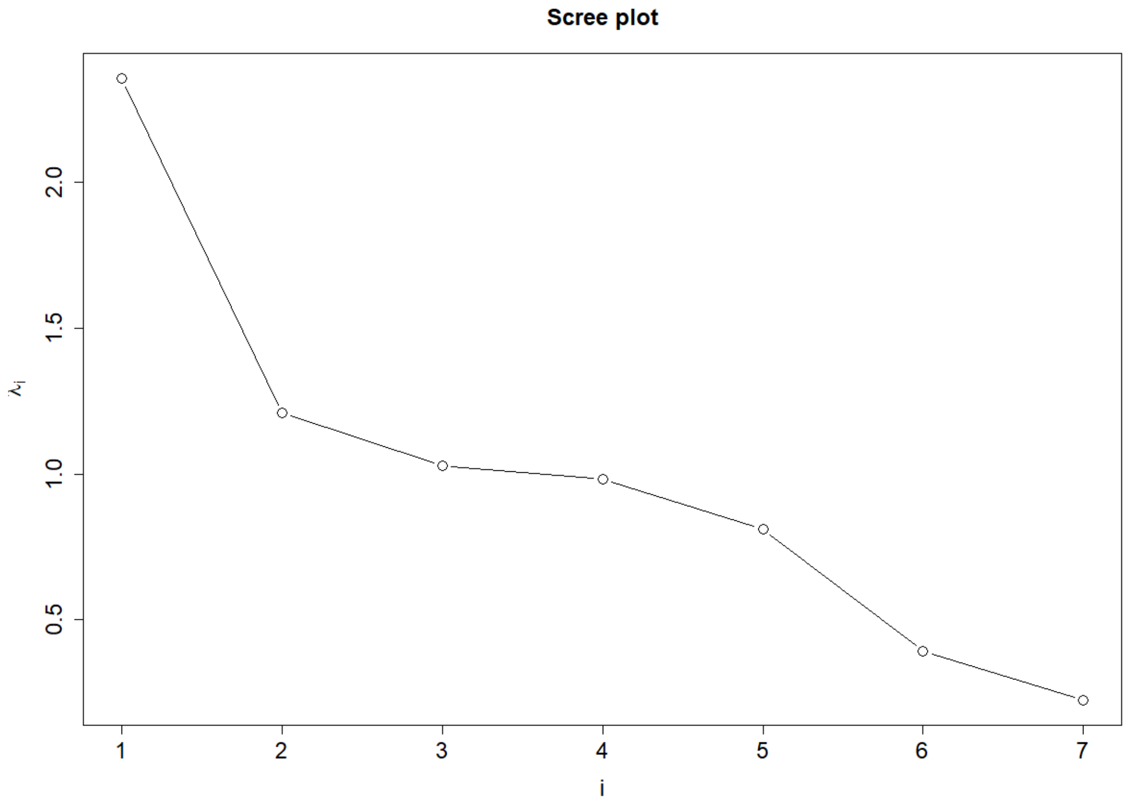


Fig. 3. *Scree plot of variances of seven principal components*

A plot of the first two principal component loadings is shown in *Figure* 4. The plot demonstrates the ability of variables to affect the component scores. It is important to understand what the first four principal components represent. The first four principal components are formulated below:

PC1 = 0.503 crop + 0.510 grazing + 0.181 forest + 0.323 fishing + 0.313 built-up + 0.500 carbon

PC2 = 0.218 Crop – 0.208 grazing – 0.370 forest – 0.213 fishing + 0.463 built-up + 0.714 population

PC3 = 0.290 crop – 0.472 grazing + 0.647 forest + 0.223 fishing + 0.295 built-up – 0.372 carbon

PC4 = 0.157 crop – 0.281 forest + 0.746 fishing – 0.417 built-up – 0.250 carbon + 0.320 population

