University of Business and Technology in Kosovo
UBT Knowledge Center

UBT International Conference

2014 UBT International Conference

Nov 8th, 9:00 AM - 9:15 AM

The analysis of compositional data, a general overview and an application with GDP data for Albanian economy in R software

Mirjeta (DËRA) Pasha
Aleksandër Moisiu University, mirjetadera@yahoo.com

Edlira Kalemi
Aleksandër Moisiu University, edlirakalemi@gmail.com

Senada Bushati
Aleksandër Moisiu University, bushatin@yahoo.com

Anisa Skandaj
Aleksandër Moisiu University, anisa.skandaj@gmail.com

The analysis of composıtıonal data, a general overvıew and an applicatıon with GDP data for Albanıan economy in R software

Mirjeta Pasha (DËRA)[1], Edlira Kalemi[2], Senada Bushati[3], Anisa Skandaj[4]
[123]University Aleksandër Moisiu Durrës, Albania
mirjetadera@yahoo.com[1], edlirakalemi@gmail.com[2], bushatin@yahoo.com[3]
anisa.skandaj@gmail.com[4]

Abstract. In the paper "The analysis of compositional data, a general overview and an application with GDP data for Albanian economy in R software" we have studied the principle rules of the compositional data analysis. We have listed some of the fields where we can find and we can apply compositional data analysis. Furthermore there have been treated the main problems that a user will have during the work with coda data. After problems there are a lot of ways and methods in order to avoid those problems and some transformations that really help the coda work. The most important part of this work will be considered the application that we have separated it into two parts. We have chosen the GDP data, because we can consider them as compositional data. From every model we have concluded some important results and we have compared some parameters and results too. As a conclusion we have introduced the idea for a further work.

Keywords: compositional data analysis, model, predictions, GDP

1 Description of the methodology used

Compositional data have been applicated in many fields, including economy. In the context of developing an application for exposing concrete compositional data analysis, we selected a key economic indicator called Gross Domestic Product GDP. Recently, performers and professionals in economy fields, consider not only economic indicators relating to domestic production, but also they consider as an indicator the Bruto Added Value. The implementation of this application, has been requested in terms of time and intellectual capacity. Taught by professionals in the field, there are several methods that can be applied to calculate approximately the GDP, and are different considerations for this indicator in combination with other economic indicators.

An economic model at the macro level (MAEMA) was used to make economic situation or scenario simulations that could really occur in certain situations, not only economic, but also political, financial, climate, etc. After preparing a strong theoretical basis and after successive searches in various literature, we have concluded that this application is accomplished as follows:

1. The data on GDP taken by INSTAT and periodic performances from the Bank of Albania, to create a simple model, an MA, AR, ARMA, ARIMA or VARIMA, considering these data as simple time series, without compositions or subcompositions.
2. Considering the data obtained by INSTAT as compositional data, we will realize the calculations and predictions for a later time period.
3. Carrying out a comparative work
    a. Among the results given by different methods
    b. Among the results given by applied methods and between actual observed data

Compositional database description: GDP is divided into several sectors that are:

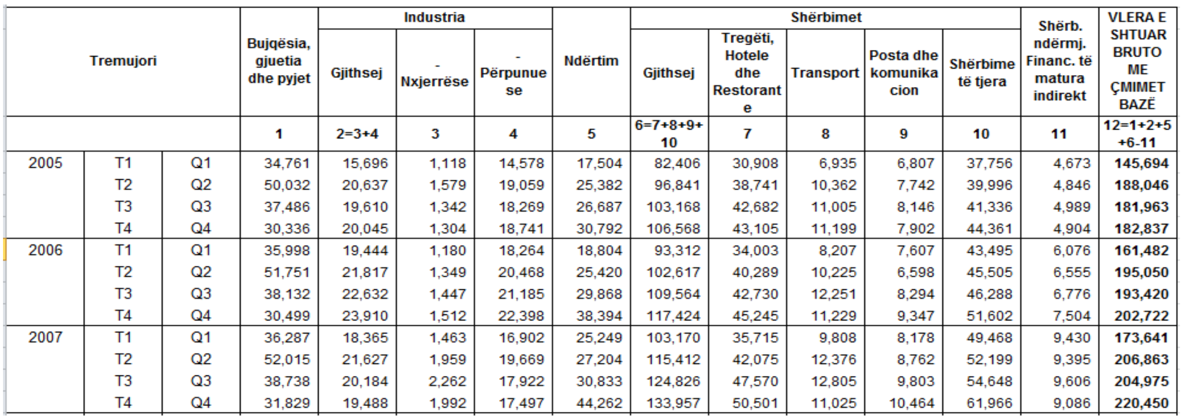
1. Agriculture, hunting and forestry
2. Industry

3. Construction
4. Services
5. Financial Services directly measured

Some sectors are subdivided into several other sub-sectors, for example services are divided into:

- ✓ Total = 2 + 3 + 4 + 5
- ✓ Market, Hotels and Restaurants
- ✓ Transport
- ✓ Post and telecommunications
- ✓ Other Services

