WWW.JAQM.RO

JOURNAL OF APPLIED QUANTITATIVE METHODS

Quantitative Methods Inquires

Vol. 9
No. 4
Winter
2014

ISSN 1842–4562

JAQM Editorial Board

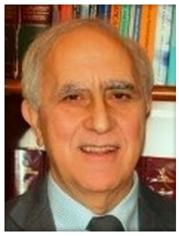
JAQM Advisory Board

Contents

PRINCIPAL COMPONENT ANALYSIS TO RANKING TECHNICAL EFFICIENCIES THROUGH STOCHASTIC FRONTIER ANALYSIS AND DEA

Sergio SCIPPACERCOLA
Associate Professor, Department of Economics, Management, Institutions
University of Naples "Federico II", Italy

E-mail: sergio.scippacercola@unina.it



Enrica SEPE
PhD Candidate, Department of Economics, Management, Institutions
University of Naples "Federico II", Italy

E-mail: enrica.sepe@unina.it

Abstract

The Stochastic Frontier Analysis permits evaluating the Technical Efficiency scores for one output variable to obtain the corresponding Technical Efficiency of n Decision-Making Units (DMU). The objective of this work is a comparison between a Stochastic Frontier Analysis, with same input and different output variables, and the Data Envelopment Analysis. You get k Technical Efficiency TE(yi) which are unified by a Principal Component Analysis and compared with the results of a DEA on the same data.

Keywords: Principal Component Analysis; Stochastic Frontier Analysis; Technical Efficiency; Data Envelopment Analysis; Secondary Schools

1. Introduction

The evaluation of Technical Efficiency (*TE*) is a fundamental tool for seeing which determinants slow down the development of production. We have two distinct approaches to evaluating Technical Efficiency, namely a parametric approach which is Stochastic Frontier Analysis (*SFA*) and a non-parametric deterministic approach which is Data Envelopment Analysis (*DEA*). DEA is an approach which uses mathematical programming to identify the efficient frontier, and does not impose functional forms (Kumbhakar, Lovell, 2003; Ray, 2004; Cooper, 2006). The main advantage of DEA is that it does not require any hypothesis about the analytical form of the production function. In DEA we have many inputs and many outputs jointly considered. DEA is based on the chosen inputs and outputs of entities that are named Decision-Making Units (DMUs). For example, all the schools (DMUs) are compared in relationship to the "best" performing schools. DEA is a non-parametric linear programming method for assessing the efficiency of (DMUs).

SFA requires strong distribution assumptions of both statistical random errors (i.e. normal distribution) and non-negative technical inefficiency random variables. SFA considers many input variables (x_1, x_2, ..., x_k) but only one output variable (y). Our proposal is very interesting when we have more than one output variable and with SFA cannot be considered jointly as happens with DEA. Our goal is therefore to make more SFAs and unify *TEs* into a single list as for DEA. So, the main objective of this work is to obtain a single ranking of different SFAs**.**

Section 2 introduces the Stochastic production frontier methodology and Section 3 the Data Envelopment Analysis. Afterwards, in Section 4 we suggest how to organize the SFA with the same input and different outputs in order to obtain a synthetic indicator of efficiency instead of many outputs in accordance with the hypothesis of the stochastic model; an application follows of the methodology on a real case (Secondary Schools) with a brief discussion of the main findings. Finally, in Section 5, our comparison between DEA and SFA is presented.

2. Stochastic Frontier Analysis

The SFA is a *parametric approach* that hypothesizes a functional form and uses the data to econometrically estimate the parameters of this function. SFA requires functional forms on the production frontier, and assumes that units may deviate from the production frontier not only owing to technical inefficiency but also to measurement errors, statistical noise or to other non-systematic factors. In addition, the SFA requires strong distribution assumptions of both statistical random errors (i.e. normal distribution) and the non-negative technical inefficiency random variables (i.e. half-normal or truncated normal distribution) (Coelli et. al., 2005). The Stochastic Frontier Analysis searches for the production function, which represents the maximum output attainable given a certain quantity of inputs (Rao et al., 2005).

The *first step* of SFA consists in the specification and in the estimation of the stochastic frontier production function as well as in the estimation of technical inefficiency effects, assuming that these inefficiency effects are identically distributed. SFA methodology allows a functional form and the breakdown of the inefficiency error term. SFA is a parametric approach that hypothesizes a functional form and uses the data to econometrically estimate the parameters of this function. A production function ***f*** is defined as the schedule of the maximum amount of output that can be produced from a specified set of inputs, given the existing technology. The model of the Stochastic Frontier Analysis is (Rao et al., 2005):

$$\ln y_i = x_i'\beta + v_i - u_i \qquad (1)$$

where $\boldsymbol{y_i}$ is the output of the *n-th* producer (i.e. DMU), $\boldsymbol{x_i}$ is a vector of inputs, $\boldsymbol{\beta}$ is a vector of $k+1$ parameters to be estimated, $v_i \approx iid\, N(0, \sigma_v^2)$ is the noise or error term or the measure of effects independent of the producer, $\boldsymbol{v_i}$ is homoskedastic; $\boldsymbol{u_i}$ is *iid*, $\boldsymbol{u_i}$ is a non-negative random variable measuring the technical inefficiency with $N^+(0, \sigma_u^2)$ (half-normal either normal-truncated model $N^+(\mu, \sigma_u^2)$ or exponential or gamma); $\boldsymbol{v_i}$ and $\boldsymbol{u_i}$ are distributed independently of each other and of the regressors. We can define the Technical Efficiency (*TE*) as the ratio of realised output to the stochastic frontier output:

$$\ln TE_i = \ln y_i - \ln y_i^* = \ln(y_i / y_i^*) = -u_i \qquad (0 \le TE \le 1) \qquad (2).$$

The parameters of stochastic frontier function are estimated by the maximum likelihood method. An estimation of stochastic frontier is the use of the γ (Battese and Corra,1977):

$$\gamma = \sigma_u^2/(\sigma_u^2 + \sigma_v^2).$$

When the parameter $\gamma = 0$ the variance of the technical inefficiency effect is zero, if γ *is close to one* it indicates the deviations from the frontier are due mostly to technical inefficiency, and if $\gamma = 1$ it indicates that one-sided error component dominates the symmetric error component.

The main hypothesis of interest of the SFA is:

$$\mathrm{H_0}: \beta_1 = \ ... = \beta_q = 0 \quad q < K \ .$$

The omission of u_i is equivalent to imposing the restriction specified in the null hypotheses i.e.

$$\mathrm{H_0}: \gamma = \delta_0 = \cdots = \delta_J = 0.$$

This indicates that the inefficiency effects in the frontier model are not present (no efficiency). Null hypotheses of interest are tested using the generalized likelihood ratio. The null hypothesis is $\mathrm{H_0}: \gamma = 0$ which specifies that technical inefficiency effects are not stochastic.

We reject the null hypothesis of no technical inefficiency effects given the specifications of the stochastic frontier and inefficiency effect model. If the parameter $\gamma = 0$ we accept null hypothesis then the variance of the technical inefficiency effect is zero and so the model reduces to the traditional mean response function. Leaving a specification with parameters that can be consistently estimated using ordinary least squares.

The *second step* of SFA involves the specification of a regression model for predicted technical inefficiency effects. OLS is inappropriate and either the dependent variable must be transformed prior to estimation or a limited dependent variable estimation technique must be employed.

3. Data Envelopment Analysis

The parametric method involves the application of econometric techniques where efficiency is measured relative to a statistically estimated frontier production function. The non-parametric method revolves around mathematical programming techniques, the most commonly applied of which is generically referred to as DEA. In this case, the former body of method imposes a particular functional form, while the latter does not. Therefore, another linear programming method for assessing the efficiency and productivity units is the Data Envelopment Analysis. In particular, DEA is a non-parametric linear programming method for assessing the efficiency and productivity units called decision-making units (DMUs) because they enjoy a certain decision-making autonomy. DEA application areas have grown since it was first introduced as a managerial and performance measurement tool in the late 1970s. The DEA approach was introduced by Charnes et al. (1978) who proposed the efficiency measurement of the DMUs for constant returns to scale (CRS), where all DMUs are operating at their optimal scale. Later Banker et al. (1984) introduced the variable returns to scale (VRS) efficiency measurement model, allowing the breakdown of efficiency into technical and scale efficiencies in DEA.

Over the last few decades, data envelopment analysis has gained considerable attention as a managerial tool for measuring the performance of organizations, and it has been used widely for assessing the efficiency of public and private sectors. This method leads to the System Selection of the optimal weights for the generic DMUs, and to the solution of a mathematical programming model in which the decision variables are represented by the weights associated with each input and output unit. DEA allows multiple inputs–outputs to be considered at the same time without any assumption on data distribution. In each case, the efficiency is measured in terms of a proportional change in inputs or outputs. A DEA model can be subdivided into an input oriented model, which minimizes inputs while satisfying at least the given output levels, and an output-oriented model, which maximizes outputs without requiring any more observed input values. The most well-known is represented by the *input oriented* CRS efficiency (Charnes, et. al., 1978), where the formulation of the linear optimization problem, for the i-th DMU, is :

$$max_{f,l}\ f, \qquad \text{subject to} \begin{cases} -fy_i + Y\lambda \geq 0 \\ -fy_i + Y\lambda \geq 0, \\ x_i - X\lambda \geq 0, \\ \lambda \geq 0, \end{cases} \tag{3}$$

where *X* is a matrix of *kxn* input and Y is a matrix of *mxn* output, with *n* equal to the number of DMU, y_i and x_i are the outputs and inputs observed for the i-th DMU, *f* is a scalar $(1 \leq f \leq +\infty)$ and λ is a constant vector of *nx1*. The score of technical efficiency for the DMU is represented by the quantity *1 / f* , and varies therefore, between *0* and *1* (*f = 1* denotes a DMU that stands on the frontier of production and is therefore technically efficient).

Another goal of the input-oriented DEA model is to minimize the virtual input, relative to a given virtual output, subject to the constraint that no DMU can operate beyond the production possibility set and the constraint relating to non-negative weights. In practice, most of the available DEA programs use the dual forms as expressed in (4), which lower the calculation burden and are virtually the same as (3):

$$min_{\theta\lambda,}\ \theta, \tag{4}$$

where λ is a semipositive vector in R^k and θ is a real variable.

In this paper we use the *input-oriented CRS* model to compare the results of the SFA; however, other variations are easily extendable and available in most DEA literature, including Coelli et al. (2005) and Cooper et al. (2006).

4. SFA with same inputs and different outputs

The Stochastic Frontier Analysis permits evaluating the technical efficiency scores for the input variables (x_1, x_2, ..., x_k) with output y_1 and to obtain a measure of the Technical Efficiency (TE_1) that can be called *TE* (y_1) i.e. a technical efficiency that is a function of y_1. We suggest performing multiple SFA with the same group of input variables (x_1, x_2, ..., x_k) but with different output variables (y_j) (*j=2, …,k*). For each *i-th* SFA we have the corresponding TE(y_i) with continuous values in [0,1]. Each indicator of efficiency *TE*(y_i)

obtained by each SFA, can be transformed into values on an ordinal scale. You obtain k rankings each due to a specific input variable used (y_j). It becomes, therefore, a problem of ordering multivariate data of an ordinal type. In a lot of applications we are interested in a unified ranking of the DMU rather than in values of the single Technical Efficiency.

In order to obtain a single graduation, you can use a Principal Component Analysis in considering the *TE* (y_j) ($j=1,2, \ldots,k$) as variables. You may grade the DMU according to the score on the first axis, but you obtain a ranking that is dependent on the first eigenvalue. The scores on the first principal component furnish an approximate indication of the probable ranking of the DMU. However, because the first principal component maximizes the weighted sum of squares of the correlation coefficients between the original variables and the first principal component, we will use this ranking that permits obtaining a unified ranking.

We use the data gathered from an official survey performed by the school management of the Campania Region *(Cometa project)*. The schools surveyed by the Regional School District will be at the end of the investigation, being more than a thousand. In this work were examined only thirty-three schools that had given coherent and validate data. The survey covers attributes regarding: environment, territorial context and economic resources. We started with the model including all variables and interactions. The choice of the model is based on the Box-Cox transformation (Box and Cox, 1964), while the choice of the functional form has been carried out under the hypothesis of a parsimonious model by likelihood ratio test and AIC criteria (Akaike,1977). After significance tests, only certain variables have been kept on the list of the potential determinants of technical efficiency, that represent characteristics of the school and of the management/production. We started with the model including all variables and interactions. The choice of the functional form has been carried out under the hypothesis of a parsimonious model. The null hypothesis of absence of random technical inefficiency is rejected in the different specifications and thus the Stochastic Frontier Analysis seems appropriate for the data. After verifying the hypothesis of asymmetry present in the residuals of the OLS and after trying several models with different dependent variables, the first model of SFA (SFA1) is:

$$ln(y_{1i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + v_i - u_i \qquad (5),$$

where i refers to the i-th school, y_{i1} is the *number of students who have passed the average score in the national test respect to the number of students,* x_{i1} is the rate of *number of teachers who have worked for more than ten years,* x_{i2} is the *rate of use of laboratories with respect to the availability,* x_{i3} is the *rate of use gyms and sports equipment*, x_{i4} *rate of implementation of projects*. Variables v_i and u_i are defined as described in Section 2.1. In Table 1 are summarized the main results of model (5), based on data of 18 schools. The second (6) and the third model (7), SFA2 and SFA3 respectively, differ from (5) only for the output variable (y_{2i}, y_{3i}) :

$$ln(y_{2i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + v_i - u_i \qquad (6),$$

$$ln(y_{3i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + v_i - u_i \qquad (7),$$

where, in (6) y_{2i} refers to the *number of students who passed the secondary school-leaving examination with a score greater than 80/100* compared to the total number of examined students, while, in (7), y_{3i} represents the *number of regular students in the study with respect to the starting lever*.

The results (Table 1) of model (5) show that the production inputs as the rate of use of laboratories with respect to the availability and the rate of use of gyms and sports equipment has a significant impact on the determination of the production frontier. Although positive, the presence is not significant of teachers with more than ten years' teaching experience as well as the rate of realization of projects. For reasons of space we do not report further comments on the results of model (6) and other models because we are interested mainly in the graduation of technical efficiency with respect to the stochastic frontier.

Table 1. Estimation results of Frontier Production with dependent variable being the number of students who have passed the average score in the national test with respect to the number of students

Input variables/Parameters	Coefficient	Standard Error	z	P>|z|	95% confidence interval	
Constant	2.2993660	.4785349	4.81	0.000	1.361455	3.2372770
x_{i1}	.0025374	.0026548	0.96	0.339	-.0026660	.0077408
x_{i2}	-.0084085	.0016368	-5.14	0.000	-.0116165	-.0052004
x_{i3}	-.0066347	.0037900	-1.75	0.080	-.0140631	.0007936
x_{i4}	.0068563	.0040808	1.68	0.093	-.0011420	.0148546
σ_u	.0930487	.0540731				
σ_v	.3810259	.0921572			γ =0.94	
Log likelihood = .760369 Prob > χ^2 = 0.0000						
Likelihood-ratio test of σ_u= 0 :	$\chi^2(01)$ = 2.72				Prob ≥ χ^2 =0.049	

Indeed, by means of the respective models (5), (6) and (7) were computed the Technical Efficiencies (Table 3) of individual schools (DMU) suitably codified.

We assume that the three Technical Efficiencies have been collected in a data matrix ***X***, in which the rows are associated with the DMU and the columns with the three Technical Efficiencies as variables. The principal components of the three Technical Efficiencies are obtained from the PCA on ***X***. We can see (Table 2) that about 47% of the total variation is explained by the first principal component indicating that there is some conflict among the individual rankings.

The first Principal Component is expected to approximate to the common ranking quite well, therefore the scores, transformed into rank (Table 3) could be used for comparison with the results from a Data Envelopment Analysis on the same data. Thus, by considering the Pearson's correlation coefficients of ***X*** (Table 4), we note that a low positive correlation exists (0.2717, 0.2500, 0.0548) among the three Technical Efficiencies. That concordance of sign of correlation, even if low, will ensure the success of the methodology.
Conversely, there is a very high correlation among the three technical efficiencies and the first principal component which reinforces the quality of the graduation carried out by the first component. Finally, the high Kendall's rank-correlation coefficient (0.8265) between the two rankings, Ist Principal Component and DEA, confirms the validity of the method shown.

