"Dimitrie Cantemir" Christian University
Knowledge Horizons - Economics
Volume 5, No. 4, pp. 56–61
P-ISSN: 2069-0932, E-ISSN: 2066-1061

www.orizonturi.ucdc.ro

Analysis of Variance (Anova) Used to Verify the Significance of the Development Regions in Characterizing the Standard of Living and Quality of Life

Claudia Gabriela BENTOIU[1], Emilia GOGU[2], Gratiela GHIC[3], Alexandra MORARU[4]

[1,3,4]Faculty of International Business and Economics, "Dimitrie Cantemir" Christian University, Bucharest, Romania, [1]E-mail: bcg@yahoo.com, [3]E-mail: gratiela72@yahoo.com, [4]E-mail: alexandramoraru2002@yahoo.com
[2]Faculty of Tourism and Commercial Management, "Dimitrie Cantemir" Christian University, Bucharest, Romania, [2]E-mail:arina_emilia@yahoo.com

Abstract *According to the specific literature, the national and international statistics for ranking countries (regions) we use the ranking method. The ranking method has the advantage that the hierarchy of territorial units can use local variables expressed in absolute values, no matter the structure or the intensity. Territorial units are complex statistical units in which the variable values are obtained either by aggregation or as relative values based on a report, either as medium measures. Being partial synthetic values calculated at the local level it is almost impossible to know the analytical function expressing the shape and intensity of the relationship that shaped distribution by regions within each region to be in the same relation of interdependence.*

Key words:
Analysis of variance, standard of living and quality of life

JEL Codes:
C1, C8, O47

1. Introduction

In order to verify the significance of territorial criteria for the classification of the indicators we used variance analysis taking on all the variables recorded. After we analyzed the variation of statistical indicators characterizing the standard of living and quality of life in the regions considered as independent variables, we aimed to analyze, regionally, the interdependencies between them. Statistical data being centralized at the level of complex units such as developing regions, the results must be interpreted properly as correlation indices which measure the tendency of achieving Statistics.

2. The analysis of variance of the population. Procedures and results

To this end we systematically grouped data according to the unifactorial model variance analysis, using the appropriate groups in developed regions for each variable separately.

For the variance analysis of the population we calculated the conditional average on the eight groups corresponding to eight regions and general multi-annual average. On this basis we determined the last two columns of Table 1 Empirical squared deviations from the mean values of the group and to the overall average.

Table 1. Analysis of regional variation in population during 2000-2010, according to the region

Development region	Year	Number of people	$(y_j - \bar{y}_i)^2$	$(y_j - \bar{y}_0)^2$
NORD-VEST	2000	2849982	8813147159	16406279569
	2001	2844430	7801546513	15014881923
	2002	2762565	41749220.04	1654067386
	2003	2750406	32463060.13	812890081
	2004	2743281	164420003.3	457370717
	2005	2742676	180301418.3	431859407
	2006	2729181	724828348.8	53089108
	2007	2729256	720795578.3	54187667