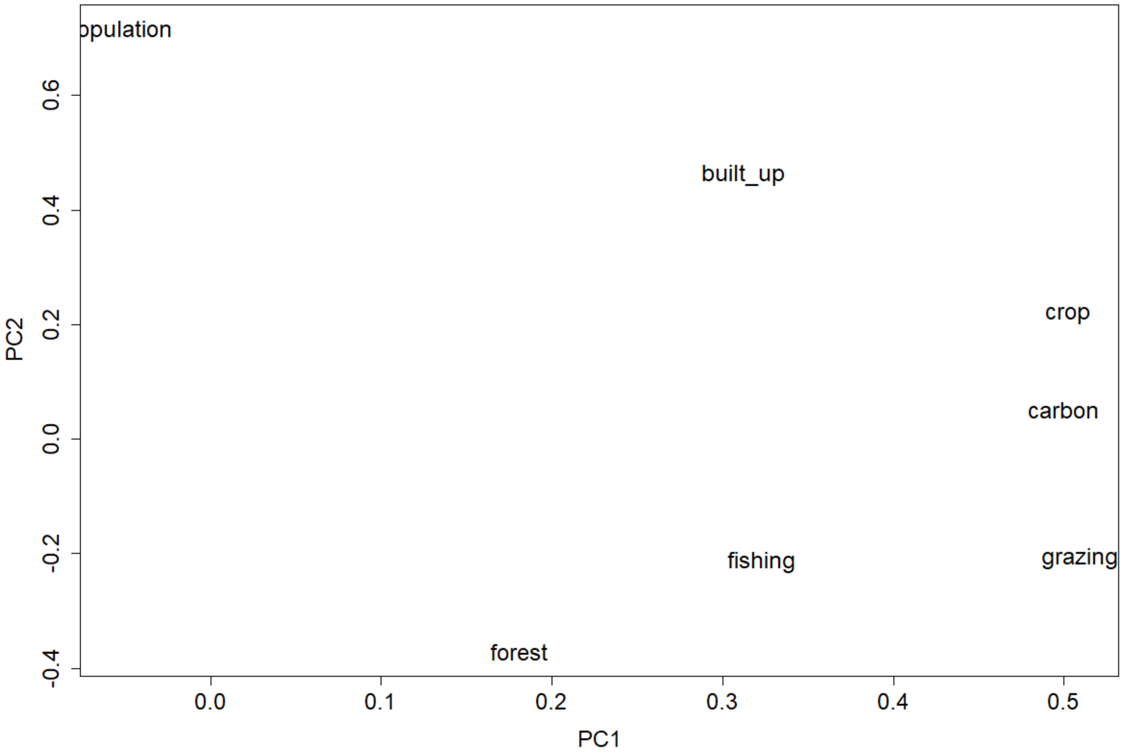


Fig. 4. *Plot of the first two principal component loadings*

Clearly, the first component (PC1) might be viewed as a weighted average of all six land type variables. The plot shows the same interpretation that *crop*, *grazing* and *carbon* contribute more to the value of PC1 and the coefficient of *population* is negative but approximately zero. High values of PC1 probably indicate relatively large hectares of land for food consumption and carbon equestrian; in other words, the country with higher PC1 score probably needs larger hectares of land type to support food consumption and carbon sequestration. The second component (PC2) essentially contrasts *population* and *built-up* with *forest* measuring the population and the land demanded for human activities. The country with higher PC2 score probably has larger population and require the global hectares of land for human infrastructure more. High values of PC2 could also indicate relatively low global hectares of *grazing*, *fishing*, and *forest* because human take most land for infrastructure and fewer biological production from grazing land, marine and forest land for human use. The third component (PC3) contrast *grazing* and *carbon* with the other land type variables, especially *forest*. The fourth component (PC4) is essentially a contrast between *fishing* and *built-up*.

A bivariate boxplot matrix plot based on the first four principal component scores is shown in *Figure 5.* The plot demonstrates that Luxembourg is an outlier and suggests that Norway and Denmark may also be suspects in this respect. Analyzed previously, the value of required global hectares of land for carbon emission by Luxembourg is very high and Luxembourgers consumed the most meat in the world. Therefore, Luxembourg has the highest PC1 score and lowest PC3 score. Norway has the highest PC4 score and relatively low PC3 score. It might be explained by the extensive coastline, facing the North Atlantic Ocean. Norwegian cuisine also focuses more on fish, which explains the high required global hectares of fishing grounds. Denmark has very high PC1 score. The fact that carbon emission of Denmark has increased sharply because many tech companies construct sites for major facilities in Denmark. It is reasonable that the demand of land for carbon emission is high for Denmark. Observing the bivariate boxplots, most countries do not have large PC1 and PC2 scores. It indicates that most European countries have small population demand low global hectares of land for consumption, wood production, infrastructure. and carbon emission.