Table 2. The results of Principal Component Analysis on the Technical Efficiencies

	Eigenvectors		
Variable	1st PC	2nd PC	3rd PC
TE (y_1)	0.6498	0.1374	0.7476
TE (y_2)	0.4785	-0.8381	-0.2619
TE (y_3)	0.5906	0.5279	-0.6104
Eigenvalues	1.5053	0.8722	0.6225

Table 3. Scores of Technical Efficiencies on the first Principal Component and rankings by DEA

SCHOOL CODE	$TE(y_1)$	$TE(y_2)$	$TE(y_3)$	I.st PC	Rank by I.st PC	Rank by DEA
S2	1.000	1.000	1.000	1.879973	1	1
S7	1.000	1.000	1.000	1.879973	1	1
S16	1.000	0.682	1.000	1.387817	3	1
S13	0.845	0.909	0.916	1.019996	4	7
S14	1.000	0.368	0.971	0.804637	5	5
S17	0.908	0.462	0.960	0.653533	6	6
S9	0.815	1.000	0.683	0.295076	7	8
S11	0.825	1.000	0.642	0.185863	8	8
S12	0.680	0.669	0.906	0.149252	9	11
S6	0.390	1.000	0.980	0.090945	10	16
S10	0.792	0.086	0.958	-0.262555	11	9
S15	0.673	0.449	0.842	-0.425536	12	12
S3	0.401	1.000	0.762	-0.608788	13	16
S8	0.854	0.504	0.505	-0.959166	14	8
S5	0.670	0.404	0.673	-1.070176	15	16
S4	0.596	0.152	0.841	-1.105909	16	16
S1	0.388	0.282	0.905	-1.277339	17	17
S18	0.232	0.097	0.716	-2.637603	18	18

Table 4. Pearson's correlation coefficients (0.05 significance level with a star)

Variable	$TE(y_1)$	$TE(y_2)$	$TE(y_3)$	Ist Principal Component
$TE(y_1)$	1.0000			
$TE(y_2)$	0.2717	1.0000		
$TE(y_3)$	0.2500	0.0548	1.0000	
Ist Principal Component	0.8045*	0.6315 *	0.5930*	1.0000

5. Discussion and Conclusions

The Data Envelopment Analysis (DEA) is a non-parametric deterministic approach that uses the mathematical programming to identify the efficient frontier, and does not impose functional forms. The main advantage of DEA is that it does not require an a priori hypothesis about the analytical form of the production function. Indeed, DEA determines the production function by applying minimization techniques on the data. Differently from regression analysis, the DEA is based on extreme observations, and this leads to the main disadvantage of DEA, i.e., that the frontier is sensitive to the extreme observations. Furthermore, DEA postulates the absence of random errors and that all deviations from the frontier denote inefficiency of the DMUs.

Vice versa, the SFA, is a parametric approach that hypothesizes a functional form and uses the data to econometrically estimate the parameters of this function. The SFA requires functional forms on the production frontier, and assumes that units may deviate from the production frontier not only due to technical inefficiency but also to measurement errors, statistical noise or to other non-systematic factors. In addition, the SFA requires strong distribution assumptions of both statistical random errors (i.e., normal distribution) and the non-negative technical inefficiency random variables (i.e., half-normal or truncated normal distribution) (Coelli et. al., 2005).

With SFA the determinants of efficiency are directly obtained by estimating the production function. With SFA you can use various models changing the response variable every time and can eventually identify the model which has greater relevance in terms of acceptance.

The method described in this work is suitable for the evaluation of efficiency. Moreover, even our partial data, the method and the results achieved already provide a useful interpretation of the efficiency frontier for the evaluation of schools. Indeed, the efficiency estimates obtained have been utilized to rank the schools according to the common efficiency index.

The comparison with the results obtained through the Stochastic Frontier Analysis with same input and different outputs correlates very well (0.8265) with the results of the DEA. This result confirms the quality of the alternative method proposed in this paper. The rankings obtained by the Stochastic Frontier, however, are more robust than those of the DEA for the very closely tested hypothesis.

Acknowledgements

The present work has been partially funded under an agreement with Nidge University (Resp. Prof. Luigi D'Ambra). We thank the School Superintendent of Campania, and in particular Angela Orabona of the "Polo Qualità di Napoli" for providing us with the data for this study.

Bibliography

1. Abdi, H. and Lynne, J. W. **Principal component Analysis, vol. 2,** John Wiley & Sons, 2010
2. Alden, H. W. **Genetic Algorithms for Real Parameter Optimization,** Foundations of Genetic Algorithms, Edited by Gregory J. E. Rawlins, Morgan Kaufman, 1991
3. Akaike, H. **Factor analysis and AIC,** Psychometrika, Vol. 52,no.3, 1987, pp. 317-332
4. Banker, R. D., Charnes, A. and Cooper, W. W. **Some models for estimating technical and scale inefficiencies in data envelopment analysis,** Management Science, Vol. 30, 1984, pp. 1078–1092
5. Battese G. E. and Corra G. S. **Estimation of a production frontier model: with application to the pastoral zone of eastern Australia,** Australian Journal of Agricultural and Resource Economics, Vol. 21, 1977, pp. 169–179
6. Box G. E. and Cox D. R. **An analysis of transformations,** Journal of the Royal Statistical Society, Series B (Methodological), 21, 1964
7. Charnes, A., Cooper, W. W. and Rhodes, E. **Measuring the efficiency of decision making units,** European Journal of Operational Research, Vol. 2, 1978, pp. 429–444
8. Coelli, T. J., Rao, D. S. P., O'Donnell, C. J. and Battese, G. E **An introduction to efficiency and productivity analysis,** Springer, 2005
9. Cooper, W. W., Seiford, L. M. and Tone, K. **Data envelopment analysis: a comprehensive text with models applications references and DEA-solver software,** Springer, 2006
10. Holland, J. **Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence,** Bradford Books, 1992
11. JGAP, **Java Genetic Algorithms Package,** http://jgap.sourceforge.net/

12. Korhonen , P. and Siljamäki, A. **Ordinal principal component analysis theory and an application,** Computational Statistics & Data Analysis, Vol. 26, No. 4, 1998, pp. 411-424
13. Kumbhakar, S.C. and Lovell, C.K. **Stochastic frontier analysis,** Cambridge University Press, 2003
14. Mizala, A., Romaguera, P. and Farren, D. **The technical efficiency of schools in chile,** Applied Economics, vol. 34, 2002, pp. 1533–1552
15. Rao, D. P., O'Donnell, C. J., Battese, G. E. and Coelli, T. J. **An introduction to efficiency and productivity analysis,** Springer, 2005
16. Ray, S. C. **Data envelopment analysis: theory and techniques for economics and operations research,** Cambridge University Press, 2004

CHOOSING APPROPRIATE METHODS FOR MISSING DATA IN MEDICAL RESEARCH: A DECISION ALGORITHM ON METHODS FOR MISSING DATA

Hatice UENAL
Institute of Epidemiology and Medical Biometry, Ulm University , Germany

Benjamin MAYER
Institute of Epidemiology and Medical Biometry, Ulm University , Germany

E-mail: benjamin.mayer@uni-ulm.de

Jean-Baptist DU PREL
Institute of History, Theory and Ethics of Medicine, Ulm University, Germany

**HU and BM contributed equally.*

Abstract

Missing data (MD) are a common problem in medical research. When ignored or treated not appropriately, MD can lead to seriously biased results. Currently, there are no comprehensive guidelines for efficiently identifying suitable imputation methods in different MD situations. The objective of the paper is to discuss various methods to handle missing data. Based on a selective literature search, common MD imputation methods were identified. A decision algorithm is presented where the considered methods are prioritized with respect to the underlying missing data mechanism and scale level of the incomplete data. Furthermore, all included imputation methods are described in more detail. No alternative decision algorithms for MD imputation methods of this complexity have been developed yet, thus it could serve as a useful tool for researchers confronted with MD.

Keywords: *Missing data, decision algorithm, imputation methods, multiple imputation*

Introduction

Missing values are a common problem in medical research and may be a considerable source of bias when ignored or not handled appropriately (Schafer & Graham, 2002). Most studies come along with some type of missing data (questionnaire data, laboratory values, missing records, subject loss to follow-up, etc.). There is often no information in publications which measures were taken into account for missing data (MD) and what impact these choices may had on the results (Eekhout et al., 2012). Furthermore, there are few guidelines in epidemiology as well as in clinical research (Little et al., 2012) which explicitly focus on the MD problem.

The capabilities of imputation might be unknown or inappropriately used for minimizing effects (reduction of power, introduction of bias) of MD on study results. A recently published systematic review on studies in three leading epidemiological journals showed that complete case analysis (CCA) (81%) and single imputation methods (15%) were most frequently used to cope with MD (Eekhout et al., 2012). Both approaches are easily applicable but often not appropriate to deal with MD, as shown in the following. In many cases the adverse effect of MD on a study's validity might not be fully aware and those who are aware may not know that validity could be improved with imputation. Finally, those who wish to use imputation may not know which method is suitable in a particular situation.

Currently, there is no detailed guidance which provides researchers from various disciplines analyzing medical data with information on choosing a proper MD technique for their analysis. In order to address this gap, a comprehensive decision-making algorithm for the application of established MD imputation methods is suggested. Statistical imputation is a powerful tool to handle with MD, so researchers have to be encouraged to deal with the available methods in order to get more reliable study results. The presented algorithm is able to facilitate this process.

Methods

This article is based on the assumption that missing values are only present in a single variable (in the following called "target variable"), which can be any variable of the dataset including the outcome variable. Classifying the cause of missing values is of fundamental importance for determining how they should be handled. Therefore, the developed decision algorithm included a distinction between established missing data mechanisms (Table 1) which were initially described by Rubin (1976). A Missing Completely At Random (MCAR) mechanism is present if the missing values of a data set are a random sub-sample of the complete data set, i.e. the probability of MD is independent of all other variables in the data set including the target variable. An example of MCAR may be a patient who dies in a traffic accident during the course of a clinical trial. Assuming a Missing At Random (MAR) mechanism, the probability of MD depends on other variables of the data set, but not on the values of the target variable. MAR is present, for example, if gender predicts the probability of response on a depression score. In case of Not Missing At Random (NMAR), the probability of MD depends on the observed as well as unobserved variables. A NMAR is present if a subject with manifest depression will probably not report about his mental constitution since he fears the consequences of doing so (e.g. inpatient treatment).

Table 1. Overview of missing data mechanisms

MCAR[a]	**MAR**[b]	**NMAR**[c]
Probability of MD is unrelated to covariates and the values of the target variable	Probability of MD could be related to covariates, but not to values of the target variable	Probability of MD relates to unobserved values of the target variable even after control for covariates

[a] Missing Completely At Random [b] Missing At Random [c] Not Missing At Random; MD=missing data

It is usually not easy to distinguish between these three concepts, so a sound knowledge on substantial coherences in the data set is necessary. The mechanisms do not always provide a logical-causal explanation for the absence of data, nevertheless they offer a mathematical approach to model the probability of MD in association with other variables in the dataset.

In longitudinal studies with MD in a single variable (or studies with MD in multiple variables in general), it is also necessary to look for patterns generated by MD in the data set. Monotone patterns have to be distinguished from arbitrary patterns (Figure 1), since the shape of the pattern affects the applicability of particular imputation approaches (Schafer & Graham, 2002). Hence, the decision algorithm also considers the missing data pattern as a relevant criterion for choosing appropriate imputation methods.

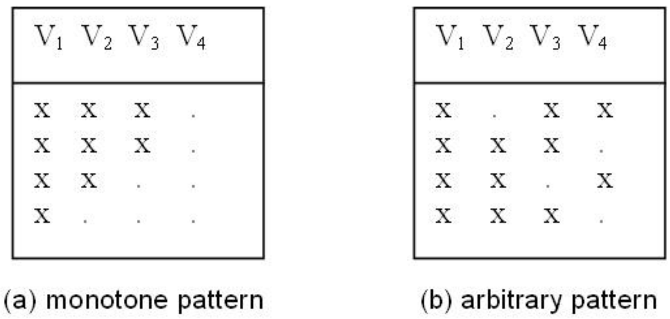


Figure 1. Missing data patterns

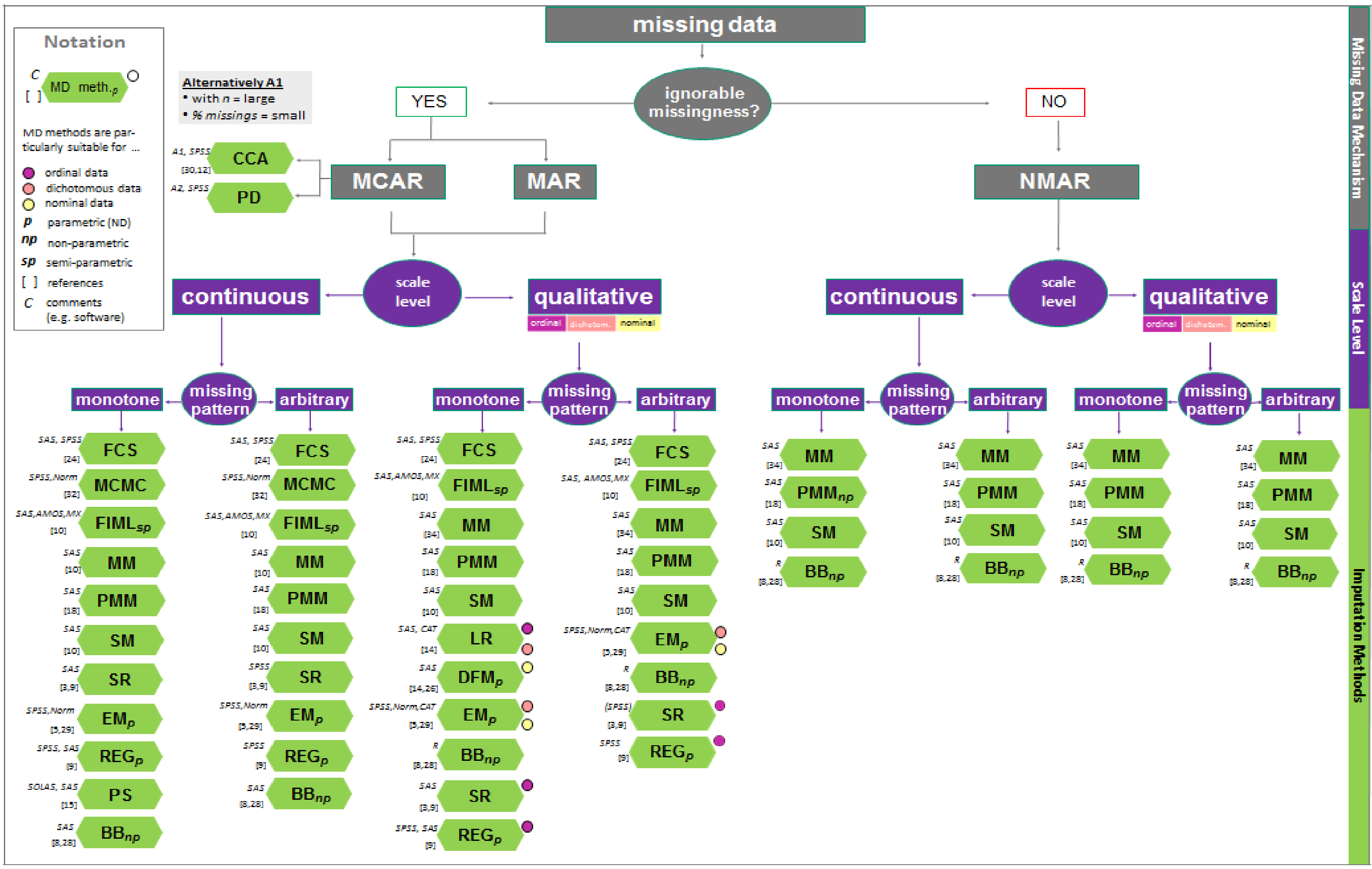
Along with the missing data mechanism and pattern the third important aspect to be considered within a decision algorithm on MD methods is the scale level of the variable to be imputed. There are methods for categorical as well as continuous variables which have to be chosen properly.

Common parametric as well as non-parametric imputation procedures were identified by selective literature search. Based on the available information in the literature, criteria for choosing adequate imputation methods were discussed and prioritized. They formed the base for the development of the proposed decision algorithm. The prioritization of methods rely on the findings of referenced studies which compared different imputation methods.

Results

The above described criteria for choosing appropriate methods to handle with MD are considered within the proposed decision algorithm (Figure 2). This guidance can efficiently assist researchers in finding appropriate imputation methods. To illustrate the use of the proposed algorithm, the following hypothetical MD situations are exemplarily considered: (i) Following a one-time survey of smoking behaviour, the data did not reveal a clear pattern of MD in the variable "current use (few, a packet or more than a packet of cigarettes per day)" and (ii) it was previously known that in a longitudinal psychiatric study patients with diagnosed depression are more likely to have MD due to a reduced willingness to report about their current mental condition (score value out of 5 items).

Considering these scenarios, an initial examination whether the MD have occurred randomly or systematically is required. Given the definitions in Table 1, the mechanisms MCAR and MAR indicate situations where the probability of MD are assumed to be non-systematic. Therefore, at least a MAR mechanism can be assumed for situation (i), since the probability of MD in the variable "current use" seems to be independent of the covariate structure of the respective patients. In contrast, a NMAR mechanism is most likely for situation (ii) since patients with previously diagnosed depression are more likely to deny reporting about their actual mental condition. Moreover, the target variable in (i) is "current use" and according to the description above an ordinal scaled qualitative outcome. The target variable "mental condition" in (ii) is a score value derived from five items, thus it can be considered as a continuous variable. Further assumptions are that in situation (i) there is an arbitrary MD pattern and in situation (ii) it is possible to induce a monotone MD pattern.