Table 1. A presentation of the data of GDP, divided by sector for each quarter



Tremujori			Bujqësia, gjuetia dhe pyjet	Industria			Ndërtim	Shërbimet					Shërb. ndërmj. Financ. të matura indirekt	VLERA E SHTUAR BRUTO ME ÇMIMET BAZË
				Gjithsej	- Nxjerrëse	- Përpunuese		Gjithsej	Tregëti, Hotele dhe Restorante	Transport	Posta dhe komunikacion	Shërbime të tjera		
			1	2=3+4	3	4	5	6=7+8+9+10	7	8	9	10	11	12=1+2+5+6-11
2005	T1	Q1	34,761	15,696	1,118	14,578	17,504	82,406	30,908	6,935	6,807	37,756	4,673	**145,694**
	T2	Q2	50,032	20,637	1,579	19,059	25,382	96,841	38,741	10,362	7,742	39,996	4,846	**188,046**
	T3	Q3	37,486	19,610	1,342	18,269	26,687	103,168	42,682	11,005	8,146	41,336	4,989	**181,963**
	T4	Q4	30,336	20,045	1,304	18,741	30,792	106,568	43,105	11,199	7,902	44,361	4,904	**182,837**
2006	T1	Q1	35,998	19,444	1,180	18,264	18,804	93,312	34,003	8,207	7,607	43,495	6,076	**161,482**
	T2	Q2	51,751	21,817	1,349	20,468	25,420	102,617	40,289	10,225	6,598	45,505	6,555	**195,050**
	T3	Q3	38,132	22,632	1,447	21,185	29,868	109,564	42,730	12,251	8,294	46,288	6,776	**193,420**
	T4	Q4	30,499	23,910	1,512	22,398	38,394	117,424	45,245	11,229	9,347	51,602	7,504	**202,722**
2007	T1	Q1	36,287	18,365	1,463	16,902	25,249	103,170	35,715	9,808	8,178	49,468	9,430	**173,641**
	T2	Q2	52,015	21,627	1,959	19,669	27,204	115,412	42,075	12,376	8,762	52,199	9,395	**206,863**
	T3	Q3	38,738	20,184	2,262	17,922	30,833	124,826	47,570	12,805	9,803	54,648	9,606	**204,975**
	T4	Q4	31,829	19,488	1,992	17,497	44,262	133,957	50,501	11,025	10,464	61,966	9,086	**220,450**

GDP database, is a compositional database, or more specifically a 5-compositional database and two of it's compositions like services and industry are compositions by itself. Services have 4 sub-compositions and industry have 2 sub-compositions.

The problem that arises is: If we have a historical performance of the parts of a composition, what can we say about their evolution in a further period? For example, if we consider the data for annual GDP, with only five main divisions, without neglecting sub-compositions will have a situation as follows: A historical evolution of GDP may be presented through pie charts

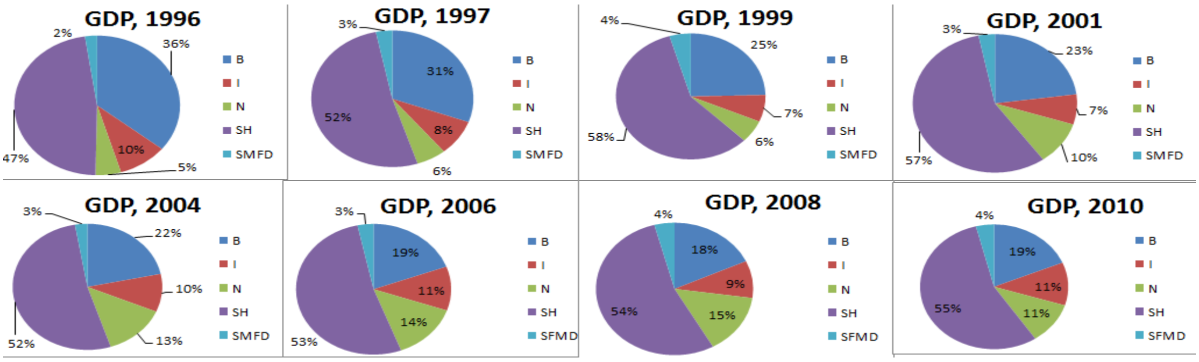


Figure 1. A historical percentage of GDP by sectors over the years.

If the situation is as follows, what can we say about the year 2011, 2012, 2013. What about a longer term o period? All these questions can be answered by a detailed analysis of the data in two perspectives: as simple data, as a time series and compositional data.

2. Simple statistical analysis considering data as compositional

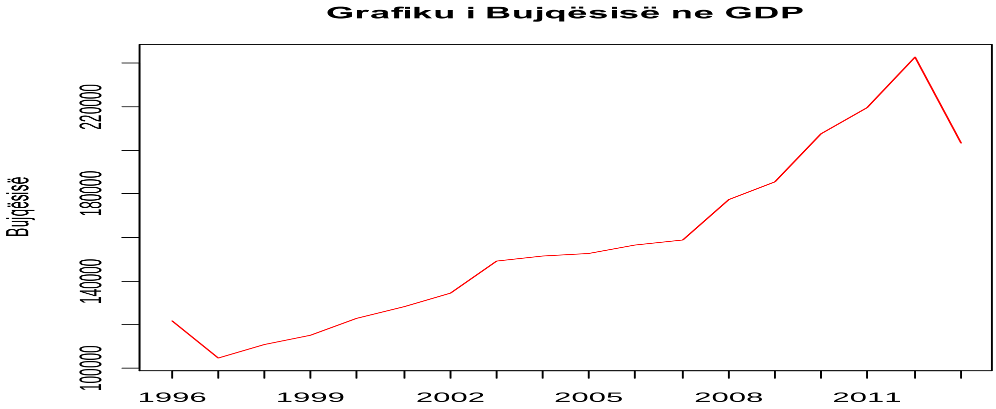


Figure. 2. Graphic performance of the agricultural sector, with contribution to GDP

The figures above show that GDP in the agriculture sector in 2007 has undergone in a deep recession. In the following years we have development, in 2012 culminated with an amount of 242950. The minimal value was in 1997, that in economy is seen as a structural fracture is 104,506. If GD is called as the starting database, which stores the annual GDP in the sector of the economy from 1996 to 2013, then we follow these commands in order to complete some simple statistics and graphical representation for some.