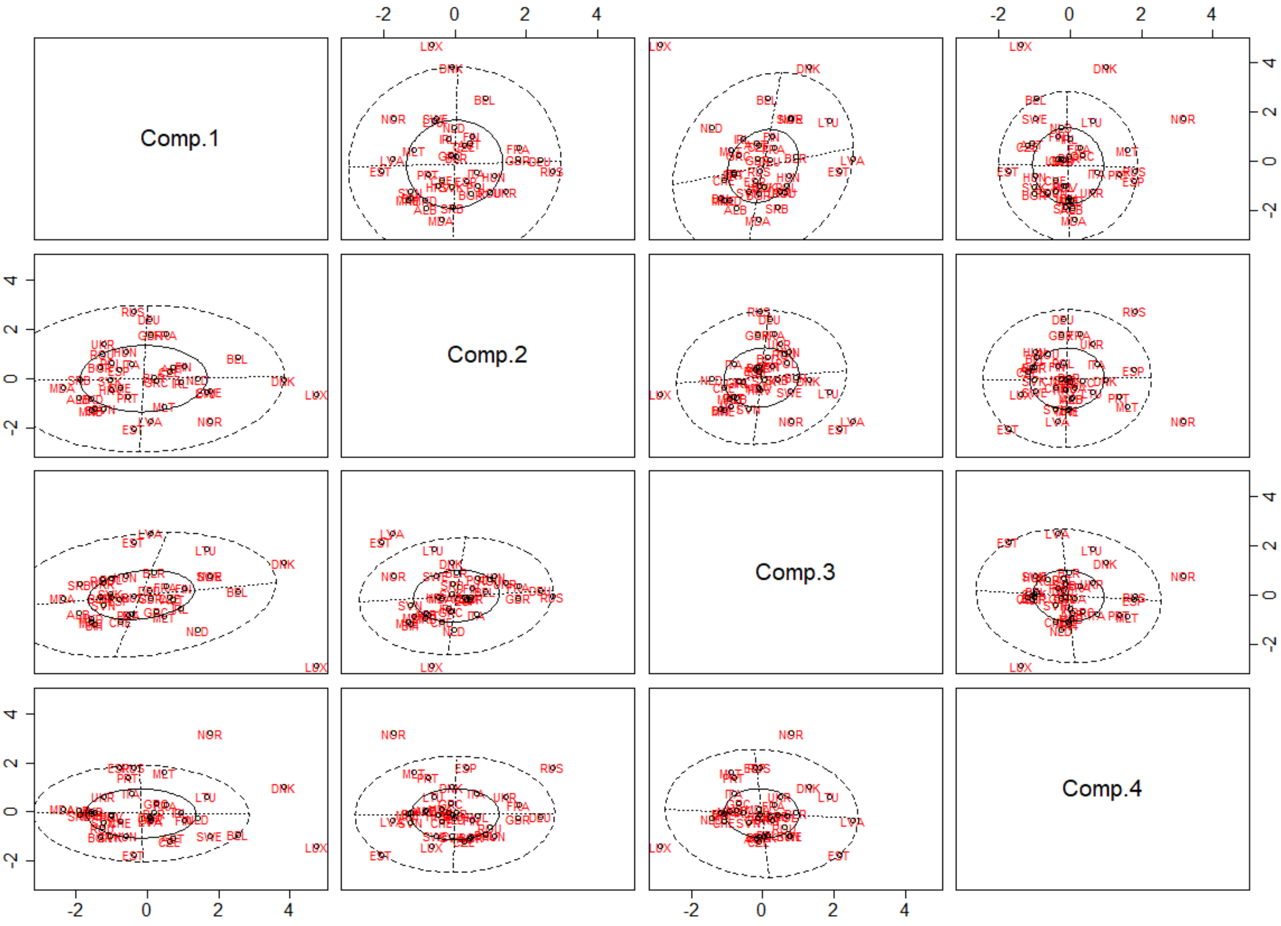


Fig. 5. *Bivariate boxplot matrix of the first four principal component scores*

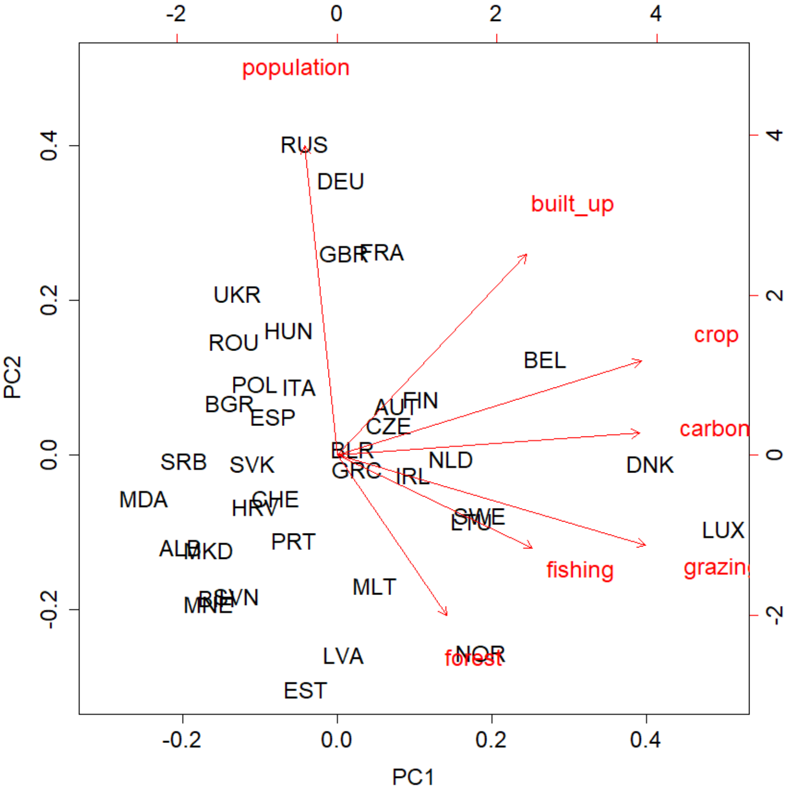


Fig. 6. *Biplot of the first two principal components*

The biplot (*Figure 6*) is plotted using the first two principal components. Although the first two components only count for approximately 50% of the variance of data, the biplot still demonstrates some useful information. Russia and Germany have the largest population in Europe, but their global hectares of land for consumption and carbon emission are not high. Instead, these two nations need high global hectares for human infrastructure. Most nations in Europe has relatively small population and low global hectares of each land type. Moreover, a plot of the first two principal component scores with each point colored by their *subregion* is shown in *Figure 7*. Countries in eastern and southern Europe tend to have low demand for all land types; however, countries in eastern Europe have large population while countries in southern Europe have small population. Countries in western and northern Europe vary from country to country, but most of them consume large global hectares of all land types.