BB = Bayesian Bootstrap, CCA = Complete case analysis DFM = Discriminant Function Method, EM = Expectation Maximization, FCS = Fully conditional specification, FIML = Full Information Maximum Likelihood, LR = Logistic Regression, MCMC = Markov Chain Monte Carlo, MM = Mixed Model, PD = Pairwise Deletion, PS = Propensity Score, SR = Stochastic Regression, REG = Regression, PMM = Pattern Mixture Model, SM = Selection Model, A1 = with large n and small number of missings; A2 = like A1, but to the contrary, AMOS (licensed for use by SPSS, Inc.), CAT (licensed for use by S-Plus, TIBCO Software, Inc), MX[36], NORM[33], R (www.r-project.org), SAS (SAS Institute Inc.), SPSS (SPSS Inc.), SOLAS (Statistical Solutions Ltd.)

Figure 2. Decision algorithm on missing data methods

Applying the presented MD decision algorithm (Figure 2), a full information maximum likelihood (FIML) procedure might be adequate to impute ordinal categorical data in (i) (Enders & Bandalos, 2010). In situation (ii) a mixed model or pattern mixture model to handle the missing score values may be appropriate (Enders & Bandalos, 2010, Verbeke & Molenberghs, 2000).

Imputation strategies and methods

In the following, a summary of the methods presented in Figure 2 is provided. An in-depth description of all methods is out of the scope of this article, but interested readers may use the cited references given in the algorithm for each suggested method to obtain more information.

Complete case analysis (CCA) discards subjects with incomplete data in any variable from analysis (listwise deletion) (Enders, 2010). This may be justifiable only in the MCAR situation when complete cases are a random sample of all cases (Vach & Blettner, 1991), but generally problematic in MAR or NMAR situations. The loss of statistical power increases as the number of MD is high. Cases with MD in at least one variable are common, especially in epidemiological studies, where it is typical to have a large number of variables. Hence, even an overall small amount of MD can lead to a dramatic reduction of evaluable cases (Figure 3).

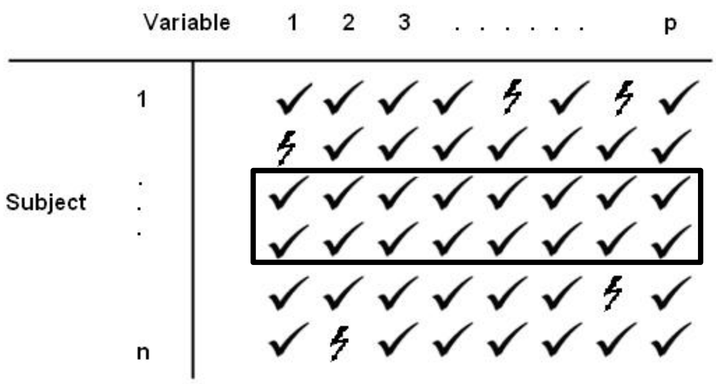


Figure 3. Loss of information with a complete case analysis

Available case analysis (pairwise deletion) is another method which may be applied in the MCAR situation if the number of MD is low. It overcomes the problem of high loss of cases and power due to MD in contrast to CCA, but has other disadvantages (Enders, 2010). This method utilizes all available data of the subsample used for the actual analysis which contains all observation units with no MD in the variables included. Therefore it is possible that every analysis including different variables has a different number of cases. This difference becomes especially obvious when calculating for example correlation coefficients on the base of subsamples with different numbers in a covariance matrix. Then correlation coefficients of over 1 could result (Enders, 2010).

Single Imputation (SI) methods are characterized by replacing each missing value of a dataset by just one imputation value. There are numerous SI methods and all of them have the tendency to underestimate variance, because the imputation itself is not explicitly considered within the imputation as an additional source of variance. Most of these approaches are not recommended and are therefore not included in the decision algorithm. The most common SI-methods are mean or median imputation and so called “hot and cold” deck imputation methods. Regression based imputation uses information of the complete cases to estimate MD in incomplete variables. The regression approach shares this “bond” of available data for MD estimation with maximum likelihood methods, though the latter use more complicated algorithms (Enders, 2010).

A more sophisticated technique is the stochastic regression imputation which overcomes the disadvantage of the simple regression imputation of predictable bias due to a lack of variability (Baraldi & Enders, 2010). Similar to the simple regression approach, this method imputes the MD of a variable according to their relationship to other covariates, but additionally incorporating a normally distributed residual to account for the reduced variability of the imputed variables. Simulations have shown that stochastic regression provides reasonable results similar to those of Multiple Imputation methods or maximum likelihood methods (Enders, 2010).

Maximum likelihood estimation is another approach to handle MD (Schafer & Graham, 2002). The Expectation Maximisation algorithm (EM) iterates maximum likelihood estimations (repeated, stepwise approximations) of a set of regression equations for predicting MD to get the best solution (Dempster, Laird, & Rubin, 1977). Another modern maximum likelihood based imputation is the full information maximum likelihood method (FIML) which is based on structure equation models and takes all available information into account (Enders & Bandalos, 2001).

The Multiple Imputation (MI) strategy incorporates several state of the art techniques. A MI analysis includes three distinct phases, the imputation phase (1), the analysis phase (2), and the pooling phase (3) (Figure 4). Initially, missing values are imputed *m* times (*m* > 1), resulting in *m* completed data sets (Rubin, 1987). Each of these datasets is analyzed separately afterwards according to the statistical model chosen to answer the research question. The *m* single results (e.g. an estimated regression coefficient) are finally combined to one MI estimator (Little & Rubin, 2002).

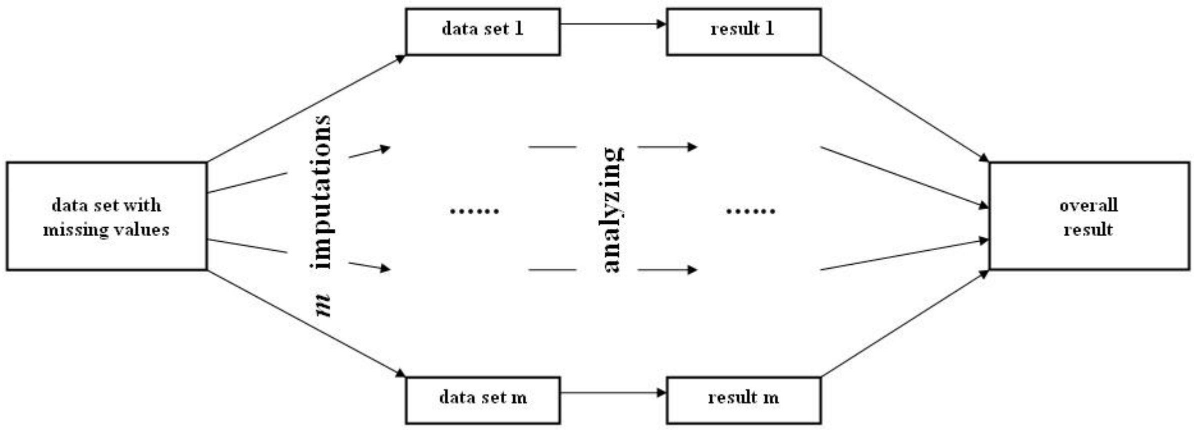


Figure 4. The principle of Multiple Imputation

From a statistical point of view, MI is able to approximate the observed data likelihood by the average of the completed data likelihood over unknown missing values (He, 2010). MI produces more reasonable estimates than SI procedures (Tanner & Wong, 1987) because the imputation process is considered as an additional source of variation.

The Markov Chain Monte Carlo (MCMC) method uses available information from the observed data to calculate a corresponding a-posteriori distribution with the help of different algorithms, e.g. data augmentation (Tanner & Wong, 1987). Here, likelihood-based sampling for the imputation of MD is performed. It is a quite robust MI procedure for quantitative variables which can be applied to arbitrary MD patterns and produces valid values even for high proportions of MD in the case of MCAR or MAR.

An alternative algorithm for more complex data structures, especially when multiple variables have MD, is known as fully conditional specification (FCS) or multiple imputation by chained equations (MICE). It enables the imputation of quantitative as well as qualitative variables (Raghunathan et al., 2001). Imputation proceeds from one variable with the least to that with the most MD. At each step random draws are made from both the posterior distribution of the parameters and the posterior distribution of the missing values. Imputed values at one step are used as predictors in the imputation equitation at subsequent steps. Once all MD have been imputed, several iterations of the process are repeated before selecting a completed data set.

The Propensity Score method is conceptualized for a MI of continuous variables and demands a monotonous MD pattern (Lavori, Dawson, & Shera, 1995). In the context of MD, the propensity score represents a conditional probability that a subject has a missing value given a structure of covariates (Lavori et al., 1995). A propensity score is calculated based on a logistic regression model for each subject and all subjects of a data set are then sorted in an ascending order and distributed to a number of subsets according to their corresponding propensity score. The MD is imputed by a random draw of observed values of a certain subset.

The logistic regression procedure is an appropriate MI approach for binary or ordinal variables with MD in the MCAR or MAR situation (Hohl, 2007; Sentas, Angelis, & Stamelos, 2013; Rubin, 1987). A monotonous MD pattern is required. First, a regression model is set up with all complete cases to estimate the regression coefficients. For given covariates the probability p is calculated that one of two possible values 0 or 1 is used for imputation of a missing value. Comparing p with a uniformly distributed random variable u, a decision is made if the imputation value is 0 or 1.

The discriminant function method imputes MD of a nominal variables and assumes MAR as well as a monotone MD Pattern (Hohl, 2007). Furthermore, the covariates in the model of the analysis have to be normally distributed (Rubin, 1987). Generation of linear discriminant functions are based on the nominal scaled dependent variable with categories 1,..., g which are affected by missing values, continuous covariates of the dataset and the a-priori likelihoods for the categories 1,...,g. The a-posteriori likelihoods $p_1,\ldots,p_g$ with $p_1 + \ldots + p_g = 1$ are then compared to a uniformly distributed random variable u on the interval 0 to 1. For $u < p_1$ the missing value is imputed by category 1, for $p_1 < u < p_1+p_2$ by category 2 and so on. Only continuous variables can be applied to the generation of the discriminant functions.

Bootstrap methods have proved to work well for the estimation of missing values (Cohen, 1997). In small or nonparametric samples it is recommended to use respective variants of this method (Efron 2011; Pajevic & Basser, 2003; Rubin & Schenker, 1991). The bootstrap is a useful method to estimate a valid variance for complex survey situations (Efron 2011; Shao & Sitter, 1996), in particular non-ignorable non-response (Cohen, 1997; Efron 2011; Rubin & Schenker, 1991). In the field of MD, bootstrap is based on the observed data. By resampling from the data set, random bootstrap samples are repeatedly induced (Efron & Tibshirani, 1993). Initially, the whole sample is stratified in complete and missing data and then the variability of the MD is estimated. In this context, the error variance allows a reliable approximation of the "correct bootstrap distribution". Then a MI follows (Hinkley & Davison, 1997). In non-ignorable MD situations bootstrap methods have the advantage to guard against misspecification of the imputation model with minimal assumptions about the distribution of the data (Sapra, 2012).

In contrast to the already discussed imputation approaches, modelling approaches estimate parameters of interest without explicitly imputing MD. Most established methods are based on mixed linear models and maximum likelihood methods (Allison 2011; Efron 2011; Verbeke & Molenberghs, 2000). These advanced methods are especially indicated in the NMAR situation where the MD mechanism has to be taken into account at the estimation process (Verbeke & Molenberghs, 2000). There are different ways to partition the common distribution (Little, 1993), leading either to so called pattern mixture or selection models (Verbeke & Molenberghs, 2000). Accounting accurately for the right mechanism can often be difficult, therefore modelling approaches should be implemented with caution. They can indeed be applied when a NMAR situation is assumed, although the results remain uncertain to some extent. Therefore, it is recommended to conduct a comprehensive sensitivity analysis by applying different approaches when missing values are supposed to be NMAR (European Medicines Agency, 2010; Verbeke & Molenberghs, 2000).

Overall, the literature search suggested a superiority of MI and maximum likelihood methods over regression based approaches in case of continuous target variables (Dempster et al., 1977; Eekhout et al, 2012; Mayer, 2011; Newman, 2003). According to Allison (2009), the two particular methods of discriminant function and logistic regression imputation are proposed for categorical target variables instead of just disregard incomplete cases. However, these methods assume a monotone pattern of the data set. No direct comparisons between imputation methods and modelling approaches could be found which address the problem of MD by considering them explicitly in a statistical model. Therefore, especially in the case of NMAR, the given prioritization of modelling approaches in Figure 2 is based on the authors' subjective understanding of the methods' properties. The listed methods for the imputation of missing values have been summarized in Table 2.

Table 2. Imputation methods and their presumptions

Methods	**Missing Data Mechanism**			**Scale level**		**Pattern**	
	MCAR	**MAR**	**NMAR**	**continuous**	**qualitative**	**arbitrary**	**monotone**
MCMC	X	X		X		X	X
FCS	X	X		X	X	X	X
FIML	X	X	X	X	X	X	X
SR	X	X		X	X		X
MM	X	X	X	X	X	X	X
EM	X	X		X	X	X	X
REG	X	X		X			X
PS	X	X		X			X
BB	X	X	X	X	X	X	X
LR	X	X			X		X
DFM	X	X			X		X
PMM	X	X	X	X	X	X	X
SM	X	X	X	X	X	X	X

BB = Bayesian Bootstrap, **DFM** = Discriminant Function Method, **EM** = Estimation Maximization, **FCS** = Fully Conditional Specification, **FIML** = Full Information Maximum Likelihood, **LR** = Logistic Regression, **MAR** = Missing At Random, **MCAR** = Missing Completely At Random, **NMAR** = Not Missing At Random, **MCMC** = Markov Chain Monte Carlo, **MM** = Mixed Model, **PMM** = Pattern Mixture Model; **PS** = Propensity Score, **REG** = Regression, **SM** = Selection Model, **SR** = Stochastic Regression

Discussion

Common MD situations and imputation strategies were reviewed. The most appropriate methods were chosen for the presented decision algorithm according to the cited literature for certain MD situations. Most SI methods were not considered since they are regarded as obsolete (Enders & Bandalos, 2001). An easily applicable algorithm (Figure 2) was suggested to assist researchers being confronted with the situation of MD in their data set in selecting an appropriate MD imputation method. Decision-making is based on MD mechanism and pattern as well as of scale level of the target variable to be imputed. NMAR is the most complicated MD mechanism to cope with. Allison (2012) described that maximum likelihood estimations were even more efficient than MI under the MAR assumption producing the same results for the same set of data. Prioritization of various available methods for specific MD situations helps to find the most suitable and, if possible, an appropriate statistical software to handle it (Mayer, Muche, & Hohl, 2012). To date this is the first decision algorithm for MD imputation methods of this complexity. Imputation of a large fraction of MD in a variable is always difficult. Then the question arises whether the chance of successful imputation is decreased. Complete case analysis then may lead to a systematic bias of study results by the loss of many observation units and the respective characteristics of other variables.

A number of articles have already proposed useful methods in several MD situations. However, finding a proper approach requires laborious and time consuming literature research. The decision algorithm was designed to efficiently support researchers in identifying appropriate imputation methods when confronted with MD. One challenge in using MD imputation procedures is that there is currently no single statistical software package in which all imputation methods are implemented (Yucel, 2011). Since not all scientists necessarily have access to commercial software packages, also non-commercial software alternatives are presented in the decision algorithm.

The stages of the algorithm provide orientation in handling various MD situations. The paths vary when the sample size is small or for different scale levels of the target variable. When applying the proposed concept, users have the possibility to impute MD in single-item instruments in several situations with various imputation methods.

Imputation methods are listed following the three stages MD mechanism, scale level and MD pattern. Methods were listed in order of their appropriateness as described in the cited literature. The suggested statistical software packages and references for each method are listed at the upper and lower left corners of each method block. Deciding between systematic missing (NMAR) and not (MCAR, MAR) is challenging and will surely demand certain understanding and consideration. However, upon choosing the initial paths, the user should obtain results efficiently via suggested statistical software packages.

Completeness of the presented methods cannot be guaranteed. The suggested prioritization is mainly based on findings in cited references, which compared different methods for imputing MD. Results of future studies comparing different imputation strategies may require a revision of this prioritization. A crude understanding of the methods described in the algorithm is advantageous, therefore a study of the cited references in the algorithm and to consult an expert is recommended. There is still no universal recommendation in relevant guidelines (European Medicines Agency, 2010) and the literature with respect to the percentage of MD up to which an imputation is suggested. Further analysis is needed to prove the applicability established methods for different amounts of MD. Furthermore, every method given in the algorithm has its limitations, for example the use of selection models in the NMAR situation: if the assumptions are satisfied, the selection model can virtually eliminate bias caused by NMAR data, but already modest correlations among the predictor variables and moderate deviation from normality assumption may produce severely biased estimates. Some limitations might will be handled by future developments.