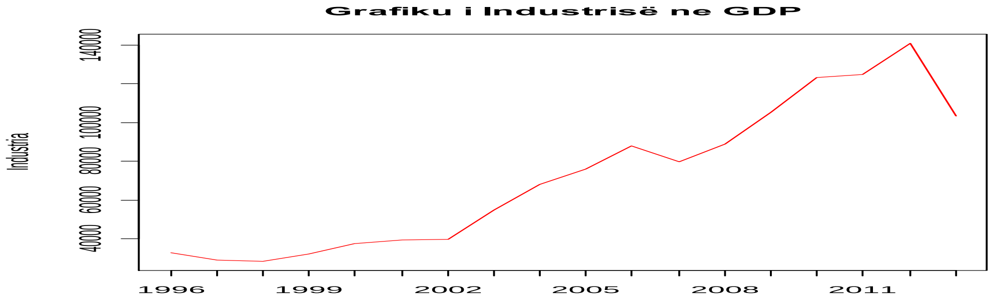


Figure. 3. Graphic performance of the industry sector, with contribution to GDP

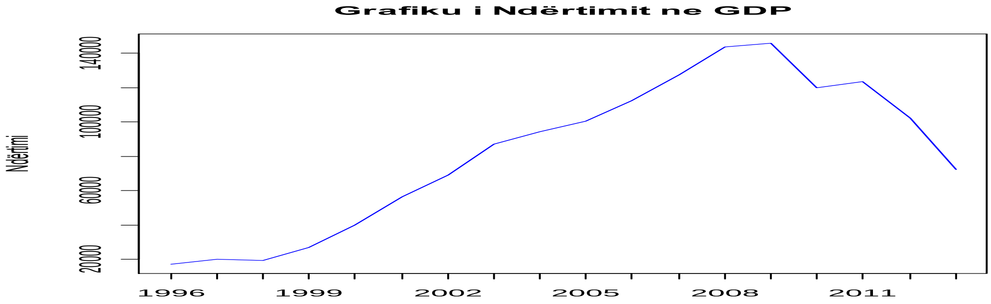


Figure 4. Graphic performance of the construction sector, with contribution to GDP

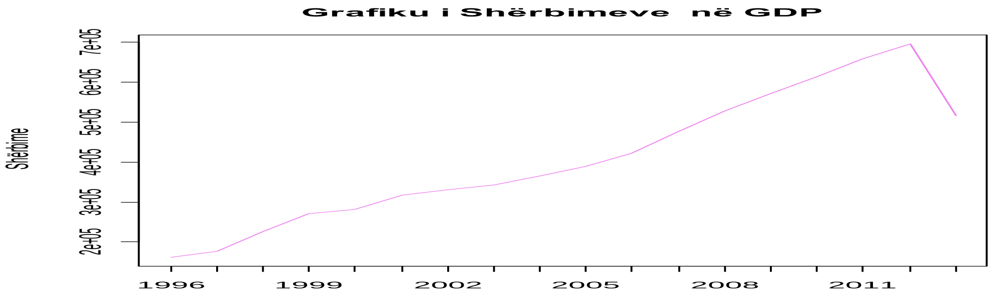


Figure 5. Graphic performance of the service sector, with contribution to GDP

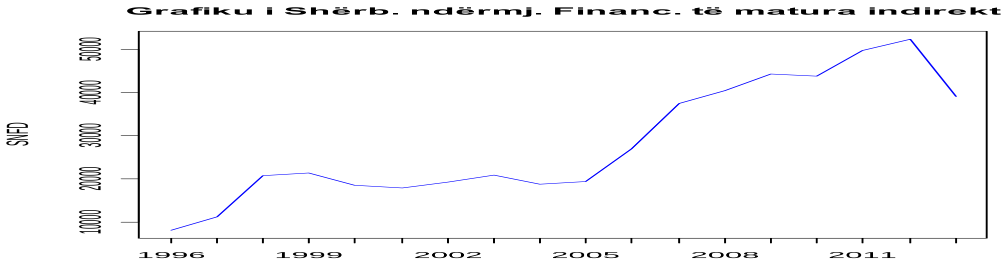


Figure 6. Graphic performance of SFMD sector, with contribution to GDP

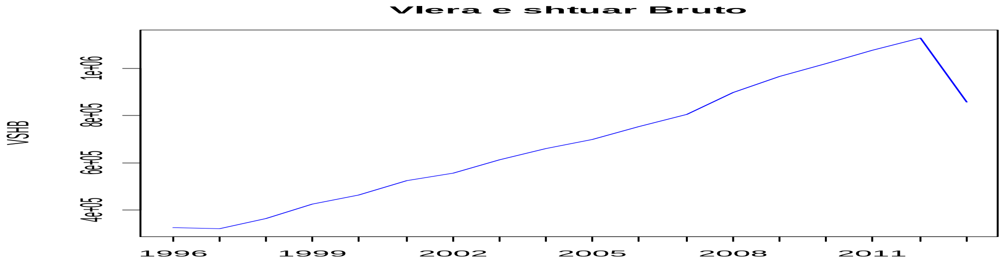


Figure 7. Graphical presentation of Gross Added Value of the Albanian economy in years

For each of the indicators that affect the GVA we will make a prediction. In order to have the opportunity to check the value, we are implementing a job with the following remarks:

1. Part of the data, until 2011, used to create the model.
2. On the built model, we predict for 2012 and 2013, on the values we have.
3. Achieve comparisons between predicted and observed values.
4. Making predictions about a longer term for a further work.