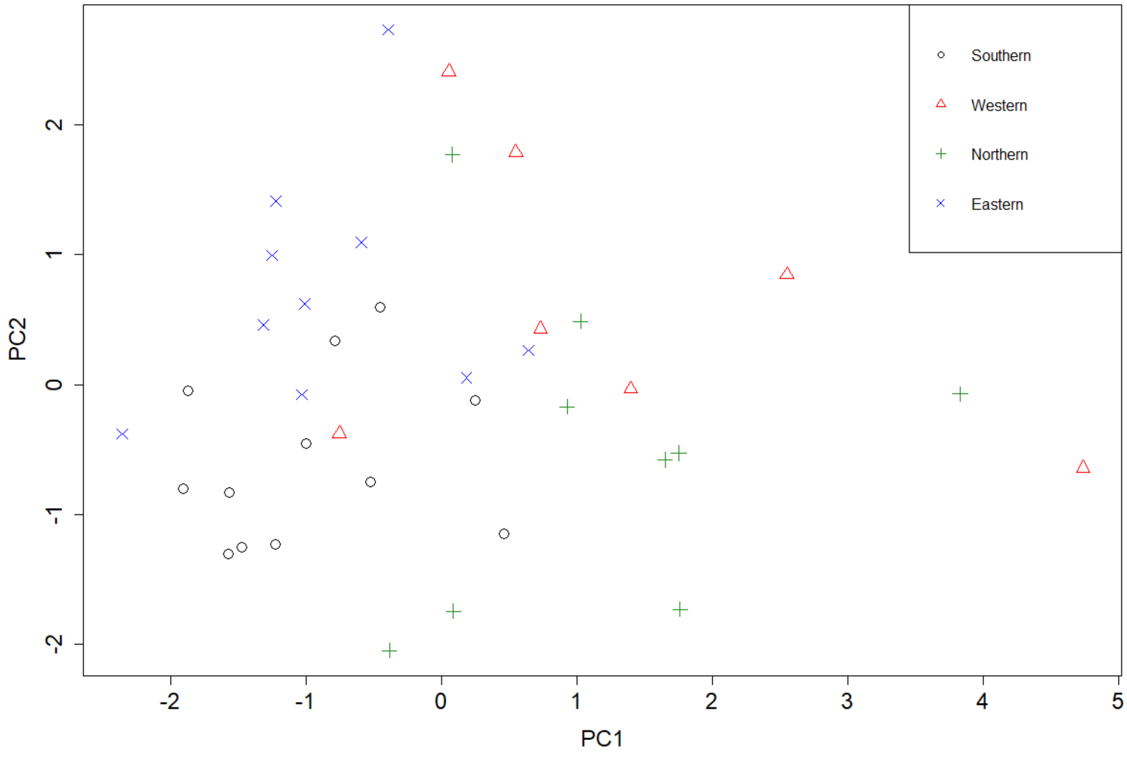


Fig. 7. *Plot of the first two principal component scores*

In summary, the demand of certain land type can be related to the country location, the economic development, and the favor of the residents. For example, Norway is developed, but it consumes more biological production from marine or fishing grounds instead of grazing land because Norway has an extensive coastline and Norwegian cuisine involves more with fish. Briefly, the first four principal components, which account for almost 80% variance of the data, are used to observe the relationship between land type variables and population. Most European countries do not demand much from environmental production and has small population. However, there are some outliers, such as Luxembourg, Denmark, and Norway. The countries in western and northern Europe are likely to demand large global hectares of land types, especially land for carbon emissions and food production. The countries in southeaster Europe require low global hectares of land type.