Overall, our extensive literature search revealed that there are numerous approaches which address the problem of MD in medical research. However, it has obviously not been possible yet to arrange them reasonably to have a universal guidance for their application in specific MD situations. This points out again the importance of making all efforts to generally prevent MD and to find ways to systemize the available methods.

Acknowledgement

The authors thank Prof. Dr. Dietrich Rothenbacher and Prof. Dr. Rainer Muche for their helpful advice and critical comments on drafts of the manuscript. This research was supported by German Research Foundation Grant awarded to Hatice Uenal and by the German Federal Ministry of Research and Education Grant awarded to Jean-Baptist du Prel.

References

1. Allison, P.D. **Missing data**. in: Millsap, R.E. and Maydeu-Olivares, A. (eds.) "The SAGE Handbook of Quantitative Research in Psychology. 1st ed. London", Thousand Oaks, New Dehli, Singapore: Sage Publications, 2009, pp. 72-89
2. Allison, P.D. **Handling Missing Data by Maximum Likelihood**. SAS Global Forum, 2012
3. Baraldi, A.N. and Enders, C.K. **An introduction to modern missing data analyses**. Journal of School Psychology, Vol. 48, 2010, pp. 5-37
4. Cohen, M.P. **The Bayesian bootstrap and multiple imputation for unequal probability sample designs.** Proceedings of the Survey Research Methods Section, American Statistical Association, 1997, pp. 635-638
5. Dempster, A. P., Laird, N. M.,and Rubin, D. B. **Maximum Likelihood for Incomplete Data via the EM Algorithm.** Journal of the Royal Statistical Society, Vol. 39, 1977, pp. 1-38
6. Eekhout, I., de Boer, R. M., Twisk, J. W., de Vet, H. C. and Heymans, M. W. **Missing data: a systematic review of how they are reported and handled.** Epidemiology, Vol. 23, 2012, pp. 729-732
7. Efron, B. **The bootstrap and Markov chain Monte Carlo.** Journal of Biopharmaceutical Statistics, Vol. 21, 2011, pp. 1052-1062
8. Efron, B. and Tibshirani, R. J. **An Introduction to the Bootstrap. 1st ed**. New York, NY: Chapman and Hall, 1993
9. Enders, C.K. **Applied Missing Data Analysis (Methodology in the Social Sciences). 1st ed.** New York, London: Guilford Press, 2010
10. Enders, C.K. and Bandalos, D.L. **The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models.** Structural Equation Modelling: A Multidisciplinary Journal, Vol. 8, 2010, pp. 430-457
11. European Medicines Agency. **Guideline on Missing Data in Confirmatory Clinical Trials**. European Medicines Agency, London, 2010

12. He, Y. **Missing data analysis using multiple imputation: getting to the heart of the matter.** Circulation Cardiovascular Quality and Outcomes, Vol. 3, 2010, pp. 98-105
13. Hinkley, D. V. and Davison, A.C. **Bootstrap Methods and their Application. 1st ed**. Cambridge, England: Cambridge University Press, Publishers, 1997
14. Hohl, K. **Handling missing data: imputation methods for missing categorical values in clinical datasets.** PhD Thesis 2007, University of Ulm (in German); http://vts.uni-ulm.de/docs/2007/6027/vts_6027_8103.pdf. Accessed June 2014
15. Lavori, P. W., Dawson, R. and Shera, D. **A multiple imputation strategy for clinical trials with truncation of patient data.** Statistics in Medicine, Vol. 14, 1995, pp. 1913-1925
16. Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T. and Stern, H. **The prevention and treatment of missing data in clinical trials**. New England Journal of Medicine, Vol. 367, 2012, pp. 1355-1360
17. Little, R. J. A. **Regression with missing X's: a review.** Journal of the American Statistical Association, Vol. 87, 1992, pp. 1227-1237
18. Little, R. J. A. **Pattern mixture models for multivariate incomplete data.** . Journal of the American Statistical Association, Vol. 88, 1993, pp. 125-134
19. Little, R. J. A. and Rubin, D. B. **Statistical Analysis with Missing Data. 2nd ed.** New York, NY: John Wiley & Sons, 2002
20. Mayer, B. **Missing values in clinical longitudinal studies – handling of drop-outs**. PhD Thesis 2011, University of Ulm. (in German); http://vts.uni-ulm.de/docs/2011/7633/vts_7633_10939.pdf . Accessed June 2014
21. Mayer, B., Muche, R. and Hohl, K. **Software for the Handling and Imputation of Missing Data – An Overview.** Journal of Clinical Trials, Vol. 2, 2012, p. 103
22. Newman, D. A. **Longitudinal Modelling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques.** Organizational Research Methods, Vol. 6, 2003, pp. 328-362
23. Pajevic, S. and Basser, P. J. **Parametric and Non-Parametric Statistical Analysis of DT-MRI Data.** Journal of Magnetic Resonance, Vol. 161, 2003, pp. 1-14
24. Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V. and Solenberger, P. **A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of regression Models.** Survey Methodology, Vol. 27, 2001, pp. 85-95
25. Rubin, D. B. **Inference and missing data.** Biometrika, Vol. 63, 1976, pp. 581-592
26. Rubin, D. B. **Multiple Imputation for Nonresponse in Surveys. 1st ed.** New York, NY: John Wiley & Sons, 1987
27. Rubin, D.B. and Schenker, N. **Multiple Imputation in health-care databases: An overview and some applications.** Statistics in Medicine, Vol. 10, 1991, pp. 585-598
28. Sapra, S. **Robust bootstrap multiple imputation with missing data and outliers.** http://ubuntuone.com/75CiQjUOIgLXFG2utL9Whr. Accessed June 2014
29. Schafer, J. L. and Graham, J. W. **Missing data: Our view of the state of the art**. Psychological Methods, Vol. 7, 2007, pp. 147–177

30. Sentas, P., Angelis, L. and Stamelos, I. **Multiple Logistic Regression as Imputation Method Applied on Software Effort Prediction.** Department of Informatics, Aristotle University of Thesaloniki. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.196.1080. Accessed June 2014
31. Shao, J. and Sitter, R. R. **Bootstrap for Imputed Survey Data.** Journal of the American Statistical Association, Vol. 91, 1996, pp. 1278-1288
32. Tanner, M. A. and Wong, W. H. **The Calculation of Posterior Distributions by Data Augmentation.** Journal of the American Statistical Association, Vol. 82, 1987, pp. 528-540
33. Vach, W. and Blettner, M. **Biased Estimation of the Odds Ratio in Case-Control Studies due to the use of Ad Hoc Methods of Correcting for Missing Values for Confounding Variables.** American Journal of Epidemiology, Vol. 134, 1991, pp. 895-907
34. Verbeke, G. and Molenberghs, G. **Linear mixed models for longitudinal data.** New York, NY: Springer Press, 2000
35. Yucel, R. M. **State of the Multiple Imputation Software.** Journal of Statistical Software, Vol. 45, 2011, pp. 1-7

JOURNAL
OF
APPLIED
QUANTITATIVE
METHODS

WHAT HAPPENED WITH THE ATTRACTIVENESS OF THE ROMANIAN COUNTIES FOR FDI DURING THE PERIOD 2001 – 2012?

Vasile Alecsandru STRAT
PhD, University Lecturer,
Department of Statistics and Econometrics,
Bucharest University of Economic Studies, Romania

E-mail: vasilestrat@csie.ase.ro; strat_vasile@yahoo.com

Abstract
The FDI have become a very important aspect in the nowadays economical and geopolitical circumstances and, therefore, the study of this phenomenon is regarded with an increased attention by scholars, by government and business representatives.
Following this direction, the study of disparities registered between different regions or between different countries when dealing with the attractiveness of these entities in the eyes of foreign investors became a topic of an increasing importance.
In the present study, using yearly data regarding the stocks of FDI at the level of the Romanian counties, for the period 2001 – 2012, I try to evaluate the evolution of the attractiveness of these entities for foreign investors using the Gini coefficient. The study reveals that the attractiveness of the Romanian counties was significantly influenced by the main events which happened during this period.

Keywords: *FDI; county level; human capital; development region; Gini coefficient; king and viceroy effect*

1. Introduction

The foreign direct investments are regarded by the governments of many developing countries as one of the most important tools that can be used for their economical development.

The former communist countries from the eastern part of Europe are no exception and, starting from the beginning of the `90, their governments have been preoccupied with forging strategies that would enable them to attract foreign direct investments. These foreign direct investments were regarded as an important source of capital, an important source of management skills, an important source for new and better paid jobs, an important source for new technologies and also an important source for new and more completive products for both the internal market and for boosting up the export capacity of the country.

Although the competition between national governments for bringing foreign direct investments inside their countries is significantly more visible, there is also an important competition inside every state between the regional or even local authorities to attract these investments into their units. Another important aspect, mentioned by the literature on this matter, is the presence of important discrepancies between the attractiveness (in the eyes of the foreign investors) of these local entities in almost every state.

Of main importance regarding this subject is also the regional policy of the EU, developed with the sole purpose of reducing the development (economic point of view) disparities among the European regions. Therefore, it is obvious that reducing the disparities between regions (and other economical entities) should be regarded with an increased attention due to the fact that it is the main tool that can be used for ensuring similar and adequate standard of living for all the inhabitants.

Romania is no exception in this regard and the presence of disparities among counties needs to be analyzed and these discrepancies need to be diminished in order to ensure a sustainable development of the national economy. Also, forging policies for diminishing these discrepancies needs to be an important priority because the enhancement of this phenomenon can create significant macroeconomic and social disequilibria with significant negative impacts in several fields.

The structure of the paper contains four main sections, as follows: literature review and theoretical background, methodology and data related issues, empirical results and conclusion.

2. Literature Review and General Framework

The subject of foreign direct investments has been one of the central topics in a large variety of scientific studies in the last 30 years. Due to its importance for the economical environment of a country or region this topic has started impressive debates and controversies among scholars, government representatives, company representatives and also NGO's representatives.

As many scholars have shown, the foreign direct investments have been regarded as the "holy grail" by the governments of the great majority of the developing countries. Starting from the '80, the phenomenon of foreign direct investments has increased its intensity due to globalization and also to the fact that governments and foreign investors have shifted their approach towards a more collaborative side. Murtha and Lenway have shown, in a research paper published in 1994, that governments lowered taxation levels and have also designed new policies and regulations in this field with the sole purpose of attracting as much foreign direct investment as possible.

Following this direction, I need to state that foreign direct investments have been regarded as a major source for fuelling the economic growth by the governments of all the ex communist countries from Europe, during the last 24 years. The foreign direct investments were regarded as a source of capital and also of other benefits which could have been easier obtained through such a method. Therefore I can state that these foreign investments were considered responsible for bringing: new and superior management skills and also new and improved technologies, in the host countries. They were also regarded as an important factor in providing new and better paid jobs, new and more competitive products and services.

Bearing in mind these benefits (and many others) brought in a country by foreign direct investments, many scholars have focused their research on studying the problematic aspect of the main determinants responsible for attracting a foreign investor in a specific location. Scholars have conducted studies, following this research direction, both at national and at regional level.

As I have stated earlier, the literature provides evidence supporting the hypothesis that foreign direct investments are an important tool which can catalyze the development at regional level and, by consequence some of the main determinants should be looked for at regional level (Porter (2003)).

One of the main determinants of the foreign direct investments identified by a large number of studies is represented by the market size. Remarkable in this regard are the studies conducted by Crozet, Mayer and Mucchielli (2004), Przybylska and Malina (2000) and Ghemawat and Kennedy (1999), Cleeve (2008) and Schneider and Frey (1985).

The infrastructure was also identified as an important determinant of the foreign direct investments by many scholars, both at national and at regional level. Such studies are those published by Wei et al in 1998, by Mariotti and Pischitello in 1995, and by Broadman and Sun in 1997. Dunning, in a study published in 1998, argues that infrastructure represents a significant advantage of a location, when talking about foreign direct investments, because it is responsible for improving the potential to exploit the available resources. Noteworthy regarding the infrastructure, is the fact that studies conducted in this field asses the importance of the communication infrastructure (Asiedu (2002) and Khadaroo and Seetanah (2009)) and also of the transport infrastructure (Khadaroo and Seetanah (2009)).

Another important determinant identified by the academicians, who have studied aspects connected with the localization process of foreign direct investments, is represented by the characteristics of the labor market. Crozet, Mayer et al, in a study released in 2004, and Lansbury et al. in a study published in 1996, stress the importance of the availability and the price of the labor force in attracting foreign investors in a location. Wheeler and Moody provide evidences, in a study published in 1992, that between the inflows of foreign direct investments and the average wage there is a positive relation. In their studies, Vijayakumar et al. (2010) and Schneider and Frey (1985) support (through their findings) the idea that foreign direct investments are attracted into locations where the labor costs are low.

Research and development level and the human capital are other important determinants which have a significant importance in attracting foreign direct investments in a host country or region. Cantwell and Iammarino, in a study published in 2001, argue that the research and development level of an economy represent an important factor considered by foreign investors who intend to locate a future investment. Evidences in the same direction are provided by Cantwell and Piscitello (2005) and also by Chung and Alcácer, (2002). Dunning, in a study released in 2001, argues that the human capital positively influences the inflows of foreign direct investments. In the same direction point the findings reported by Cleeve (2008) and Al-Sadig (2009) which show that the secondary school enrolment has a positive impact on the inflows of foreign direct investments.

The literature also indicates: trade openness, government regulations, corruption, political stability and macroeconomic stability as key drivers of the foreign direct investments attraction. Al-Sadig (2009), Cleeve (2008) and Asiedu in 2002, all support the hypothesis that between trade openness and the inflow of foreign direct investments exists a positive correlation. Vijayakumar et al reach the same conclusion in a study published in 2010. Important for the literature concerning the interdependencies between the government regulations and the inward foreign direct investments is the study published by Morrissey and Udomkerdmongkol (2012) who repots a positive link between these two aspects. Schneider and Frey (1985) show that incoming foreign direct investments are encouraged by political stability. Their results are further confirmed by Asiedu in his study published in 2006. Noteworthy regarding our topic is that Al-Sadig (2009) and Asiedu (2006) provide evidences, in their studies, supporting the idea that increasing corruption level is regarded by the foreign investors as a significant disadvantage of a potential location. These findings are reinforced by the results obtained by Cleeve in his study, released in 2008, and by Wei in 2000. Scholars used unemployment level or the inflation rate as a proxy for the macroeconomic stability and proved that foreign investors incline to locate their future investments in countries with a higher stability level.

However, I need to clearly state the fact that, even though the determinants described above were all identified and studied in a large variety of studies, some of them manifest their influence mainly at country level and less at regional level. Nevertheless these phenomena should be studied by the Romanian policy makers when trying to identify the determinants of the foreign direct investments, at regional level.

Also, when talking about foreign direct investments at regional level, of significant importance is the study of the disparities registered between different regions and their underlying causes. The registered disparities concern different domains like: labor market (Taylor and Bradley (1997)), tourism (Xue (2005), Soukiazis and Proença (2008)), infrastructure (Démurger (2001)) and other important socio-economic aspects (Singh, Kogan, et al (2008)).

At national level, even though the literature is not very vast, the disparities among regions are the main topic of the studies published by Boldea, Parean et al in 2012 and by Goschin, Constantin et al in 2008. Also noteworthy in the context of our study is the research paper published by Danciu and Strat (2012) where, based on micro economic level data, the authors analyze the potential of the Romanian regions in attracting foreign direct investments in the manufacturing sector.

2. Methodology, Research Goal and Data Issues

Three important aspects of the research are discussed along this section. First of all, the main objective of the research is presented and described. In the second part of the section the focus is on the employed methodology and on aspects related to the administrative divisions of Romania. Finally the third part of the section deals with issues related to the data used in this study.

2.1. Research Goal

The main goal of the present research is to analyze the evolution of the localization process of foreign direct investments at the level of the Romanian counties, during the period 2001 – 2012. Therefore, I will try to emphasize any changes in the trend of the localization process and I will also try to connect these turning points with the most important events which took place in Romania in that particular period (events that might have had an impact on the attractiveness of the Romanian counties in the eyes of foreign investors). Following this approach, the analysis will be conducted with respect to the three important milestones that have occurred during this period, namely: the year 2004 when Romania became a member of NATO, the year 2007 when Romania became a member of the European Union and the year 2009 when the economical crisis brought its effects in Romania.

2.2. Administrative divisions of Romania

Romania is organized, from an administrative perspective, into 41 counties and a capital city named Bucharest. These counties serve as NUTS III units. The capital city is also divided into six administrative entities named Sectors.