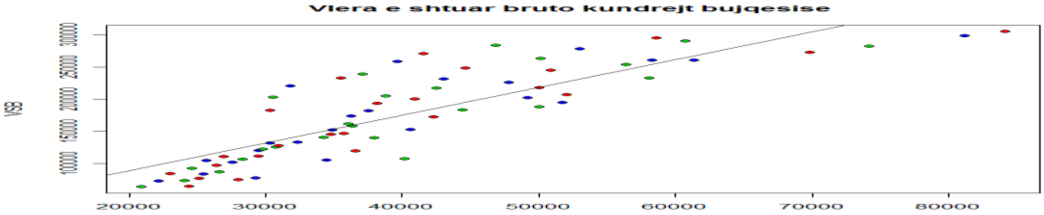


Figure 8. Graphical presentation of the relationship between agriculture and GVA

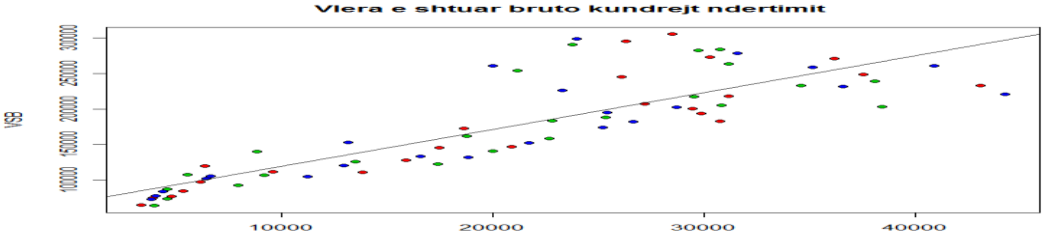


Figure 9. Graphical presentation of VSB towards building

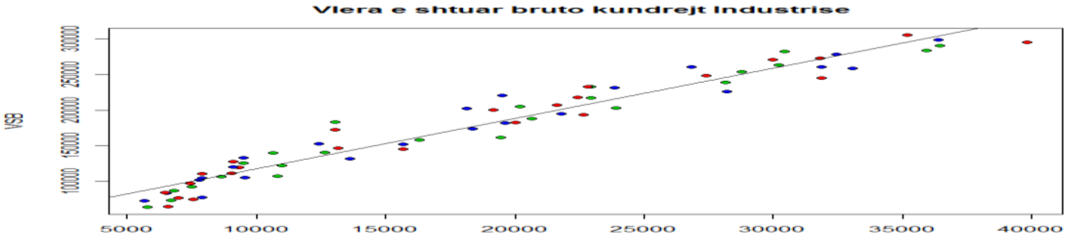


Figure 10. Graphical presentation of GVA versus industry

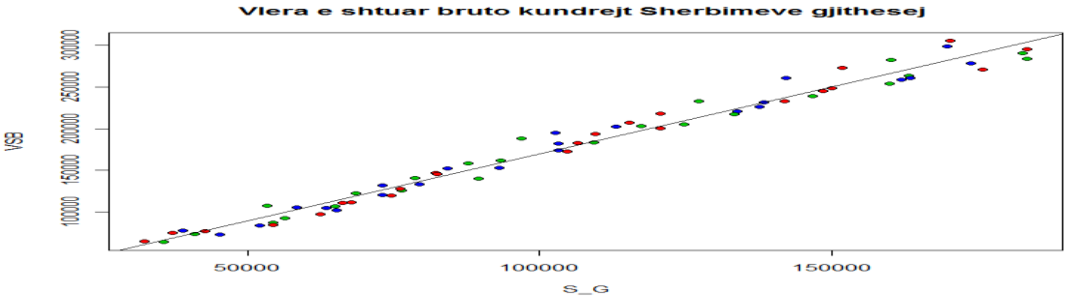


Figure 11. Graphical presentation of gross value added versus total services

We will build a model on the time series, with 70 data.