II. Multiple linear regression

In the data, there are six variables measure the ecological assets that a given population requires to product the natural resources it consume and to absorb it waste, especially carbon emissions. Moreover, the development of a country is usually associated with demand for land. A less developed country needs land for its own food consumption and production. A developing country needs land for infrastructure. Highly developed countries, like Luxembourg and Denmark, encounter the large carbon emissions caused by technique development and hence require large global hectare of forest land for sequestration. A measure of the economic development of a country is per capita GDP. The multiple linear regression is used to determine which of the land type and population variables are the best predictors of the standard of living in a country or the development of a country.

Because the land type variables are moderately correlated, principal components are used as explanatory variables in the regression. Before regressing the *GDPPC* variables on all seven principal components, the plots of regressing the *GDPPC* variables on each land type and population variables are shown in *Figure 8.* The plot demonstrates that all six land type variables are positively correlated with *GDPPC* with *carbon* the most correlated and *forest* the least correlated.

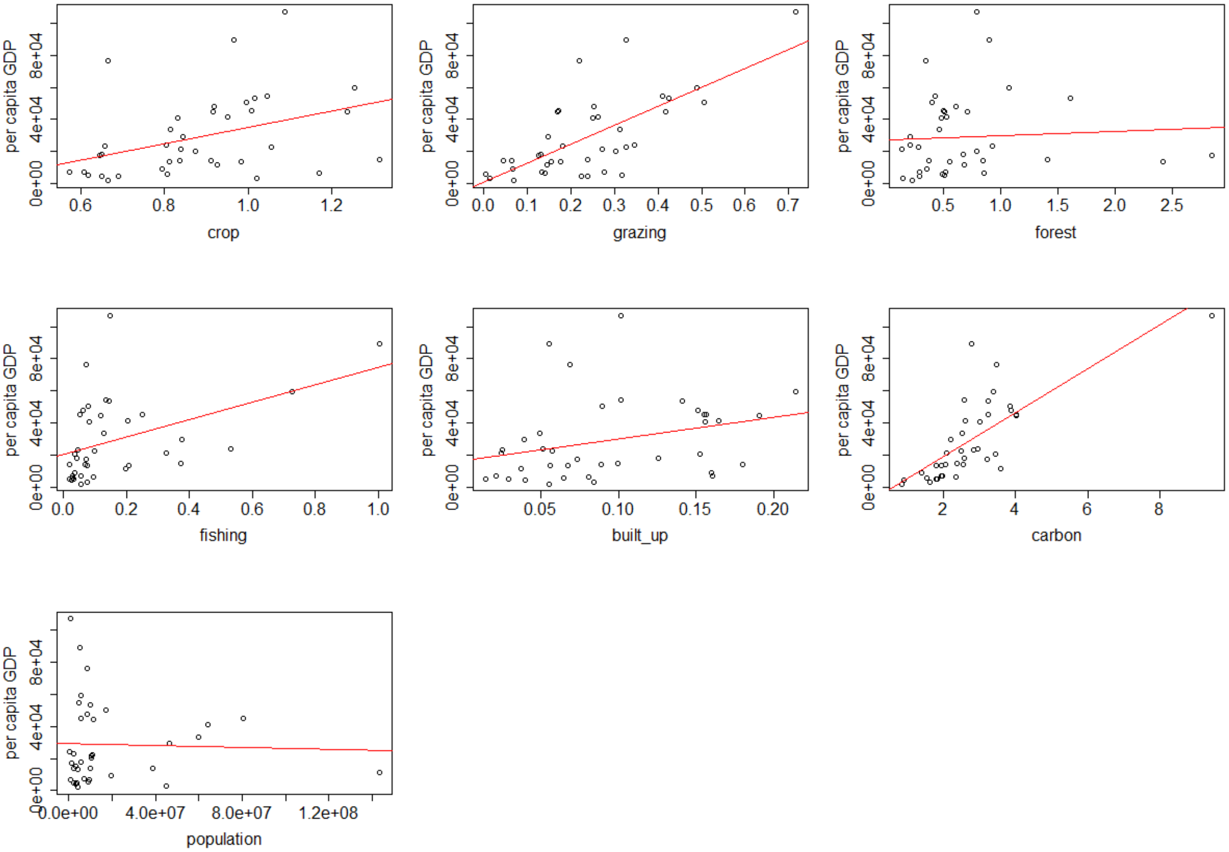
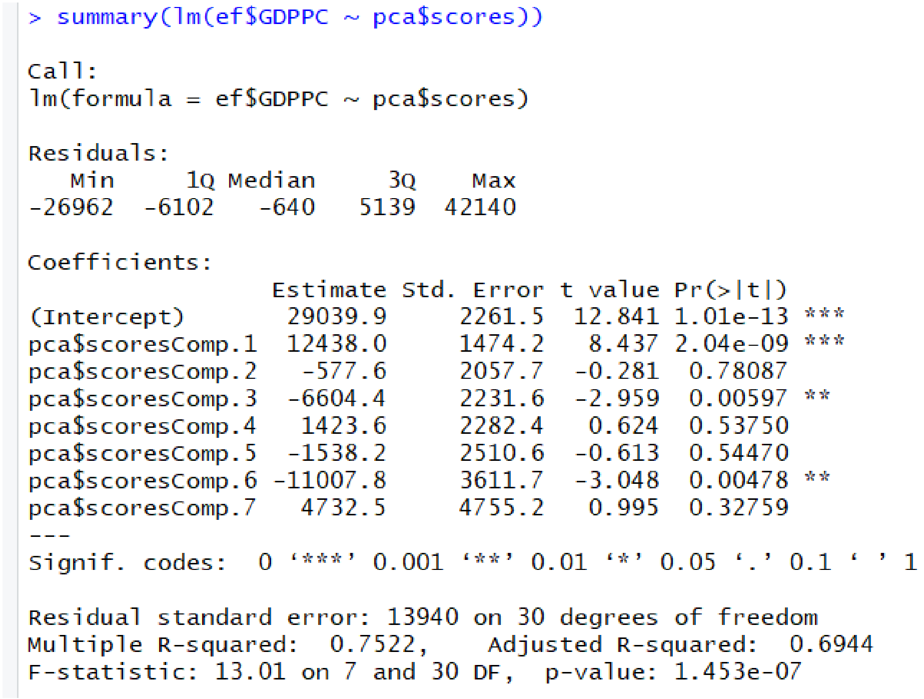