After the end of the communist era, which took place in 1990, Romania decided to redesign its administrative and spatial organization from a highly centralized model to a new framework based on a regional perspective. An important milestone during this transformation process was the year 1998 when eight development regions were created. This eight development regions, serve as NUTS II units and their names are: North - East development region, South - East development region, South development region, South - West development region, West development region, North-West development region, Centre development region and Bucharest - Ilfov development region. Important to mention is the fact that, after the crucial moment which took place in 1998, no other significant events were registered in this regard. Moreover, I need to mention that these development regions are not fully functional administrative regions even though Romania became a member of the European Union since the 1st of January 2007.

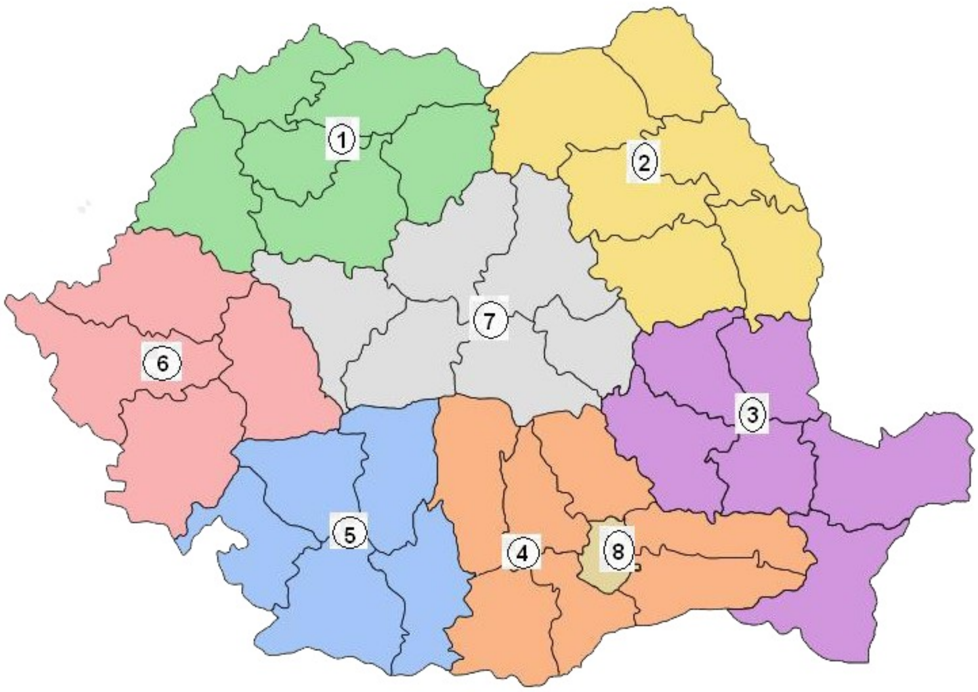


Figure 1. The administrative organization of Romania

2.3. Data Issues

The analysis presented in this paper is conducted on the series of stocks of foreign direct investments registered at the county level for the period 2001 – 2012. The national value of the stock used is calculated by summing up the individual values, for each year. The data were gathered from the database of the Romanian National Trade Register Office (the database is available online on the website of the institution). Due to comparability reasons and also in order to improve the relevance of our results all the values (for all the counties) were expressed as percentages from the stock registered at national level.

3. Empirical Results

The evolution of the stocks of foreign direct investments at the level of the Romanian counties might be regarded as an important indicator of their economical development and, therefore, studying this aspect should be considered as being very important by the policymakers.

First of all, when talking about the inward foreign direct investments, it is important to mention that the stocks of foreign direct investments for Romania (calculated as a sum of the stocks registered at county level) have increased from 9119942.4 Euro to a value of 32939762.5 Euro over the period 2001 – 2012. The entire evolution of the growth rhythms is displayed in the chart from Fig. 2. Until 2003 the stocks have slightly decreased and afterwards the trend was constantly positive. For the period between 2004 and 2007 the growth rate was decreasing constantly from 25.7% to little over 15.5%. Therefore, when shifting our perspective from absolute values to percentages, I can assert that, even though the period between the NATO accession and the European Union accession has been a favorable period, Romania's attractiveness for foreign investors has diminished. Going further, I notice that the growth rhythm has increased in 2008 (the last year before the economical crisis started to affect the Romanian economy) at a value over 22%. In 2011 and 2012 the growth rhythm has been significantly lower, with values under 10%, but the positive trend has reappeared.

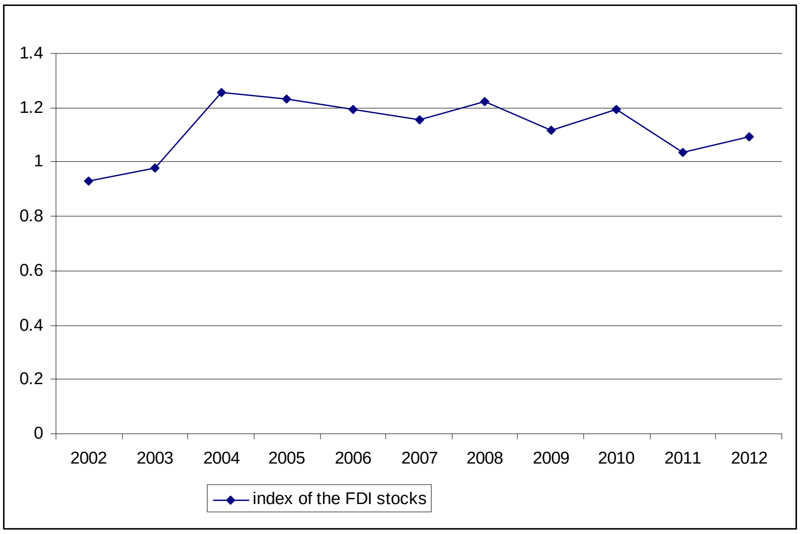


Figure 2. The dynamic of the FDI stocks (index with a moving base) for Romania
Source: Author's work

Taking the analysis further I will continue to asses the evolution of the stocks of foreign direct investments at the county level. In order to facilitate the comparison I have expressed the stock of foreign direct investment for each county as a percentage from the stock of FDI registered at national level.

Before moving on with the analysis I will present the evolution for the development region Bucharest Ilfov, due to its essential particularities. First of all, it is very important to mention that this region has received over 50% of the total foreign direct investments located in Romania. The percentage has increased from little over 50% in 2001 to over 59% in 2012. The entire evolution is displayed in the chart listed in the figure number 3.

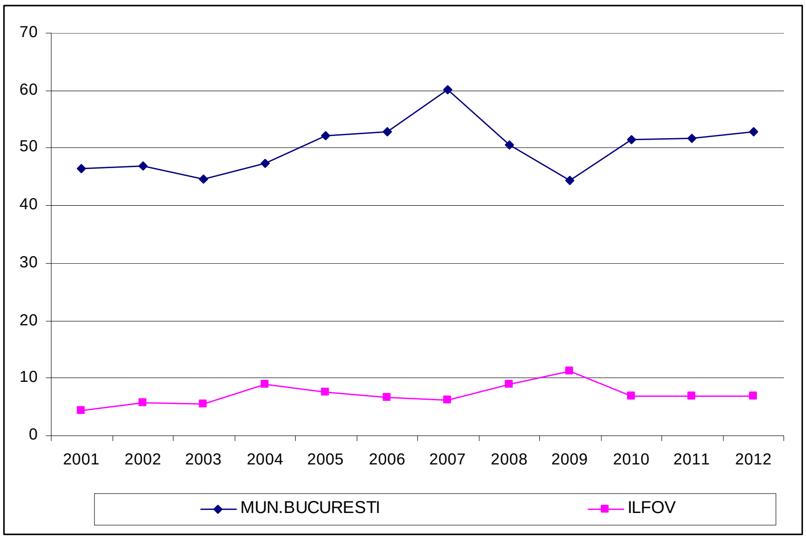


Figure 3. The dynamic of the FDI stocks (% from national value) for Bucharest and Ilfov
Source: Author's work

In the evolution of the stocks of foreign direct investments for Bucharest, an important milestone represents the first year (2007) after Romania was accepted as a member of the European Union. In 2007 the stocks attracted in Bucharest represented over 60% from the national level stocks. The lowest level for Bucharest was registered in 2009 when the stocks for Ilfov reached their highest value (over 11%). After 2009, the stocks for Ilfov constantly represented around 6.8% from the national stock and those for Bucharest increased easily from 51% to 52%.

The particularities presented by Bucharest might be described as "king effect", an effect which was observed and described by Jefferson (1939), by Laherrere and Sornette (1998) and by Roy Cerqueti and Marcel Ausloos (2014). Thus, Bucharest, even though is ranked first (it has the highest attractiveness) it attracts a percentage much, much larger, having in this way the behavior of an outlier.

Therefore, due to the fact that Bucharest represents an outlier among the Romanian counties due to its attractiveness for foreign investors, I have decided to continue the analysis without the entire Bucharest-Ilfov development region, namely Bucharest and the county Ilfov.

After dropping the Bucharest-Ilfov development region from our analysis, the percentages reported for each county were calculated based on the national stock's value calculated by summing up the stocks for all the 40 Romanian counties (except Ilfov and the capital city Bucharest).

In the table number 1 I have displayed the evolution of the stocks of foreign direct investments (presented as percentages from the national value) for the best performing five counties for each year covering the period 2001 – 2012.

Table 1. The best performing five counties
(Stocks of FDI expressed as % from national value)

	2001		2002		2003		2004
Galati	13.87	Galati	12.37	Galati	12.75	Arges	15.04
Timis	11.47	Constanta	10.34	Arges	11.90	Galati	11.73
Prahova	9.02	Timis	10.09	Constanta	9.63	Constanta	8.64
Arges	8.32	Prahova	8.17	Timis	9.41	Timis	7.92
Cluj	5.13	Arges	6.93	Prahova	7.40	Cluj	6.73
	2005		**2006**		**2007**		**2008**
Arges	14.13	Arges	12.83	Arges	12.89	Arges	10.23
Galati	10.68	Timis	10.01	Timis	9.47	Mures	8.65
Timis	10.12	Galati	9.28	Galati	8.66	Timis	8.48
Constanta	8.90	Constanta	7.47	Constanta	6.65	Bacau	7.25
Cluj	6.10	Cluj	5.68	Cluj	5.60	Galati	6.96
	2009		**2010**		**2011**		**2012**
Timis	8.81	Timis	8.74	Timis	9.14	Timis	9.18
Mures	7.04	Bihor	7.87	Bihor	7.65	Bihor	7.66
Cluj	6.71	Mures	7.06	Mures	6.93	Mures	6.66
Brasov	6.24	Brasov	5.87	Constanta	6.00	Brasov	6.37
Arges	6.09	Constanta	5.84	Brasov	5.75	Constanta	5.77

Source: Author's work

During the analyzed period the top five modified significantly. Galati County who was leading the hierarchy in the first three years became less attractive with time and finally exited the top five after 2008. Another notable evolution was registered by Arges County who lead the hierarchy between the years 2004 and 2008, due to the investment made by Renault in the Dacia factory from Mioveni. Starting from 2009, Arges County's attractiveness declined, and then it finally left top five in 2010. The best performing county for the last four years is the Timis County which was present in the top for the entire analyzed period. Another important aspect that emerges is the fact that most of the counties present in the top five are located in Transylvania. The only counties outside Transylvania are Galati, Prahova, Arges and Constanta, all of them the being located in the south and south east of the country (an exception is Bacau who appears in the top in 2008).

Notable is the fact that the leading five counties decrease their importance (as percentage of the stocks in the national stock) at national level for the analyzed period suggesting that other parts of the country have become more attractive for the foreign investors. The negative trend starts in 2003 and it ends in 2008 (at the debut of the economical crisis) when the trend becomes positive. Another noteworthy aspect is the fact that the leading five counties account for over 35% of the FDI stocks in the present, after their importance reached in 2003 values over 50%.

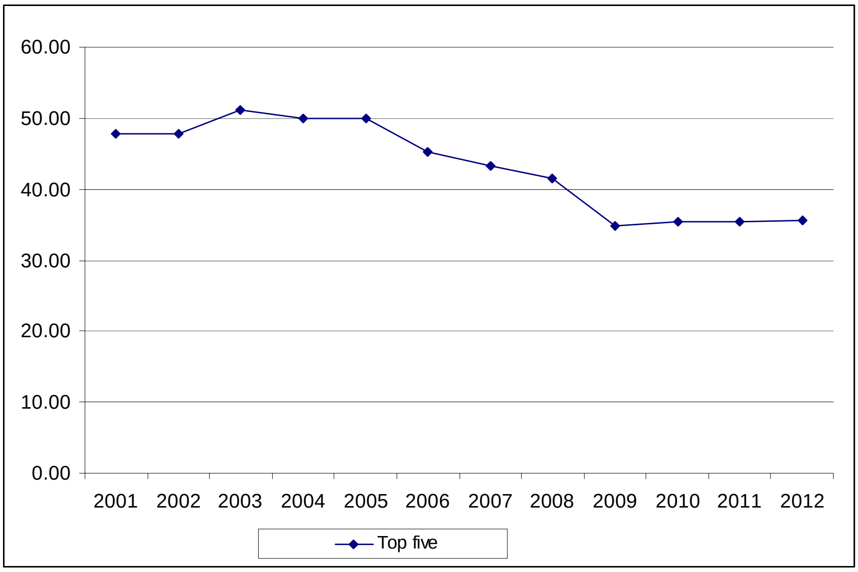


Figure 4. The importance of the leading five counties (Stocks expressed as % from the national value)

Source: Author's work

The following table displays the evolution of the stocks of foreign direct investments (presented as percentages from the national value) for the least attractive five counties, for each year covering the period 2001 – 2012.

Table 2. The least attractive five counties (Stocks of FDI expressed as % from national value)

	2001		**2002**		**2003**		**2004**
Salaj	0.32	Tulcea	0.35	Tulcea	0.25	Salaj	0.29
Bacau	0.23	Salaj	0.29	Giurgiu	0.24	Giurgiu	0.21
Giurgiu	0.12	Giurgiu	0.20	Ialomita	0.21	Ialomita	0.20
Botosani	0.10	Botosani	0.14	Botosani	0.12	Botosani	0.10
Gorj	0.05	Gorj	0.04	Gorj	0.04	Gorj	0.04
	2005		**2006**		**2007**		**2008**
Salaj	0.26	Bistrita-Nasaud	0.37	Bistrita-Nasaud	0.39	Bistrita-Nasaud	0.37
Botosani	0.22	Vrancea	0.30	Vrancea	0.32	Vaslui	0.36
Giurgiu	0.22	Botosani	0.25	Botosani	0.26	Ialomita	0.33
Ialomita	0.20	Ialomita	0.24	Ialomita	0.22	Gorj	0.21
Gorj	0.03	Gorj	0.02	Gorj	0.02	Botosani	0.21
	2009		**2010**		**2011**		**2012**
Valcea	0.48	Valcea	0.37	Valcea	0.36	Ialomita	0.32
Teleorman	0.44	Ialomita	0.33	Ialomita	0.32	Botosani	0.26
Tulcea	0.38	Vaslui	0.27	Vaslui	0.26	Vaslui	0.25
Vaslui	0.32	Botosani	0.25	Botosani	0.24	Gorj	0.14
Botosani	0.22	Gorj	0.15	Gorj	0.14	Vrancea	0.11

Source: Author's work

As it is visible from the data presented in the table, Gorj County is the least attractive county for foreign investors in nine of the analyzed years. This position is occupied by Botosani in two years and by Vrancea in 2012. Noteworthy is the fact that in 2001 Bacau was among the least performing five counties and in 2008 it entered top five, best performing counties on the fourth position (it was present among the best performing counties for only one year).

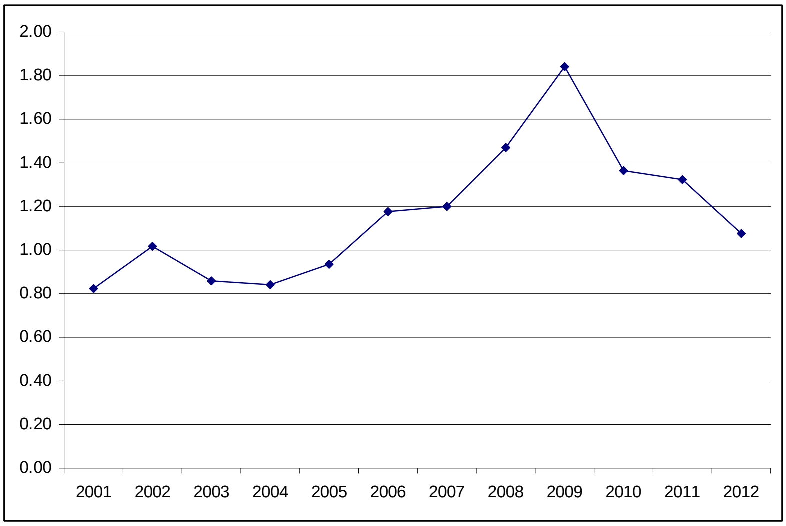


Figure 5. The importance of the last five counties
(Stocks expressed as % from the national value)
Source: Author's work

Important to mention is the fact that these five least attractive counties are responsible constantly, for under 2% of the entire stock of FDI attracted at national level (except Bucharest-Ilfov development region). Their importance increases until 2009 (almost 1.85%) and afterwards decreases sharply to 1.08%. Therefore, it is obvious that since the economic crisis has appeared their attractiveness has decreased significantly. Summarizing these results, I can state that, starting with 2009, the remaining 30 counties account for around 64% from the entire national stock of foreign direct investment (except Bucharest-Ilfov development region).