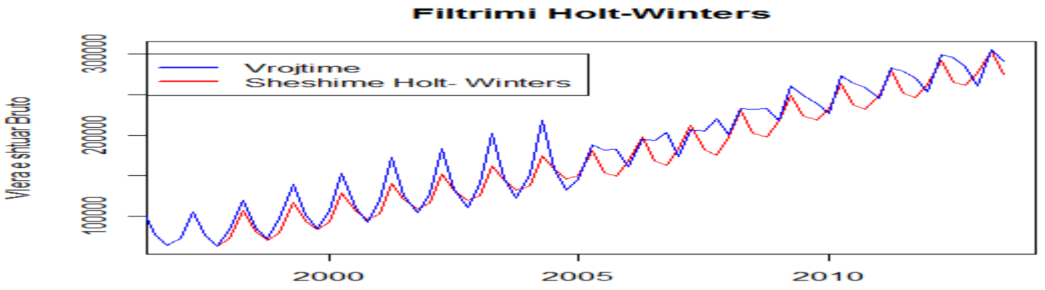


Figura 12. Graphical presentation of the observations with Holt- Winters

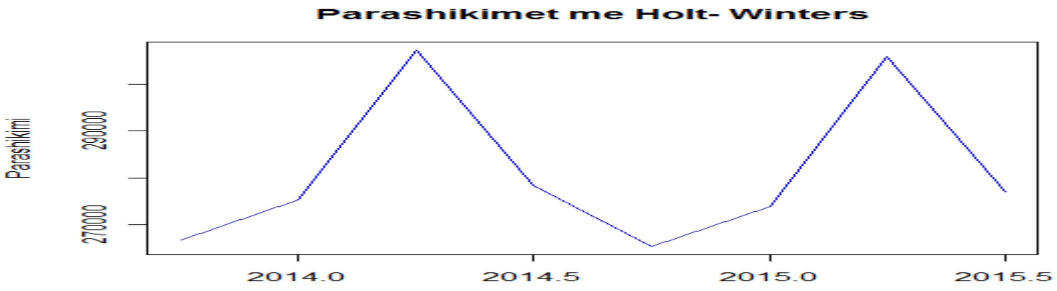


Figure 13. Graphical presentation of the predictions with *Holt- Winters*

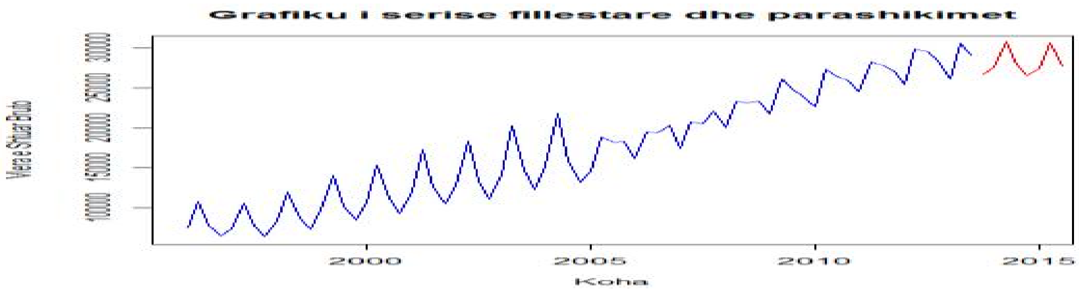


Figura 14. Graphical presentation of the observations and predictions for the GVA

If we don't take all the database as we have done above, but take the data until 2012, we build the model with the given data, and later compare predictions.

> VSB1 = ts (VSB1, start = 1996, frequency = 4). Holt- Winters builds a model on these data through the model and realize the predictions. The observed values are:
> Qtr1 Qtr2 Qtr3 Qtr4
> 2013 260546.25 305065.34 290789.02
> Building ARMA model is:
> \> X = auto.arima (VSB1)
> \> Predict Y = (X, n.ahead = 2 * 4)

With data from 1996 through 2012 ARMA adapts a model.

2.1. Comparative analysis modeling with Holt- Winters and ARIMA

In the processed data are presented predictions obtained with Holt- Ëinters model, predictions obtained with ARIMA model, and in the last column are given observations.

H = c (267700.7, 312736.2, 309250.0, 297585.3, 281869.2, 326904.7, 323418.6, 311753.8)
H = ts (H, start = 2013, frequency = 4)
full (H, main = "Graphical presentation of the observations and forecasts for some models", xlab = "Years", ylab = "observations or predictions")

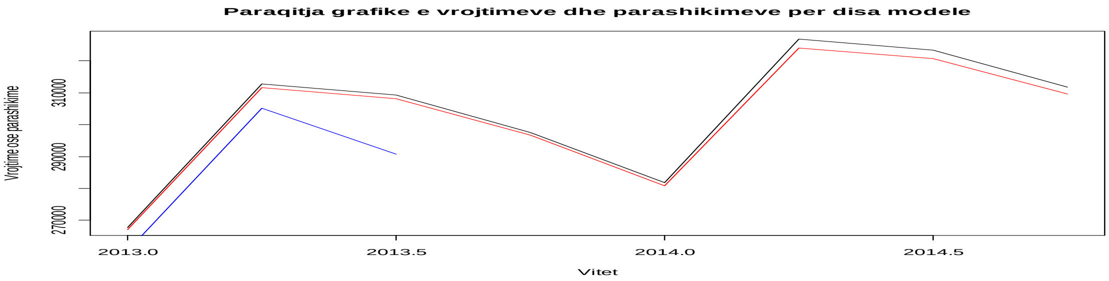


Figure 15. Graphical presentation of the observations and predictions for some models

Note: prediction of a very important economic indicator is a difficult and delicate procedure, which requires a lot of professionalism and care. Also, GBP (gross domestic product) or Gross added value that we have considered, are the economic indicators that depend on many factors. Our aim in this paper is to highlight the model and not the predictions. Its accuracy and reliability are features that require more care, starting from the number of observations and take or not consideration some important factors.

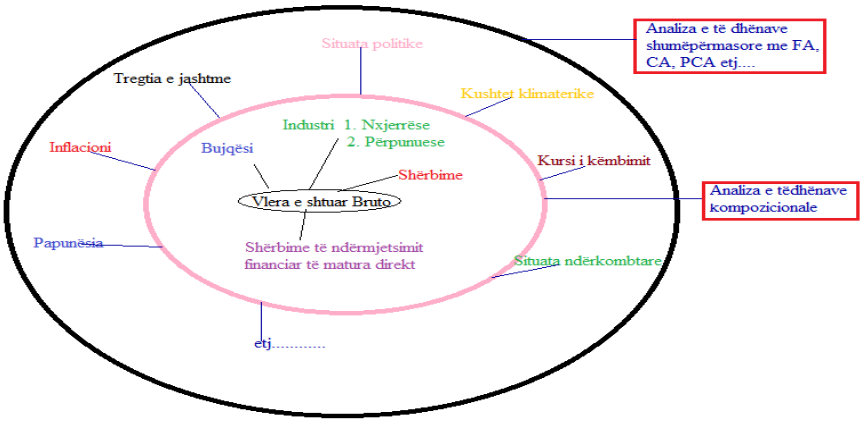


Figure 16. Schematic presentation of the difference between compositional data analysis and multidimensional data analysis

The figure above clearly shows the difference between compositional data and multidimensional data. Obviously compositional data consist of several parts and merging them gives the compositional variable under study, while the multi-dimensional data in a certain way are some external factors that affect the performance of this variable.