Fig. 8. *Plot of regressing GDPPC on each variable (above)*

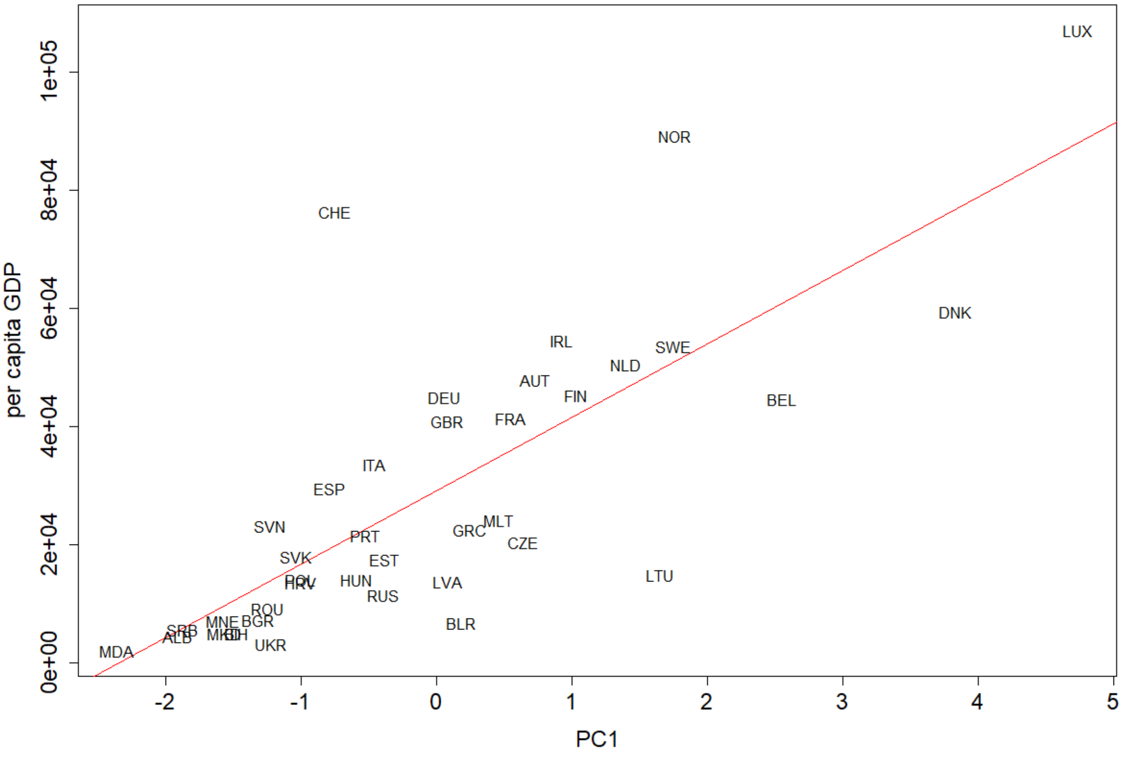
Table 5. *Summary of multiple linear regression of GDPPC on all PCs (below)*



```
> summary(lm(ef$GDPPC ~ pca$scores))

Call:
lm(formula = ef$GDPPC ~ pca$scores)

Residuals:
   Min     1Q Median     3Q    Max 
-26962  -6102   -640   5139  42140 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       29039.9     2261.5  12.841 1.01e-13 ***
pca$scoresComp.1  12438.0     1474.2   8.437 2.04e-09 ***
pca$scoresComp.2   -577.6     2057.7  -0.281  0.78087    
pca$scoresComp.3  -6604.4     2231.6  -2.959  0.00597 ** 
pca$scoresComp.4   1423.6     2282.4   0.624  0.53750    
pca$scoresComp.5  -1538.2     2510.6  -0.613  0.54470    
pca$scoresComp.6 -11007.8     3611.7  -3.048  0.00478 ** 
pca$scoresComp.7   4732.5     4755.2   0.995  0.32759    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13940 on 30 degrees of freedom
Multiple R-squared:  0.7522,    Adjusted R-squared:  0.6944 
F-statistic: 13.01 on 7 and 30 DF,  p-value: 1.453e-07
```