Before going further, I consider necessary to mention that for the first years of the analyzed period, when analyzing the behavior of the best five performing counties we might identify a "king plus viceroy effect" which was also mentioned by Roy Cerqueti and Marcel Ausloos (2014). This effect is similar with the "king effect" mentioned earlier in this paper, but it involves the presence of more outliers. The "king plus viceroy effect" fades away with the passage of time. Due to the fact that the "king effect" that we encounter in the case of Bucharest is obvious and constant, we will present a graphical description of the "king and viceroy effect" indentified for the best performing counties in Appendix A.

An even better description of how the attractiveness of the Romanian counties modified in the eyes of foreign investors is visible when analyzing the evolution of the Gini coefficient. The values of the coefficient range from 0 to 1, and higher values, close to 1 show an important concentration of the analyzed phenomenon, indicating that it is possible to talk about some important concentration poles regarding this phenomenon.

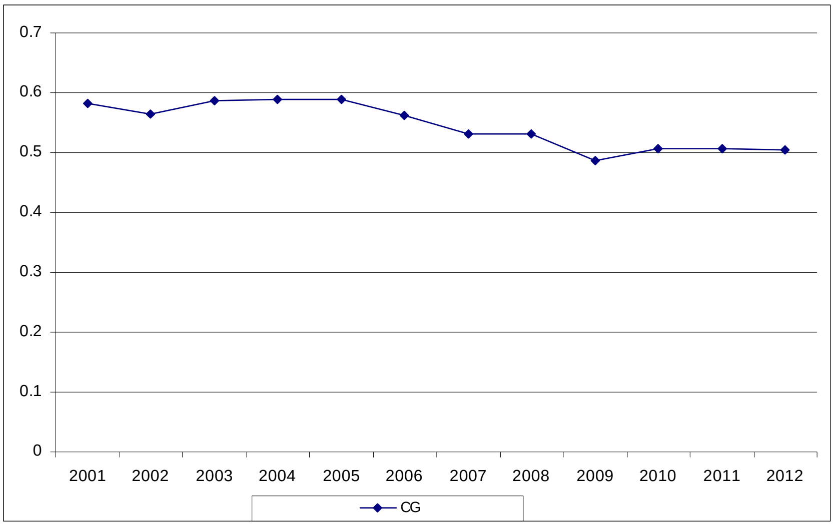


Figure 6. The evolution of the Gini coefficient
Source: Author's work

As it is visible from the chart displayed in figure number 6 the concentration level has decreased from values little under 0.6, during the period 2003 – 2005, to values around 0.5, during the last three years. The lowest concentration level is registered in 2009, which is the year when the global crisis affected the Romanian economy. In the last analyzed three years, the concentration level is somehow constant suggesting that after the year 2009 no significant events were registered regarding the evolution of the attractiveness of the Romanian counties in the eyes of foreign investors.

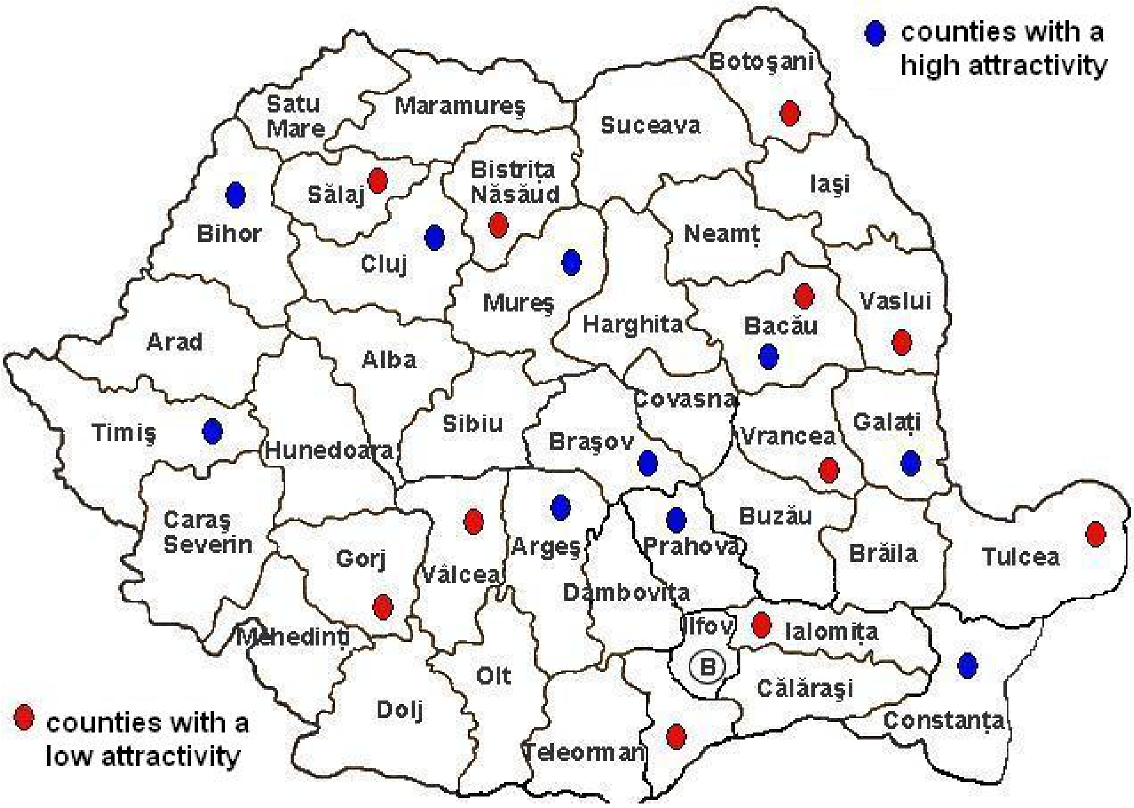


Figure 7. The most attractive and the least attractive counties

Summarizing the entire analysis, I can assert that the analyzed period might be divided into three main parts: 2001 – 2003 (2004), 2004 – 2009 and 2010 – 2012. In the first period, the coefficient was somehow stable with high values, around 0.58, suggesting that there are some poles that attract the majority of the inward foreign direct investments. This aspect is clearly visible when analyzing the figures displayed in Table 1 and Table 2. In the second period, two main events occurred, namely: Romania became a member of NATO and Romania became a member of the European Union. The value of the coefficient decreased constantly during this period showing that the importance of the concentration poles was decreasing. Therefore I can assert that these two events increased the investors' confidence in the potential of the Romanian counties' economy. Finally, the third period (after the economical crisis brought its effects in Romania) is characterized by a stability of the coefficient around the value 0.5 (the coefficient increases from 0.48 in 2009). Therefore, noteworthy for this period is the fact that the economical crisis had a significant impact on the process registered during the previous period and diminished significantly its intensity.

Using the territorial display of the information, presented in Figure 7, it becomes easily observable that the most attractive counties and the least attractive counties tend to agglomerate. Therefore, it becomes obvious that the policymakers need to construct a new regional policy which should be designed with the clear purpose to increase the attractiveness of the poorer counties (and provide therefore better opportunities for them). Following this direction, I can suggest that the administrative policy, based on counties (very small entities in the present European context), needs significant improvement due to the fact that it facilitates the increasing of the disparities in this field.

4. Conclusions

Summarizing the study, I can say that the present paper should be included among scientific works who analyze the discrepancies registered at regional level regarding the attractiveness of the Romanian counties in the eyes of the foreign investors.

An important aspect described in the present study (an aspect which confirms other analysis conducted earlier) shows clearly that, in Romania, significant discrepancies are registered between counties when talking about the stocks of received foreign direct investment. Noteworthy is the fact that, the five leading counties are responsible for over 35% ("king and viceroy effect") of the received foreign direct investments while the least attractive five counties are only responsible for under 1.8% of the national stock (national stock except the stock of FDI attracted by Bucharest-Ilfov development region). Also, following the same logic, I need to mention the fact that Bucharest ("king effect") - Ilfov development region has attracted almost 60% of the total stock of foreign direct investment during the analyzed period, showing that the remaining seven development regions have a very low attractiveness level. Therefore, a significant disequilibrium in the spatial structure of the Romanian economy might be suggested.

The most important piece of information brought by this study is represented by the description of the evolution of the spatial concentration of the stocks of foreign direct investment described with the help of the Gini coefficient. Thereby, during the analyzed period, I can identify three tendencies regarding the evolution of the FDI stocks, at county level. Until 2004 the coefficient was stable around high values, about 0.58, suggesting that some counties were receiving the majority of the foreign direct investment, being therefore important development poles. Following this stage, the period between 2004 and 2009 (Romania became NATO member in 2004 and EU member in 2007) was characterized by a decreasing trend suggesting that the importance of the concentration poles was decreasing and that the investor' confidence in the potential of the other counties was increasing. Finally, after 2009 the impact of the global crisis was significant, leading to an increase of the concentration, suggesting that the importance of the concentration poles increased again.

This phenomenon, which appeared in this field in Romania after the year 2009, might be compared with the one identified by researchers at global level, namely that the developed entities become more attractive for foreign investors and the poorer ones loose their attractiveness slowing in this way the convergence process (sometimes the discrepancies increase).

Concluding, I might state that the difference between the leading counties and the others is too important to be recovered in a medium term time period. Following the same logic I might say the same thing about Bucharest and the rest of the country. In these conditions I suggest that constructing a new, viable and functional regional policy might be an appropriate solution to tackle this problem.

Knowing the magnitude of the phenomenon, it becomes obvious that the counties are too small (economies) and their power to implement policies which will boost up their attractiveness in the eyes of the foreign investors is limited. Therefore constructing functional regional administrative units with a greater strength might be the solution.

As a final remark I can state that constructing viable regional development units should be the first priority for the Romanian government in order to increase the competitiveness of the Romanian economy in the European context. By doing so, the central authorities will also give the possibility to the regional/local authorities to construct integrated policies at the regional level, in order to ensure a sustainable development (economical and also social) for the entire community.

Acknowledgments

This work was supported from the European Social Fund through Sectorial Operational Programme Human Resources Development 2007 – 2013, project number POSDRU/ 159/1.5/S/142115, project title "Performance and Excellence in Postdoctoral Research in Romanian Economics Science Domain"

References

1. Al-Sadig, A. **The effects of corruption on FDI inflows**. The Cato Journal, 29, 2009, pp. 267
2. Asiedu, E. **On the Determinants of Foreign Direct Investment to Developing Countries: Is Africa Different?** World Development, Vol. 30, 2002, pp. 107-119
3. Asiedu, E. **Foreign Direct Investment in Africa: The Role of Natural Resources, Market Size, Government Policy, Institutions and Political Instability**. World Economy, Vol. 29, 2006, pp. 63-77

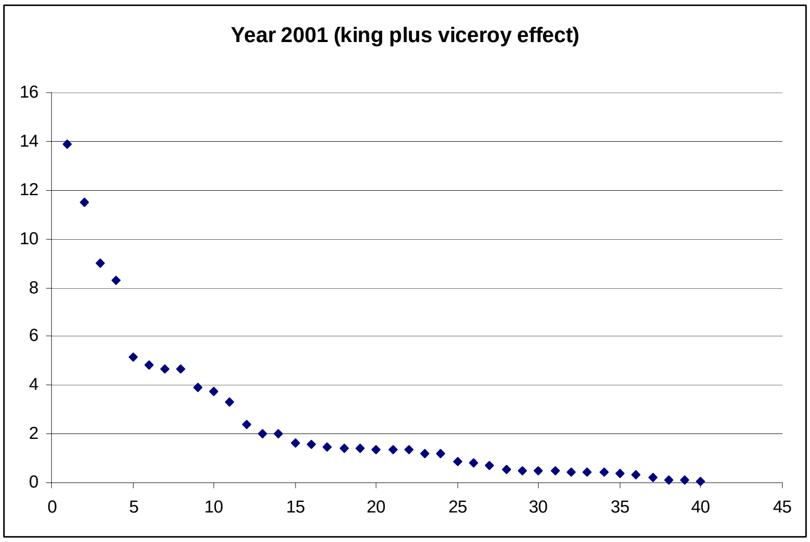
4. Boldea, M., Parean, M. and Otil, M. **Regional Disparity Analysis: The Case of Romania.** Journal of Eastern Europe Research in Business & Economics, 2012, pp. h1-10
5. Broadman, H. G. and Sun, X. **The Distribution of Foreign Direct Investment in China.** The World Economy, Volume 20, Issue 3, 1997, pp 339–361
6. Cantwell, J.A. and Iammarino, S. **EU regions and multinational corporations: change, stability and strengthening of technological comparative advantages** Industrial and Corporate Change, Vol. 10, 2001, pp. 1007–1037
7. Cantwell, J.A. and Piscitello, L. **Recent Location of Foreign-owned Research and Development Activities by Larger Multinational Corporations in the European Regions: The Role of Spillovers and Externalities.** Regional Studies, Vo. 39, No. 1, 2005, pp. 1-16
8. Cerqueti, R. and Ausloos, M., **Evidence of economic regularities and disparities of Italian regions from aggregated tax income size data.** Physica A: Statistical Mechanics and its Applications, Vol. 421, 2015, pp. 187-207
9. Cleeve, E. **How Effective Are Fiscal Incentives to Attract FDI to Sub-Saharan Africa?** The Journal of Developing Areas, Vol. 42, 2008, pp. 135-153
10. Chung, W. and Alcácer, J. **Knowledge Seeking and Location Choice of Foreign Direct Investment in the United States.** Management Science, Vol. 48, No. 12, 2002, pp. 1534-1554
11. Crozet, M. & Mayer, Th. & Mucchielli, J-L. **How do firms agglomerate? A study of FDI in France.** Regional Science and Urban Economics, Elsevier, vol. 34(1), pp 27-54, January, 2004.
12. Danciu, A. R. and Strat, V. A. **The FDI profile in the Romanian manufacturing sector**. Review of Applied Socio-Economic Research, vol. 4, Vol. 2, 2012, pp. 57-64
13. Danciu A. R. **The Ranking of the Romanian regions based on the potential to attract FDI**, Procedia Social and Behavioral Science, vol.62, 2012, pp 40-44
14. Alexandru(Davidescu) A. A. **Estimating the size of the Romanian shadow economy a labour approach**, Journal of Social and Economic Statistics, vol.3, no.3, 2014, pp. 25-37
15. Davidescu A., Strat, V. A. **Coordinates of a New Romanian Regional Policy-Identifying the Development Poles. A Case Study.** Informatica Economica, vol 18, no. 2, 2014, pp. 88-99
16. Démurger, S. **Infrastructure Development and Economic Growth: An Explanation for Regional Disparities in China?** Journal of Comparative Economics, Vol. 29, No. 1, 2001, pp. 95–117
17. Dunning, J. H. **Location and the Multinational Enterprise: A Neglected Factor?**. Journal of International Business Studies, Vol. 29, 1998, pp. 45-66
18. Dunning, J. H. **The eclectic (OLI) paradigm of international production: Past, present and future**. International Journal of the Economics of Business, Vol. 8, 2001, pp. 173-190