Development region	Year	Number of people	$\left(y_j - \bar{y}_i\right)^2$	$\left(y_j - \bar{y}_0\right)^2$
	2008	2724176	1019373964	5203998
	2009	2721468	1199627306	182135
	2010	2719719	1323841763	4733987
	TOTAL	**30317140**	**22022094335**	**34894745977**
CENTRU	2000	2644115	8075718225	6049693046
	2001	2642475	7783650625	6307500300
	2002	2551234	9096256	29125099348
	2003	2548331	35034561	30124383203
	2004	2543512	115304644	31820413606
	2005	2533421	433847241	35522363006
	2006	2534378	394896384	35162540054
	2007	2524176	904445476	39092713089
	2008	2524628	877462884	38914179622
	2009	2526062	794563344	38350474874
	2010	2524418	889948224	38997075767
	TOTAL	**28096750**	**20313967864**	**329466435915**
NORD-EST	2000	3820101	5101959184	1206056917621
	2001	3831216	6813346849	1230593585278
	2002	3743950	22306729	1044596887596
	2003	3746330	5489649	1049467534877
	2004	3742868	33698025	1042386330808
	2005	3735512	173211921	1027419883424
	2006	3734946	188430529	1026272789079
	2007	3727910	431102169	1012066637505
	2008	3722553	682254400	1001316887809
	2009	3717621	964226704	991470719679
	2010	3712396	1316020729	981092681229
	TOTAL	**41235403**	**15732046888**	**11612740854903**
SUD-EST	2000	2936219	6021830546	45934874396
	2001	2934890	5817334778	45366966841
	2002	2872007	179250715.1	22533680777
	2003	2863406	22919721.02	20025427444
	2004	2855044	12777375.21	17728716723
	2005	2849959	74987727.48	16400446307
	2006	2843624	224836393.4	14818004772
	2007	2834335	589690579.8	12642804709
	2008	2825756	1079946894	10787154531
	2009	2818346	1621877917	9302839242
	2010	2811218	2246811709	7978638930
	Total	**31444804**	**17892264357**	**223519554673**
SUD-MUNTENIA	2000	3471322	14837233158	561641168978
	2001	3467375	13891258180	555740769255
	2002	3383573	1160027866	437818076447
	2003	3368615	364855146.9	418247052364
	2004	3350248	539022.9421	394827778224
	2005	3338195	128115645	379825970136
	2006	3321392	790836657.9	359396925508
	2007	3304840	1995750031	339825138000
	2008	3292036	3303699583	325061019036
	2009	3279786	4861968628	311242621468
	2010	3267270	6764045629	297434138523
	Total	**36844652**	**48098329548**	**4381060657938**
BUCURESTI - ILFOV	2000	2285544	2390836596	190401996860
	2001	2272972	1319446185	201531655873
	2002	2213782	522845641.1	258178589809
	2003	2208150	812125641.1	263933691505
	2004	2208254	806208910.9	263826843408
	2005	2209768	722524625.5	262273831344
	2006	2215701	438769191.9	256232135548
	2007	2232162	20122564.76	239838188683
	2008	2242002	28667262.94	230297073316
	2009	2253093	270444005	219775102112
	2010	2261698	627511609.1	211781069629
	Total	**24603126**	**7959502234**	**2598070178086**
SUD-VEST OLTENIA	2000	2403632	7483539692	101291192504
	2001	2399333	6758229999	104046097225
	2002	2348337	974217318.8	139545409565
	2003	2336018	356962624.7	148900883730

Development region	Year	Number of people	$(y_j - \bar{y}_i)^2$	$(y_j - \bar{y}_0)^2$
	2004	2325020	62338202.48	157509585227
	2005	2313903	10378355.12	166457286613
	2006	2301833	233831362.4	176451892907
	2007	2285733	985429126	190237091989
	2008	2270776	2148187666	203508147107
	2009	2257752	3525099154	215428513475
	2010	2246033	5054007835	226444426743
	Total	**25488370**	**27592221335**	**1829820527085**
VEST	2000	2044570	8161890501	458768847750
	2001	2037766	6978794899	468032177673
	2002	1958035	14502248.76	583481752391
	2003	1951518	7337695.942	593480371958
	2004	1943025	125480730.6	606638122868
	2005	1935094	366064731.6	619055455964
	2006	1929158	628445645	628431590834
	2007	1926707	757340392.8	632323593895
	2008	1926700	757725719.2	632334726573
	2009	1924488	884397306.9	635857561191
	2010	1919434	1210540197	643943291766
	Total	**21496495**	**19892520068**	**6502347492864**
Total general		**239526740**	**179502946626.9**	**27511920447440**