2.2.. The package used for the realization of work with compositional data

There are numerous possibilities for the realization of work and research in terms of Compositional Data Analysis. Starting from John Aitchison, who is recognized as one of the founders of this direction, which has implemented a package in R, that has all the necessary functionalities for all compositional data analysis.

- CoDa of John Aitchison, 1986, its written in Quick Basic and it's available along with book. It has been improved from John Bacon- Shone.
- CoDaPack freeware of Santiago Thio and Martin Fernandez, 2001 is available in excel.
- In R

1. MixeR of Batagelj and Bren, 2003
2. The package compositionas of K. Gerald van den Boogaart and Raimon, Tolosana Delgado, June 2005.

In this work we have used the CoDaPack version v2.01.14. In some cases, for verification purposes and variety of results are used the results obtained from the package Compositions in R. Recently a new package in R is developed, or more specifically a newer version.

Why did we used CoDaPack in the beginning :

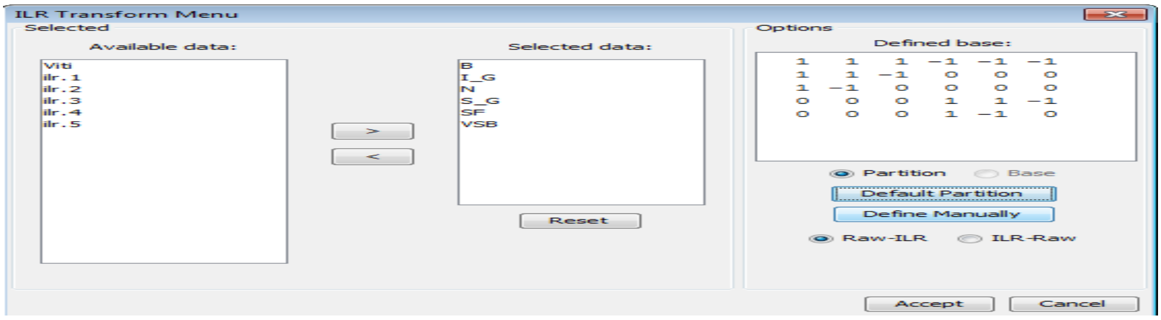
1. It's easier to be used, as SPSS.

2. Provides available commands instead of code in the command form as in R and everything is easly accomplished.

3. A couple of months ago, in (March- April 2014), in this period a lot of researches are made and it has been chosen which software was going to be used, the package CoDaPack was ready and complete, while the package Compositions in R was unusable and it can be downloaded only one of it's old versions.

2.2.1. Binary sequential separation

A useful way to build a base of simplex coordinates which can be easily interpretable is to build a sequential binary partition (SBP) of compositional vector. Each row corresponds to an order of the divisions, +1 means involvement in the part of groups G_{i1}, -1 and 0 for non-inclusion. In CodaPack software that we have used in this project, its used the binary sequential separation, whether it will be manual, so the user has to chose the binary separation.



ILR Transform Menu

Selected

Available data:

Viti
ilr.1
ilr.2
ilr.3
ilr.4
ilr.5

Selected data:

B
I_G
N
S_G
SF
VSB

Options

Defined base:

1	1	1	−1	−1	−1
1	1	−1	0	0	0
1	−1	0	0	0	0
0	0	0	1	1	−1
0	0	0	1	−1	0

● Partition ○ Base

● Raw-ILR ○ ILR-Raw

Figure 7.17. A presentation by CoDaPack software, showing possibilities to choose manually or randomly a basis for transforming ILR

2.2.2. Balances

Balances are coordinates that represent an element of simplex, of the ortho-normal basis determined by the SBP. In practice, there is no need to know exactly the expression of the base, while the coordinates are calculated using a transformation with a (ILR) and the values that we are interested for, used the inverse transformation. For the i-th partition the balance is:

$$b_i = \sqrt{\frac{r_i s_i}{r_i + s_i}} \log \frac{\left(\prod_{x_j \in G_{i1}} x_j \right)^{1/r_i}}{\left(\prod_{x_l \in G_{i2}} x_l \right)^{1/s_i}}$$

where r_i and s_i are the number of parts in the +1 and -1 group respectively. In other terms, the balance is defined as the natural logarithm of the relationship of division of the geometric mean, parts in each group.

2.2.3. Dendogram with Coda analysis

A graphical representation of a binary sequential division, together with statistical summary balances, forms a CODA dendogram. Elements of a CODA dendogram are:

1. Sharing binary sequential introduced by dendogramë type connections between parts. Columns describe groups formed by each sequence separation. The length of the lines does not present any quantitative information;

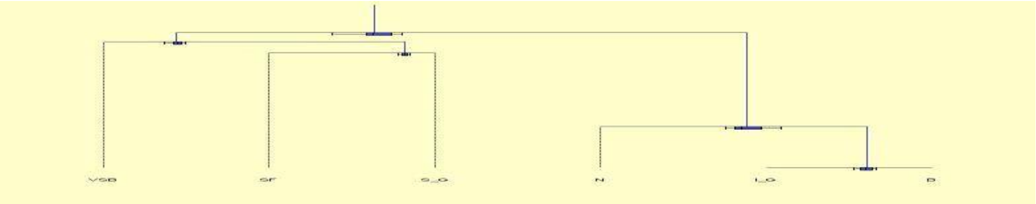


Figura 18. A dendogram for GVA data (compositional data)

2. Location of the average balance, which is determined by the grid, the vertical segments with horizontal ones.
3. The decomposition of the total variance of each choice and balance variability, presented by the length of the vertical columns. The sum of all vertical columns represents the total variance of choice. A short vertical column means that the balance has a small variability of choices, so few of the total variance explained. In contrast, a long vertical column indicates that the balance explains a good part of the total variance.