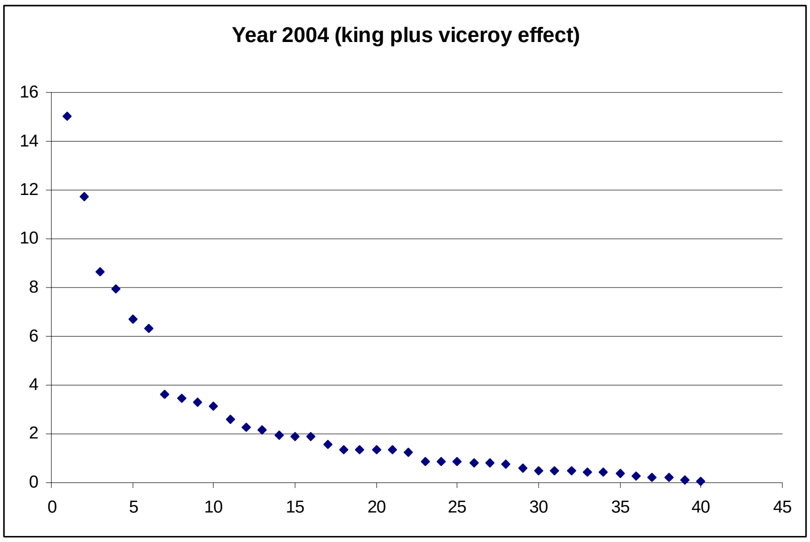
19. Ghemawat, P. and Kennedy, R. **Competitive Shocks and Industrial Structure: The Case of Polish Manufacturing**. International Journal of Industrial Organization , Vol. 17, No. 6, 1999, pp. 847-867
20. Goschin, Z., Constantin, D.-L., Roman, M. and Ileanu, B. **The current state and dynamics of regional disparities in Romania**, Romanian Journal of Regional Sciences, Vol. 2, no 2, 2008, pp. 80-105
21. Jefferson, M. **The Law of Primate City**. Geographical Review, Vol. 29, No. 2, 1939, pp. 226-232
22. Khadaroo, J. and Seetanah, B. **The Role of Transport Infrastructure in FDI: Evidence from Africa using GMM Estimates**. Journal of Transport Economics and Policy (JTEP), Vol. 43, 2009, pp. 365-365
23. Laherrere, J. and Sornette, D. **Stretched exponential distributions in nature and economy fat tails with characteristic scales**. European Physics Journal B, Vol. 2, No. 4, 1998, pp. 525-539
24. Lansbury, M., N Pain, N., and Smidkova, K. **Foreign direct investment in Central Europe since 1990: an econometric study**. National Institute Economic Review, No. 156, 1996, pp.104-113
25. Mariotti, S. and Piscttello, L. **Information cost and location of FDIs within the host country: empirical evidence from Italy**, Journal of International Business Studies, Vol. 26, 1995, pp. 815-840
26. Morrisey, O. & Udomkerdmongkol, M. **Governance, Private Investment and Foreign Direct Investment in Developing Countries**. World Development, Vol. 40, 2012, pp. 437-445
27. Murtha, T. and Lenway, S. **Country Capabilities and the Strategic State: How National Political Institutions Affect Multinational Corporations' Strategies**, Strategic Management Journal, Vol. 15(summer), 1994, pp. 113-129
28. Popovici O. C., Calin A. C. and Simionescu M. **Investigating FDI inflows in Romania through an ARMA model, Internal Auditing & Risk Management**, No. Nr.3 (35), 2014, pp. 49-65
29. Porter, M. **The Economic Performance of Regions**, Regional Studies, Vol. 37, No. 6&7, 2003, pp. 549-578
30. Przybylska, K. and Malina, A. **The Determinants of Foreign Direct Investment in Transforming Economies: Empirical Evidence From Poland**, Statistics in Transition, Vol. 4, No. 5, 2000, pp 883-899
31. Schneider, F. and Frey, B. S. **Economic and political determinants of foreign direct investment**. World Development, Vol. 13, 1985, pp.161-175
32. Singh, G. K., Kogan, M. D., van Dyck, P. C. **A Multilevel Analysis of State and Regional Disparities in Childhood and Adolescent Obesity in the United States**, Journal of Community Health, Vol. 33, No. 2, 2008, pp. 90-102
33. Soukiazis, E. and Proença, S. **Tourism as an alternative source of regional growth in Portugal: a panel data analysis at NUTS II and III levels**, Portuguese Economic Journal, Issue 7, 2008, pp. 43–61

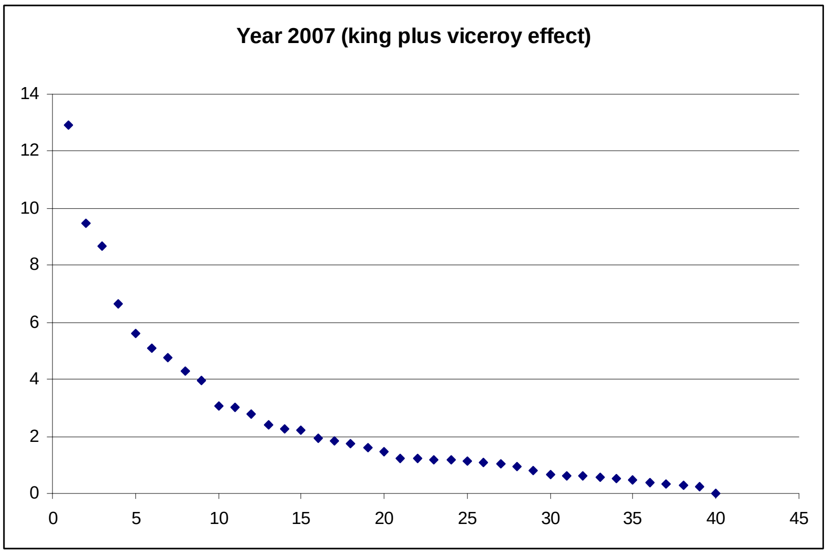
34. Taylor, J. and Bradley, S. **Unemployment in Europe: A Comparative Analysis of Regional Disparities in Germany, Italy and the UK**, International review for Social Sciences, Volume 50, Issue 2, 1977, pp 221–245
35. Vijayakumar N., P, S. and Rao, K. C. S. **Determinants of FDI in BRICS Countries: A panel analysis**. International Journal of Business Science and Applied Management, Vol. 5, 2010, pp. 1-13
36. Wei, Y., Liu, X., Parker, D. and Vaidya, K. **The regional distribution of Foreign Direct Investment in China**. Regional Studies, Vol. 33, No. 9, 1999, pp. 857-867
37. Wheeler, D. and Mody, A. **International investment location decisions**. Journal of International Economics, Vol. 33, 1992, pp. 57-76
38. Xue, Q. **An Analysis on the Regional Differences of China Tourism**, Journal of Guilin Institute of Tourism, Issue 1, 2005, pp. 004

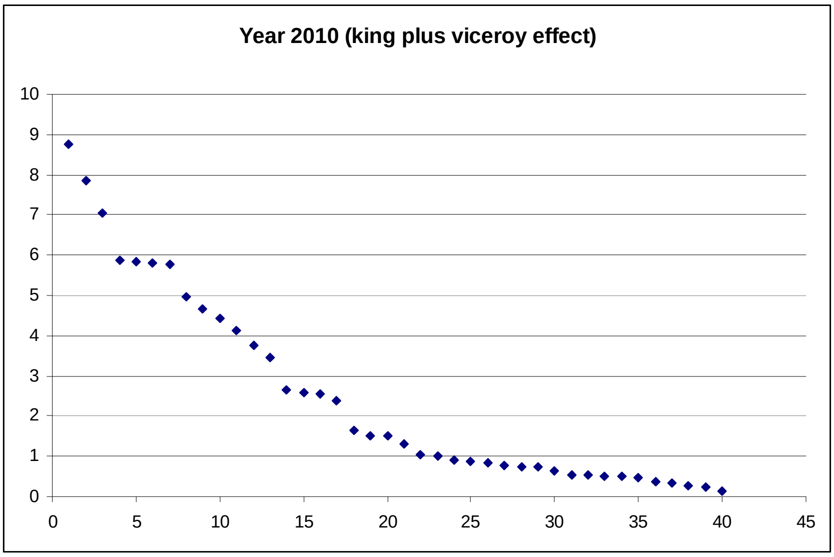
Appendix A

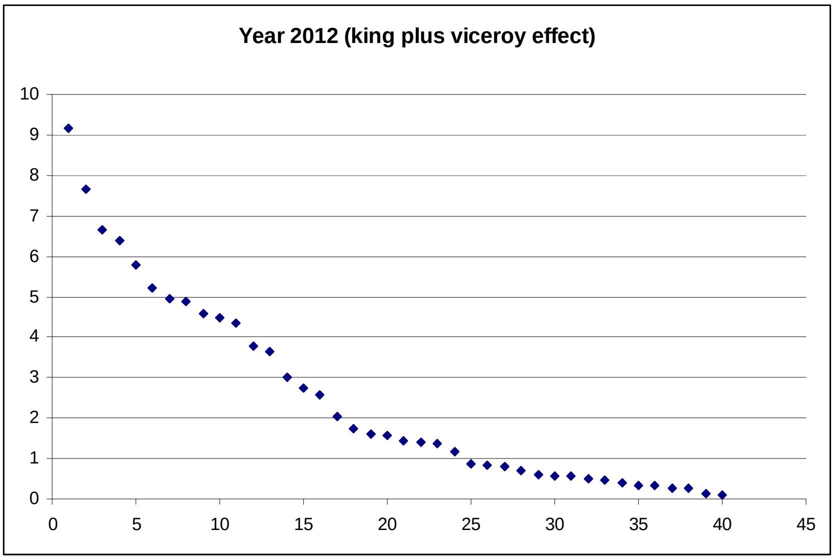
This Appendix contains figures which display the evolution of the "king plus viceroy effect" regarding the stocks of FDI received by the Romanian counties











Source: Author's work

VALUE-AT-RISK ESTIMATION ON BUCHAREST STOCK EXCHANGE

Olivia Andreea BACIU
PhD Candidate, Babes Bolyai University, Cluj Napoca, Romania

E-mail: oli_baciu@yahoo.com

Abstract

As an important tool in risk management, Value-at-Risk is estimated on the Romanian stock market based on single assets and a weighted portfolio of them. Because this is a measure of the extreme tails, several approaches are used to compute Value-at-Risk by taking in consideration the distribution of the data, namely the generalized hyperbolic distribution, Normal-Inverse Gaussian and asymmetric t-Student in comparison with the normal distribution. The considered period is divided into an analyze period and a test one, where based on the rolling windows approach are estimated the VaR values and then tested with the help of Kupiec's and Christoffersen's backtests. The choice of the time period affects the estimation due to the events that took place on the market and the approach based on the normal distribution predicted best the VaR values by underestimating the risk compared to the other distributions. This approach fits better the considered period because the analyzed period covers moments of severe economical crisis while the test period goes over a period of recovery.

Keywords: *Value-at-Risk; generalized hyperbolic distributions; heavy tails; stock market data*

1. Introduction

The instability of financial markets in the recent years has led to the need of a general accepted measure that quantifies risk of the loss when looking to invest or holding a portfolio. This measure of risk management has known various forms over time. A measure accepted and used both in practice and theory is value-at-risk (VaR). VaR was easily accepted for two reasons: the ease of computations and the weight of the information brought by a single number.

As defined by J.P. Morgan (1996), VaR represents "a measure of the maximum potential change in value of a portfolio of financial instruments over a pre-set horizon. VaR answers the question: how much can I lose with x% probability over a given time horizon?". In other words, VaR defines through a single number the maximum loss we should expect over a period of time except for x% of the cases, when the loss can exceed VaR. Due to this two inconveniences, the arbitrary choice of the probability and of the time interval, authors like Einhorn (2008) consideres VaR as being inefficient because it ignores mainly those x%

extreme situations, which represent the highest looses and compares the method with "an airbag that works all the time, except when you have a car accident".

Prause (1997) used the family of generalized hyperbolic distributions to model the returns of the German stocks and US index and based on them evaluated the risk of loss corresponding to each of the two markets. The analysis revealed that the family of generalized hyperbolic distributions provide a better fit to the empirical VaR. Bauer (2000) performed VaR computations on German stocks and DAX, Dow Jones and Nikkei market indexes over 1987- 1997 and found that the symmetric hyperbolic distribution outperforms the normal model. Huang et al. (2014) evaluated the performance of the hyperbolic distribution, Normal- Inverse Gaussian and asymmetric t- Student through VaR estimates on daily JSE Mining Index. All distributions outperformed the normal one both through means of goodness-of-fit and VaR estimates.

To overcome the shortcomings of the method, in this study are used daily returns and are applied several methods to compute VaR by taking into consideration the distribution of the returns. VaR estimates are computed using the rolling window approach and the performance of each method is tested through backtests. As shown in Prause (1999), Sognia and Wilcox (2014), Baciu (2014), appropriate distributions in modeling financial data are the generalized hyperbolic distribution or Normal- Inverse Gaussian distribution which account for the heavier tails and asymmetric distribution that characterize stock market returns. Although Vee et al. (2012) concluded that one specific distribution won't describe well indices from different markets and the choice of the distribution depends on the market, the performance of the generalized hyperbolic distribution on the index of the Romanian market was shown in Necula (2009) or Baciu (2014). This study uses the findings of Baciu (2014) in modeling financial returns and fitting the distribution to data, which show that the generalized hyperbolic distribution approximates the best the distribution of the returns of the five Romanian Investment Funds, followed by the Normal-Inverse Gaussian and asymmetric t-Student distributions. The purpose of this paper is to develop the previous study, in which the performance of each distribution was characterized through plots and goodness-of-fit means and focus on the distribution of the tails.

The remainder of this paper is organized as follows. Section 2 links the study with a previous research in which has been shown the performance of the generalized hyperbolic distribution on Romanian stock data in comparison to the normal distribution. Section 3 provides a description of the methodology used to compute VaR and the backtests of the VaR estimates. Section 4 introduces the data and presents the descriptive statistics of it. The VaR estimates and the results of the backtests are discussed in Section 5. Finally, Section 6 concludes the paper with the main findings and directions to develop the study.

2. Choice of the distribution

As documented in Baciu (2014), among the distributions that approximate well stock market returns is the family of the generalized hyperbolic distributions. Studies such as Prause (1997), Necula (2009), Rege and Menezes (2012), Sognia and Wilcox (2014) have shown the performance of this family of distributions. Because VaR accounts only for the tails, such a distribution should be more appropriate than the normal one to estimate the potential risk of loss. Based on the results in Baciu (2014) on the same data set as in the present paper, VaR will be computed on three of the generalized hyperbolic family of distributions: the generalized hyperbolic distribution, Normal- Inverse Gaussian and assymetric t- Student. The parameters of these distributions are estimated using the maximum likelihood estimation, implemented based on the EM scheme of Dempster et al (1977) in R software.

For a vector of observations $x_1, x_2, \dots, x_n$, the maximum likelihood estimation of the parameters $\lambda, \alpha, \beta, \delta, \mu$ is obtained by maximizing the log-likelihood function:

$$L(x_1, x_2, \dots, x_n;\ \lambda, \alpha, \beta, \delta, \mu) = loga + \frac{\lambda - \frac{1}{2}}{2} \sum_{i=1}^{n} \log(\delta^2 + (x_i - \mu)^2) +$$

$$+ \sum_{i=1}^{n} logK_{\lambda - \frac{1}{2}}(\alpha\sqrt{\delta^2 + (x_i - \mu)^2} + \sum_{i=1}^{n} \beta(x_i - \mu),$$

where a and K_λ are defined as:

$$a(\lambda, \alpha, \beta, \delta) = \frac{(\alpha^2 - \beta^2)^{\frac{\lambda}{2}}}{\sqrt{2\pi}\alpha^{\left(\lambda - \frac{1}{2}\right)}\delta^\lambda K_\lambda(\delta\sqrt{\alpha^2 - \beta^2})}$$

$$K_\lambda(x) = \frac{1}{2}\int_0^\infty y^{\lambda - 1} \exp\left(-\frac{1}{2}x(y + y^{-1})\right) dy.$$

In the previous study, the choice among distributions was done based on plots and goodness-of-fit measures: Kolmogorov-Smirnov distance, Log-likelihood and Akaike Information Criterion. From the beginning, plots ruled out the Variance-Gamma distribution and have shown that the family of the generalized hyperbolic distributions offers a better and more appropriate fit to the Romanian market data than the normal distribution. Based on the goodness-of-fit measures, the generalized hyperbolic distribution seems to represent the best fit to the given data.

3. Methodology

3.1. VaR estimation methods

The VaR methodology has been easily adopted and accepted both in theory and practice due to the ease of implementation and interpretation. Among the most popular methods of computing VaR can be mentioned the methods based on the historical data, Monte Carlo simulations or the assumption of normal distribution.

A survey realized by Perignon and Smith (2006) has shown that 73% of the interogated banks are using the historical method for computing the risk of loss. This method estimates the future loss based on past returns. Besides its simplicity, among the benefits of this method is that it allows for skewed and leptokurtic distributions of data. On the other side, the method takes into consideration only events that took place in the analyzed period of time and can not capture the effect of any other type of events. If the analyzed period has gone only through low fluctuations of volatility, then the computed loss will be too small in case the fluctuations grow.

The normal VaR estimation method assumes a normal distribution of the returns. But it is a known fact that equity returns exhibit heavier tails than in the case of the normal distribution. To avoid the shorcomings of the normal VaR method, it was introduced the Cornish-Fisher method, described in Favre and Galeano (2002) which accounts for distributions other than the normal one by considering the third and fourth moments.

If it is considered

$$CF_p = Q_p + \frac{S({Q_p}^2 - 1)}{6} + \frac{K({Z_p}^3 - 3Z_p)}{24} - \frac{S^2(2{Z_p}^3 - 5Z_p)}{36}$$

where S stands for assimetry , K for excess kurtosis, p the selected probability level and Q_p the corresponding quantile to the selected probability level, then, through the Cornish-Fisher method, VaR reduces to:

$$VaR_{CF} = -mean(R) - \sqrt{\sigma} CF_p.$$

Brown (2008) comments on the performance of VaR and sustains that the results are trustful only if they are backtested. For this reason, in this study data is divided into an analyze period and a test one. As sugested by Brown (2008), for a considered probability of 1% greater looses than estimated by VaR for daily returns, the analysis should be performed on 3 years of historical data.