Table 2. Determining variance indicators

Development region	Average level $\bar{y}_i$	$\sum(y_j - \bar{y}_i)^2$	$\sum(y_j - \bar{y}_0)^2$	$\sum(\bar{y}_i - \bar{y}_0)^2$
Nord Vest	2756103.6	22022094335	34894745977	12872661972
CENTRU	2554250.0	20313967864	329466435915	309152568637
NORD-EST	3748673.0	15732046888	11612740854903	11597008191948
SUD-EST	2858618.5	17892264357	223519554673	205627071559
SUD-MUNTENIA	3349513.8	48098329548	4381060657938	4332961700771
BUCURESTI - ILFOV	2236647.8	7959502234	2598070178086	2590111161099
SUD-VEST OLTENIA	2317124.5	27592221335	1829820527085	1802228953383
VEST	1954226.8	19892520068	6502347492864	6482455740464
TOTAL	**2721894.8**	**179502946626.9**	**27511920447440**	**27332418049833**

Thus, according to the previous calculation:

- the general average is:

$$\bar{y}_0 = \frac{239526740}{88} = 2721894{,}8$$

- the residual deviation is:

$$\Delta^2_{y/r} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_j - \bar{y}_i)^2 = 179502946626{,}9$$

- and the overall deviation is:

$$\Delta^2_y = \sum_{j=1}^{n_i}(y_j - \bar{y}_0)^2 = 27511920447440$$

- the factorial deviation ai calculated by the formula:

$$\Delta^2_{y/x} = \sum_{i=1}^{k}(\bar{y}_i - \bar{y}_0)^2 \cdot n_i = 27332418049833$$

With the help of the three deviances we calculated the dispersion analysis indicators.

Table 3. The dispersion analysis indicators

Indicators	Nature of variation		
	Factorial	Residual	Total
Deviance	$\Delta^2_{y/x} = 27332418049833$	$\Delta^2_{y/r} = 179502946626.9$	$\Delta^2_y = 27511920447440$
The degree of influence	$\eta^2_{y/x} = \frac{\Delta^2_{y/x}}{\Delta^2_y} = 0{,}9935$	$\eta^2_{y/r} = \frac{\Delta^2_{y/r}}{\Delta^2_y} = 0{,}0065$	$\eta^2_{y/x} + \eta^2_{y/r} = 1$
Corrected variance	$s^2_{y/x} = \frac{\Delta^2_{y/x}}{n_x} = 2484765277258$	$s^2_{y/r} = \frac{\Delta^2_{y/r}}{n_r} = 2039806212$	$s^2_y = \frac{\Delta^2_y}{n_y} = 3438990055930$

Based on these results, we can say that in the period 2000-2010, 99.35% of the total variance of the population is determined by the variation produced by the grouping factor (development region) and the remaining 0.65% is the relative influence of time variation.

To check the significance of the grouping factor we shall compare the coefficient F calculated with the spreadsheet.

Thus, the coefficient F is calculated as the ratio between two corrected dispersions: $F_{calc} = \frac{s^2_{y/x}}{s^2_{y/r}} = 121,13$. For a significance level of 1%, $F_{tab(n_x=11,n_r=88)} = 3.60$.

As F_{calc} is higher than F_{tab}, we can state that the development region (grouping factor) is relevant for the population variation in the region.

Two- factor ANOVA without replication allows us to add a second dimension to the comparison of means across treatment groups accomplished in single- factor ANOVA. This analysis assumes the two dimensions are independent of one another. Like single- factor ANOVA, two- factor ANOVA without replication is an additive model in that the total sum of squares is parsed between explained variation and unexplained variation. By adding a second factor to the experimental design, two- factor ANOVA without replication allows us to account for additional fluctuation in the dependent variable that we suspect is associated with a second variable but that is not fully explained by the primary variable. In adding a second factor to the analysis, we further reduce the unexplained within- group variation, which in turn, which leads to a smaller mean square error in the denominator to the F-statistic, and a stronger test result.[1]

Considering the second systematic factor (years), we applied two-factor ANOVA model using EXCEL and have achieved the same level of significance results in the table below.