The summary of multiple linear regression of *GDPPC* on all seven PCs is shown in *Table 5*. The graph demonstrates that the first principal component score is the most predictive of per capita GDP. Although PC3 and PC6 are also significant, their variances are relatively small and probably has small correlations with the response. The plot of regressing *GDDPC* on PC1 is shown in *Figure 9*. The first principal component is mainly a weighted average of all six land type variables. *Crop*, *grazing*, and *carbon* contribute the most to the values of PC1 score. It also can be noticed that highly developed countries Luxembourg, Denmark, and Norway have high per capita GDP while they also demand large global hectors of grazing land and land for carbon emissions. Thus, after regressing of *GDPPC* on all seven PCs, it can be observed that per capita GDP is positively correlated with the weighted average of all six land type variables. Specifically, the country with large demand of cropland, grazing land, and land for carbon emission tends to have higher standard of living.

Fig. 9. *Plot of regressing GDDPC on PC1*



III. Canonical Correlation Analysis

The canonical correlation analysis is applied to the data to find to what extent, the global hectares of demanded for human consumption are related to the global hectares of demanded for air and shelter.

A. Description of scaled variables

The global hectares of demanded *grazing* land and *fishing* grounds are combined because they both measure the world's annual amount of biological production of meat and seafood for human consumption. The global hectares required *forest* for timber products and sequestration and for carbon emission are combined because they both measure wood production and carbon absorption.

Variables	Variables in Dataset	Descriptions
X_1	Global hectares of demanded *crop* land	Amount of biological production of food crops for human consumption
X_2	Global hectares of demanded *grazing* land and marine or inland *fishing* grounds	Amount of biological production meat and seafood for human consumption
Y_1	Global hectares of demanded *forest* land for timber products, *carbon* emission, and sequestration	Amount of biological production for wood production and carbon absorption
Y_2	Global hectares of demanded *built-up* land for human infrastructure	Amount of biological production for human infrastructure

B. Interpretation of canonical correlations and variables

The first canonical correlation is 0.528 and tested by Bartlett's test at 0.01 level, there is strong evidence that the first canonical correlation is significant. The corresponding canonical variables are formulated below:

$U_1 = 0.993\ X_1 – 0.117\ X_2$

$V_1 = – 0.698\ Y_1 – 0.716\ Y_2$

U_1 is mainly a contrast between the biological production for human crops consumption and the production for human meat consumption, but the weight on X_1 is much larger than on X_2. V_1 is mainly a function of demanded forest land and demanded built-up land. The moderately positive correlation shows that the country that has a high global hectares of cropland tends to have low global hectares of forest and built-up land.