Criteria to determine a part: The initial approach is based on intuition. The question is: how should it be done when there are no criteria on how to proceed? Two auxiliary tools are quite useful:

1. Variance vector as shown in Table *
2. Biplot, as shown in Figure below.

Table. 2. % Of total variance. In the upper triangle, the variance of the parts of log-ratio, and the lower triangle, average parts of log-ratio.

	Variance ln(Xi/Xj)						
Xi\Xj	**B**	**I_G**	**N**	**S_G**	**SF**	**VSB**	**clr**
B		0.1195	0.3328	0.0718	0.132	0.052	0.0737
I_G	-0.8923		0.137	0.0406	0.0832	0.0323	0.0244
N	-0.8213	0.071		0.1471	0.2149	0.1445	0.1184

S_G	0.9045	1.7968	1.7258		0.0323	0.0044	0.0051
SF	-1.8027	-0.9104	-0.9814	-2.7072		0.0512	0.0413
VSB	1.4398	2.3321	2.2611	0.5353	3.2425		0.0031
	Mean ln(Xi/Xj)						0.2659

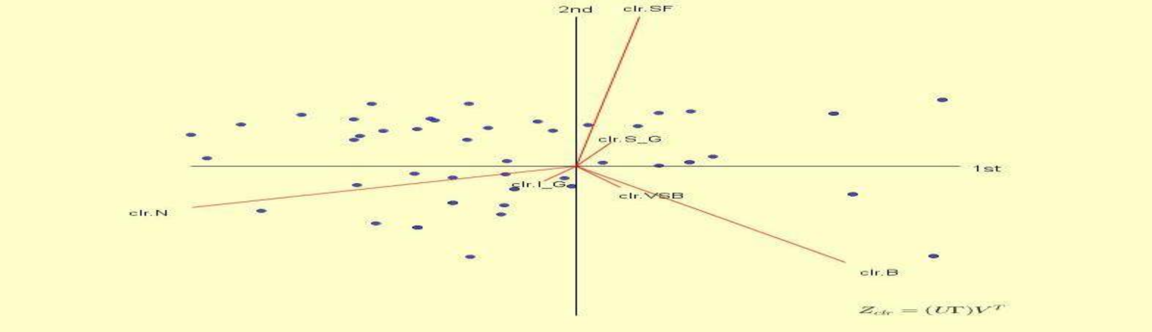


Figure 19. Biplot.

The radius length is approximately proportional with variance of CLR transformed parts. The length of the connections between the ends of the radius, are approximately proportional with the variance of the corresponding log-ratio. Compositional Biplot, is taken as a standard biplot covariance for centralized data log ratio (CLR).

Conclusion 1: The above analysis conducted in CoDaPack ends here, because all further analyzes will be performed by the package, "Compositions in R '.

2.3. Descriptive statistics of compositional data

Below we are creating a detailed analysis of compositional data. We find arithmetic average of the transformed data with acomp.

```
> x=acomp(x)

> mean(x)

          B            I_G              N             S_G             SF            VSB

 0.114183619   0.046783525   0.050226198   0.282114747   0.018824198   0.481835696
```

In the multi-dimensional real analysis, is typical to center data by subtracting the average: in compositional data analysis, we realize acting opposite of the center.

```
> mean(x-mean(x))

        B            I_G            N             S_G             SF           VSB

 0.1428571   0.1428571    0.1428571   0.1428571    0.1428571    0.1428571
```

Average of centralized database is a neutral element of the simplex, which is a vector with the same value in each component.

2.4. Matrix of variances

Variance metric does not contain information about the dependence of components. Additional information gives the variance matrix: > variation(x)

Table 3. Variance matrix of GVA data

	B	I_G	N	S_G	SF	VSB
B	**0.00000000**	0.12117281	0.3375652	0.072780553	0.13383998	0.052769855
I_G	0.12117281	**0.00000000**	0.1389336	0.041177923	0.08433911	0.032757769
N	0.33756515	0.13893361	**0.0000000**	0.149196371	0.21795077	0.146536401
S_G	0.07278055	0.04117792	0.1491964	**0.000000000**	0.03280897	0.004506815
SF	0.13383998	0.08433911	0.2179508	0.032808973	**0.00000000**	0.051958977
VSB	0.05276985	0.03275777	0.1465364	0.004506815	0.05195898	**0.000000000**

Components of the matrix are $\tau_{ij} = \mathrm{var}\left(\ln \frac{x_i}{x_j}\right)$

The matrix is symmetric, a small value indicates a small variance. For interpretation, Aitchison has proposed an index, which is interpreted as the correlation coefficient. If we seek a statistical summary acomp, we will have: > summary(x)