Table 4. Two-factor ANOVA model (region and years) for the number of staff

Anova: Two-Factor without Replication

ANOVA			
Source of Variation	SS	df	MS
Rows	88844288839	9	9871587649
Columns	27332418049833	7	3.54387E+12
Error	23683887221	63	375934717.8
Total	27511920447440	79	

ANOVA			
Source of Variation	F	P-value	F crit
Rows	26.2587816	4.13573E-18	2.032242211
Columns	9426.809865	1.36389E-92	2.158828996
Error			
Total			

From the table it is observed that the values of the coefficient F calculated in both the year and depending on the development region is smaller than the corresponding F spreadsheet. It follows that there is no significant difference in the time evolution of the population on a regional basis.

Based on the results from the application of analysis of variance for the indicators previously characterized by regions, the interdependence between them was examined. Next, we proceeded to estimate the intensity of the relationship with parametric correlation.

The complexity of dependencies in the economy implies identifying the existence of links. Statistical expression of links by the regression analysis and by calculating the correlation indicators is possible only if the anticipated establish causal links between investigated phenomena. The correlation method applies when the variables of a financial – economic model presents causal relations of a probabilistic – statistical nature. The practicality of this approach, justifying its usefulness both in post-factum analysis and forecasting, with the aim of extrapolating trends of future economic phenomena. Due to the complexity of the case, the mathematical formulation of statistical dependencies is inextricably linked with some simplifications. Thus, the correlation model includes one or more key impact factors. When seeking evidence of causal relationship between economic phenomena, we can meet different situations: there is a direct connection between phenomena, unilateral or reverse reactions (interdependence), a covariance due to common causes of phenomena, a simple parallelism chance to change two or more events. Probabilistic relationships between financial – economic phenomena and default correlation calculations that are necessary in this context can be conclusive only on the basis of statistical homogeneous data and that these phenomena reflect developments over longer periods of time[2].

Based on a qualitative analysis we considered that there are causal links between the population and the territorial area.

[1] Priscilla Chaffe-Stengel; Donald N. Stengel (2011). *Working With Sample Data: Exploration and Inference*, Business Expert Press

[2] Cojocaru C. (1997). *Analiza economico–financiară a exploatațiilor agricole şi silvice*, Ed. Economică, Bucureşti, p. 54

Table 5. Distribution area, population and population density by regions in 2010

Region	Area (km²)	Population (pers.)	Population density (pers./ km²)
1. Nord-Est	36850	3712396	100.74
2. Sud-Est	35762	2811218	78.61
3. Sud-Muntenia	34453	3267270	94.83
4. Sud-Vest Oltenia	29212	2246033	76.89
5. Vest	32034	1919434	59.92
6. Nord-Vest	34159	2719719	79.62
7. Centru	34100	2524418	74.03
8. Bucureşti-Ilfov	1821	2261698	1242.01
Total	**238391**	21462186	90.03

Although the Bucharest-Ilfov greatly outdistances the others according to territorial density, it should not be allowed out as outliers because its situation among other regions does not change link.

From the chart link it is estimated correlation is linear, and therefore the average equation of the regression function[3] is: $Y = a + bx$

Where:

y – is the value of the dependent variable;

x – is the independent variable;

a, b – the regression parameters (have average size) and show what level would be reached by the dependent variable if all factors had constant action.

The normal equation which is based on determining two parameters *a* and *b* is as follows:

$$na + b\sum x = \sum y$$

$$a\sum x + b\sum x^2 = \sum xy$$

By solving this system of equations we obtain:

$$b = \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$a = \bar{y} - b\bar{x}$$

After determining the parameters a and b and solving the system of normal equations we can calculate theoretical values of the dependent variable y for each concrete value of the x factor, other factors being considered constant.