The second canonical correlation is 0.349 and tested by Bartlett's test at 0.05 level, there is strong evidence that the second canonical correlation is significant. The corresponding canonical variables are formulated below:

$U_2 = – 0.496\ X_1 + 0.868\ X_2$

$V_2 = – 0.712\ Y_1 + 0.702\ Y_2$

U_2 is mainly a contrast between the biological production for human crops consumption and the production for human meat consumption, but the weight on X_2 is twice as much as X_1. V_1 is mainly a contrast between demanded forest land and demanded built-up land. The moderately positive correlation shows that the country that has a high global hectares of grazing land and fishing grounds tends to have large global hectares of built-up land.

C. Plot of the first canonical variables

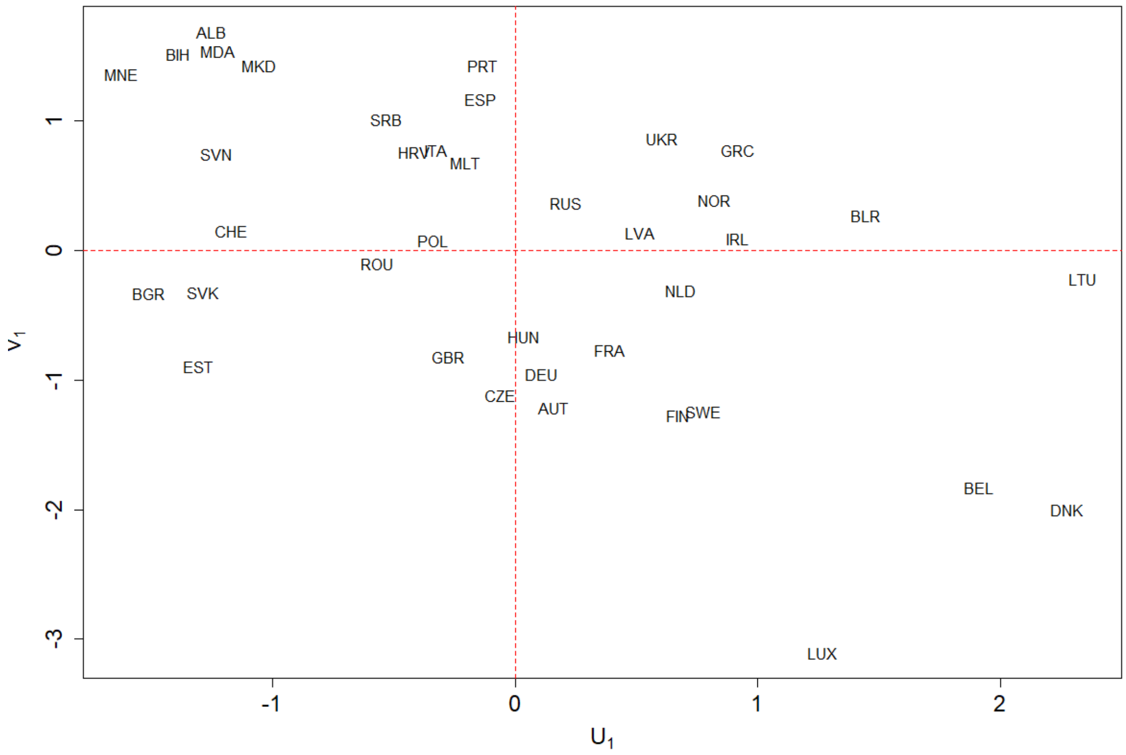


***Fig. 10.** Plot of the first canonical variables*

The vertical and horizontal lines in *Figure 10* divide the plot into four parts. 14 countries have negative U_1 value and positive V_1 value, which means these countries demand much large global hectares of land type for meat and seafood production. Relatively small value of V_1 indicates that these countries require low global hectares of land for wood production and carbon emissions. 11 countries have positive U_1 value and negative V_1 value, which means theses countries demand large global hectares of land type for food consumption and moderate amount for wood production, carbon emissions, and infrastructure.

Conclusions

Many multivariate analysis methods and figures are taught in the class. The hardest part is to determine what information or solution should be found and how to investigate the data with appropriate model or methods. With many variables, the dimensionality of the figure is high. A two-dimensional plot might not have enough dimensions to express all the relationships and distances. Regressing *GDPPC* on all seven principle components illustrates that PC1 is the best predictor of a country's per capita GDP. PC1 is a weighted average of all six land type variables with much weight on cropland, grazing land, and land for carbon dimensions. A country with high PC1 score probably requires large hectares of land for carbon dimension and food production and has larger per capita GDP. Most countries in southeastern Europe consume less from all land types and are less developed. It was expected to discover the relationship of the demand of land for food production and the demand of land for shelter and sequestration. The canonical correlation variables are hard to interpret. All the interpretations are made separately, and it is hard to combine them together. Overall, this project helps understanding the material. Instead of giving a dataset and following the instructions of several computations, it is more the process of observing the data and apply the appropriate method.