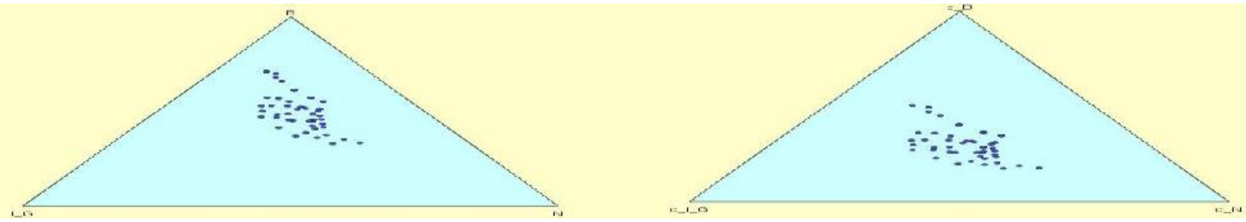


Figure 20. The difference between the uncentered and centered data

2.5. Predictive modeling, using log ratio transformations

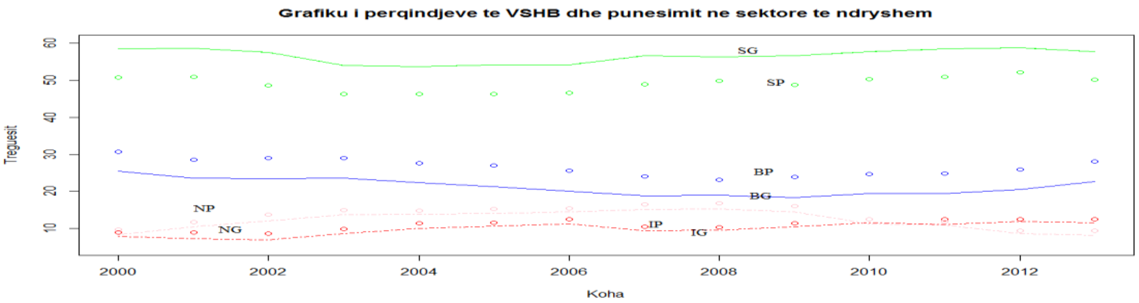


Figure 21. Graphical presentation of employment and GAV indicators in different sectors from 2000 to 2014

Build a linear regression model where employment in agriculture depends on the percentage of GAV in all other sectors.

B_P = 6.08 +1.059* B_G +0.086* I_G - 0.051* S_G (1)

Reliability of the model (1) is 97.8%. We Build a model that shows the dependence of regression employment in the Industry sector of GAV in other sectors:

I_P=4.157+ 0.866*I_G-0.026*S_G-0.019*N_G (2)

Reliability of the model (2) is 97.4%. We Build another model that shows the dependence of regression employment in construction sector of GAV in other sectors:

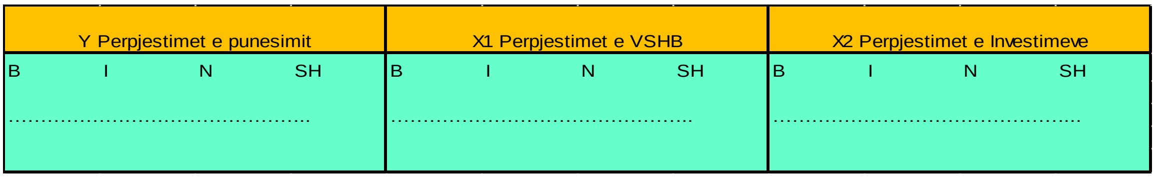
N_P= -2.341-0.073*I_G+0.065*S_G+1.046*N_G (3)

Reliability of the model (3) is 99.4%. We Build a model that shows the dependence of regression employment the service sector of GAVfrom other sectors:

S_P=-13.838+0.181*I_G+1.072*S_G+0.032*N_G (4)

2.6. The challenge in compositional data analysis.

Construction of the above models did not change from the analysis of real multidimensional data, but if we consider a partition of the data according to the following table, then the situation is very difficult:



Y Perpjestimet e punesimit				X1 Perpjestimet e VSHB				X2 Perpjestimet e Investimeve			
B	I	N	SH	B	I	N	SH	B	I	N	SH
....				....				....			

Figure 22. Presentation of structured employment data, GAV and Investments to build a multidimensional model of compositional data type

Compositional data analysis is a new area and is needed to do more in order to improve and spread in different applications. Compositional analysis is a difficult operation on the concept of full implementation of all the theoretical side. Also, we encountered difficulties in interpretation of results.

Further study would be a complete analysis including hypothesis testing and confidence intervals and other analyzes that are already known or time series multidimensional data.

3 Conclusions

As a conclusion we have achived to show which is the difference between compositional data and simple data and not only to study the nature of this data, but even to give situations and propose ways haw to deal with them.

Compositional data analysis is a difficult analysis compared with other type of analysis, because it shows a lot of unpredictive situations and problems. This are working with zeros, spurious correlations and transformations of data.

We have presented some transformations in this paper and we have applied them in the data we have analysed. As a conclusion we can say that those transformations allow us to make predicitions and all statistical analysis over those transformed data. The only problem that we are going to have with those transformed data is the interpretation, which is very difficult and delicat. It requires a lot of care.

Furthermore we need to add that we can not talk about precise values when we talk for GDP, because it is a very delicat economic indicator. All our results are in order to show and to make known this new kind of data and some ways how to deal with them.

References

[1] Aitchison, J., A concise guide to Compositional Data Analysis

[2] Bren, M., Batagelj, V., Compositional Data Analysis with R

[3] Buccianti, A., Mateu- Figueras, G., Pawlowsky- Glahn, V., Compositional Data Analysis in the Geosciences

[4] Comas, M., Thio- Henestrosa, S., CoDaPack 2.0: a stand- alone, multi- platform compositional software

[5] Gerald van den Boogart, K., Tolosana- Delgado, R., Analyzing Compositional Data with R

[6] INSTAT, Produlti i Brendshwm Bruto Tremujor, Tetor 2013

[8] Pawlowsky- Glahn, V., Egozcue, J. Tolosana- Delgado, R., Lecture Notes on Compositional Data Analysis

[9] http://data.worldbank.org/country/albania

[11] http://ima.udg.edu/codapack/assets/codapack-manual.pdf

[12] http://www.instat.gov.al/al/figures/statistical-databases.aspx

[13] http://www.bankofalbania.org/