Accessing the Data Regression function in Excel Analysis, we determined the normal equation system in which we obtained the results presented below.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.99042406
R Square	0.98093981
Adjusted R Square	0.97776311
Standard Error	999726.017
Observations	8

ANOVA

	df	*SS*
Regression	1	308623059053480
Residual	6	5996712651802
Total	7	314619771705282

ANOVA

	MS	F	Significance F
Regression	308623059053480	309	0
Residual	999452108634		
Total			

	Coefficients	*Standard Error*	*t Stat*
Intercept	29307.098	449228.2	0.065239
36850	88.5989635	5.041914	17.57249

	P-value	*Lower 95%*	*Upper 95%*
Intercept	0.950103	-1069915	1128529
36850	2.18E-06	76.26184	100.9361

RESIDUAL OUTPUT

Observation	*Predicted 3712396*	*Residuals*	*Standard Residuals*
1	3197783.23	-386565	-0.41765
2	3081807.19	185462.8	0.200378
3	2617460.02	-371427	-0.4013
4	2867486.3	-948052	-1.02429
5	3055759.09	-336040	-0.36306
6	3050531.75	-526114	-0.56842
7	190645.811	2071052	2.237605
8	21150502.6	311683.4	0.336749

The table shows that the parameters have the values: *a* = 29307.098 şi *b*= 88.5989635

So, the equation for estimating average linear relationship between the population and area, by region, is: Y_x = 29307,098 + 88,5989635*x* .

Based on the average trend equation it can be said that, in 2010, in an area of one km2 average density is 89 people on a regional basis.

The Standard Error of the two estimated parameters are: s_a =449228.2 and s_b =5.041914. These errors are used to calculate statistical significance for testing estimators t (t_a =0.065239 şi t_b =17.57249). Because *p* associated values are very close to zero, it can be said that the estimators are significant, the confidence interval for the parameter *b* is (76.26184; 100.9361).

To measure the intensity of the relationship to the ensemble we aim at the correlation ratio as the main synthetic indicator of correlation: $R_{y/x}$ =0.99042406

[3] Based on theoretical values of the equation we can make judgments about the phenomenon of change and the trend of the analyzed determining factor undergoing variation and the existence and direction of the correlation between the two variables studied (depending on the size and direction of the regression coefficient)

(Multiple R). Considering all other factors constant, the size of the population is due at a rate of 98.09% to changes in local surface.
The validity of the regression model is tested using the F test. The ANOVA table presents three variations: one explained by the model, the residual and the total. With their help calculate F statistics (309). Since F is greater than F calculated spreadsheet (Significance F) we can validate the regression model.

3. Conclusions

We used the analysis of variance with the factor-group-development region for the time variation of the studied characteristics. Based on the results from the application of analysis of variance for the indicators previously characterized by regions, we examined the interdependence between them. We estimated the intensity of the relationship with parametric correlation. Based on a qualitative analysis there are causal links between the population and the territorial area and we considered necessary to use the analysis of variance (ANOVA) in order to study the degree of variation of population and area, depending on the variation produced by the grouping factor (development region) for the period 2000-2010.

Notes (Translation studies)
The present article is a functional texts and its purpose is to inform the audience by the help of mathematical demonstration. The difficulty of the translation process consisted mainly in rewriting the word order for the target language and avoiding Latin phrase structures which are specific to the Romanian language (the source language), therefore avoiding the law of interference. Basic terminology has created few problems, as the interpreter is not familiar with the specific terms used in statistics and mathematics. Therefore there has resulted a words breviary consisting of: ranking method, aggregation, variance analysis, factor-group-development region, unifactorial model variance analysis, empirical squared deviations, overall average, dispersion analysis indicators, and probabilistic relationships.

References

1. Begu, L.B. (1999). *Statistică internațională*, Editura All Beck, Bucureşti.
2. Cojocaru C. (1997). *Analiza economico – financiară a exploatațiilor agricole şi silvice*, Ed. Economică, Bucureşti.
3. Priscilla Chaffe-Stengel; Donald N. Stengel (2011). *Working With Sample Data: Exploration and Inference*, Business Expert Press
4. Mărginean, I., Bălaşa A., (2002). *Calitatea vieții în România*, Editura Expert, Bucureşti.
5. M. Mureşan, E.Gogu R. Irimia (2010). „*Elaborarea strategiei de dezvoltare regională - proces participativ bazat pe cunoaştere*"- manual bilingv română –engleză Editura Pro Universitaria Bucureşti.