Each VaR is computed based on a rolling windows approach on 750 historical daily returns, counting for about three years of daily trading data. Data is divided into two periods of time: an analysis one and a test one. The analysis period represent the windows of the rolling windows approach and at each step the oldest return is dropped and added the return of the following day of the last observation included in the previous window. The test period accounts for the remaining data of more than two years of tradings. Over this period are compared the number of days with excess loss than estimated through VaR with the expected number of such days, according to the considered probability.

Two of the most used backtests are the ones of Kupiec (1995) and Christoffersen (1998).

3.2. Backtesting

According to Christoffersen (2012), one can only estimate with the chosen probability if the actual loss will exceed the computed VaR or not. This set of events is similar with a Bernoulli trial, with a probability *p* that the loss exceeds VaR and *(1-p)* that it does not. The total number of days in which the loss is greater than VaR, denoted by *x*, follow a Binomial (n,p) distribution, where *n* is the total number of observations.

Kupiec backtest

The Kupiec backtest, introduced by Kupiec (1995), is a test of the failure rate. The null hypothesis assumes a rate of failures equal to the expected one.

$$H_0: p = \widehat{p},$$

where, $\hat{p} = x/n$.

As mentioned in Nieppola (2009), if the number of days with excess loss is too high compared to the chosen probability level, it indicates an underestimation of the risk, while a low number of exceptions suggest an overestimation of the risk.

The Kupiec backtest is constructed as a likelihood-ratio test, with the statistic given by:

$$TS_{exc} = \frac{p^x(1-p)^x}{(\frac{n-x}{n})^{n-x}(\frac{x}{n})^x}.$$

Under the null hypothesis $TS_{exc} \sim \chi^2_{(1)}$.

Christoffersen backtest

Aditional to Kupiec backtest, Christofferson (1998) highlights the importance of the exceptions to be independent. Otherwise, if the excess loss occurs in a short period of time the company might be affected unlike if they occur occasionally, over a longer period. The new test keeps track of both events: the number of exceptions and their independence.

As defined in Christoffersen (1998), Christoffersen (2012) or Nieppola (2009), let

$$I_{t+1} = \begin{cases} 1 & pentru\ R_{t+1} < -VaR_{t+1} \\ 0 & pentru\ R_{t+1} \geq -VaR_{t+1} \end{cases}$$

be the sequence of exceptions and

$$\pi_1 = P(I_{t+1} = 1 | I_t = 1)$$
$$\pi_0 = P(I_{t+1} = 1 | I_t = 0)$$

the probability that tomorrow's loss will exceed VaR knowing that today's loss exceeds VaR, respectively the probability that tomorrow's loss exceeds VaR knowing that today's loss does not exceed VaR.

Let:

- n_{00} the number of days in which $I_{t+1} = 0$ based on $I_t = 0$
- n_{01} the number of days in which $I_{t+1} = 1$ based on $I_t = 0$
- n_{10} the number of days in which $I_{t+1} = 0$ based on $I_t = 1$
- n_{11} the number of days in which $I_{t+1} = 1$ based on $I_t = 1$.

Then, the probabilities π_0 and π_1 are:

$$\pi_0 = \frac{n_{01}}{n_{00} + n_{01}}, \pi_1 = \frac{n_{11}}{n_{10} + n_{11}}, \pi = \frac{n_{01} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}}.$$

Because the test combines the hypothesis of the Kupiec test with the one the independence of the exceptions, the test statistic is defined as:

$$TS = TS_{exc} + TS_{ind}$$

where,

$$TS_{ind} = -2\ln\left(\frac{(1-\pi)^{n_{00}+n_{10}}\pi^{n_{01}+n_{11}}}{(1-\pi_0)^{n_{00}}{\pi_0}^{n_{01}}(1-\pi_1)^{n_{10}}{\pi_1}^{n_{11}}}\right)$$

Under the null hypothesis, TS $\sim\chi^2_{(2)}$.

4. Data

The risk management measure, VaR is applied on the Romanian market on five individual equities, namely the Investment Funds: SIF1, SIF2, SIF3, SIF4, SIF5 and on an equally weighted portfolio of them (Ristea *et al.*, 2010; Dumitrana *et al.*, 2010).

The choice of these equities is based on the fact that they hold shares on the most important domains in the Romanian market, the interest of the investors in them and in Baciu (2014) has been shown the performance of the generalized hyperbolic distribution in approximating their returns.

Daily returns were gathered for more than five years, starting the day of the maximum closing price in 2007 until the last trading day of 2012. The choice of the time period was due to the economical crisis that affected Romania at the middle of 2007. The period of economical crisis brought changes in the characteristics of each time series.

The time period is divided into an analysis period of 750 days and a test period. The analysis is done using the rolling windows approach, in which at each step is created a window of 750 observations by dropping the oldest return and adding the return of the next day.

Returns are computed as the difference between natural logarithm of current day closing price and natural logarithm of previous day closing price:

$$R_t = \ln(\frac{Y_t}{Y_{t-1}})$$

Table 1 presents the descriptive statistics of the daily returns for the five investigated investment funds. It can be noticed that the returns are skewed and leptokurtic. The hypothesis that data is following a normal distribution is rejected for all equities, as given by the Jarque- Bera test results. All equities have a positive mean return.

Table 1. Descriprive statistics

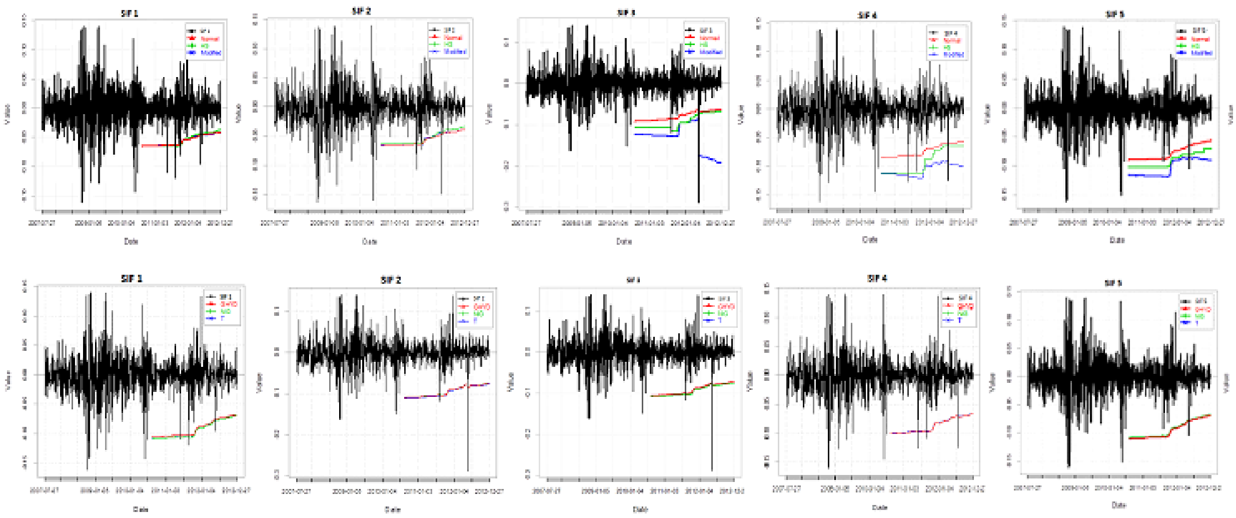
Equity	Sample size	Mean	Standard deviation	Skewness	Kurtosis	Jarque-Bera
SIF 1	1352	0.0946	3.2077	0.1802	4.1727	997,7*
SIF2	1353	0.0752	3.2264	0.2385	4.5521	1191.9*
SIF 3	1372	0.1053	3.3111	0.6811	7.5740	3460.8*
SIF 4	1362	0.0987	2.9963	0.0860	4.9256	1369.4*
SIF 5	1351	0.0924	3.1551	0.1243	4.1248	965.5*

Note: * denotes statistical significance at 5%

5. Results

VaR is computed for each asset for a period of about two years, based on three years of historical data. Several approaches were used to compute the loss of each equity and also of the portfolio: normal VaR, historical, Cornish- Fisher but also based on the distribution of the returns, namely, were used the generalized hyperbolic distribution, Normal- Inverse Gaussian and asymmetric t-Student.

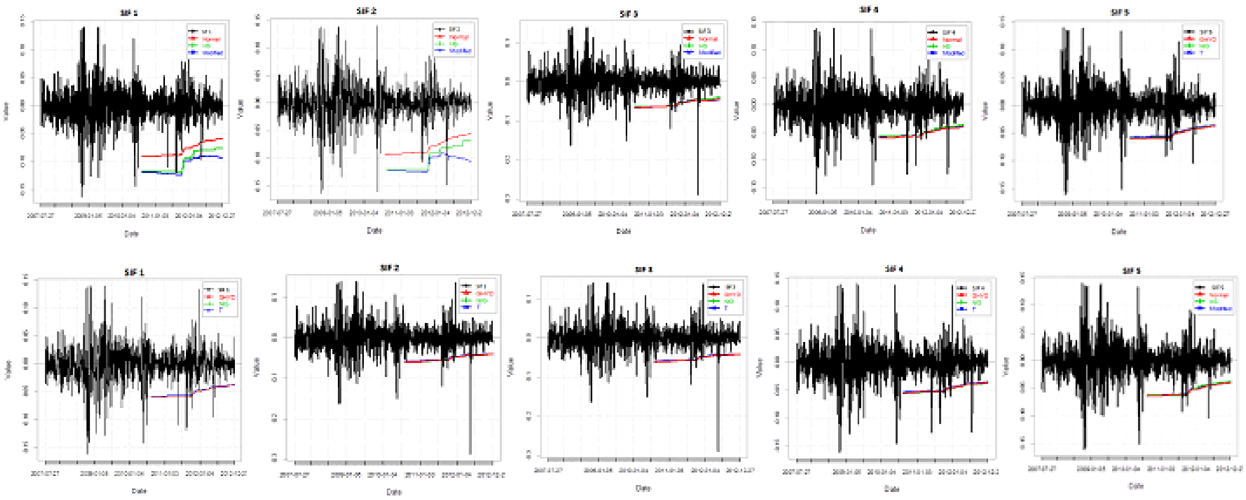
The below plots present estimated VaR values through all considered methods at probability levels of 1% and 5%.



Graph 1. 1% VaR estimates

For SIF1, normal, historical and Cornish- Fisher VaR return similar values at a 1% level of probability while at 5% level of probability normal VaR return smaller looses that the other two methods. When the distribution of the standardized returns is considered, the estimated VaR values are close for all three distributions at the two considered levels of probability.

In the case of SIF2, at 1% probability, all methods return close VaR values while at 5% probability, the Cornish- Fisher method returns the highest looses.



Graph 2. 5% VaR estimates

Estimated VaR for SIF 3 at 1% level of probability returns the smallest values through the Cornish- Fisher method. At 5% level of probability all estimated VaR values are similar. Based on the distribution of the standardized returns, VaR values obtained are similar at 1% and 5% levels of probability.

As in the case of SIF2 and SIF3, for SIF4, estimated VaR values for a 1% level of probability suggest higher looses through Cornish- Fisher method and the smallest ones through the normal method. For a level of probability of 5%, all methods return similar VaR values.

For a 1% level of probability, normal VaR, historical and Cornish- Fisher returns different VaR values for SIF5, while all methods return similar values at a 5% level of probability.

As mentioned by Brown (2008) that" Value-at-Risk is only as good as its backtest. When someone shows me a VaR number, I don't ask how it is computed, I ask to see the backtest" the below table presents the results of the two backtests: Kupiec and Christoffersen. There are also included the expected number of losses greater than VaR based on the chosen probability and the actual number of looses greater than VaR.

Table 2. Backtests results for normal VaR, historical and Cornish- Fisher

Activ	p	Normal				Historical				Cornish-Fisher			
		Expected	Actual	Kupiec p-value	Cristoffersen p-value	Expected	Actual	Kupiec p-value	Cristoffersen p-value	Expected	Actual	Kupiec p-value	Cristoffersen p-value
SIF1	0.05	29	5	0	-	29	5	0	-	29	5	0	-
	0.01	5	3	0.174	-	5	2	0.057	-	5	1	0.011	-
SIF2	0.05	29	4	0	-	29	5	0	-	29	4	0	-
	0.01	5	2	0.057	-	5	1	0.011	-	5	1	0.011	-
SIF3	0.05	29	10	0	0	29	9	0	0	29	10	0	0
	0.01	5	2	0.057	-	5	2	0.057	-	5	1	0.011	-
SIF4	0.05	29	7	0	0	29	7	0	0	29	7	0	0
	0.01	5	3	0.174	-	5	2	0.057	-	5	2	0.057	-
SIF5	0.05	29	8	0	0	29	8	0	0	29	8	0	0
	0.01	5	2	0.057	-	5	1	0.011	-	5	1	0.011	-

For normal VaR, Kupiec backtest rejects the hypothesis that the expected and actual number of looses are equal for a 5% level of probability while at 1% level of probability, Kupiec test fails to reject it. At 1% probability, the loss was overvalued for the equities SIF2 and SIF5 when VaR was computed using the historical method. As it was the case for normal and historical VaR, for Cornish- Fisher VaR, the loss is overvalued at a probability of 5%. At 1% level of probability only for SIF4 the expected number of cases of a greater loss than VaR was well estimated.

Christoffersen backtest does not offer any additional information. In all cases when it was computed, it rejects the hypothesis that the number of days when the loss exceeded VaR was well estimated and these exceedences are independent of each other.

Table 3. Backtests results based on generalized hyperbolic distribution, Normal- Inverse Gaussian and asymmetric t-Student

Activ	p	GHYD				NIG				T			
		Expected	Actual	Kupiec p-value	Cristoffersen p-value	Expected	Actual	Kupiec p-value	Cristoffersen p-value	Expected	Actual	Kupiec p-value	Cristoffersen p-value
SIF1	0.05	29	7	0	0	29	7	0	0	29	8	0	0
	0.01	5	3	0.174	-	5	3	0.174	-	5	3	0.174	-
SIF2	0.05	29	11	0	0	29	10	0	0	29	13	0	0
	0.01	29	2	0	-	29	2	0	-	29	2	0	-
SIF3	0.05	29	10	0	0	29	10	0	0	29	12	0	0
	0.01	5	2	0.057	-	5	2	0.057	-	5	2	0.057	-
SIF4	0.05	29	9	0	0	29	9	0	0	29	10	0	0
	0.01	5	3	0.1742	-	5	3	0.174	-	5	3	0.174	-
SIF5	0.05	21	9	0	0	29	9	0	0	29	9	0	0
	0.01	5	1	0.011	-	5	1	0.011	-	5	1	0.011	-

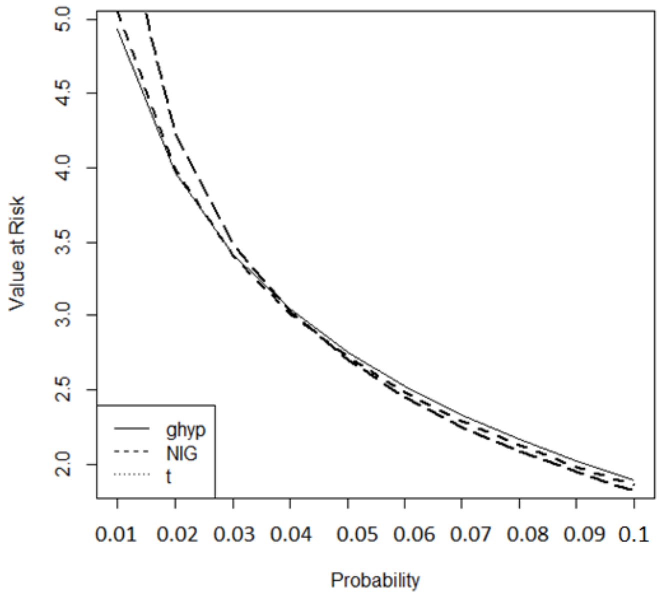
According to Kupiec backtest, VaR estimated based on the distribution of the returns, gives proper estimations of the loss for SIF1, SIF3 and SIF4 at a probability of 1%.

Table 4 presents the results of the backtests for the estimated loss of the portfolio of the five equally weighted equities. The best performance in estimating the loss of the portfolio is of the normal method, for which the loss excess situations are well estimated and are independent at 1%. Also, VaR computed through the historical method or based on the generalized hyperbolic distribution and Normal-Inverse Gaussian is well estimated at 1% probability, while VaR estimated through Cornish-Fisher method or based on asymmetric t-Student distribution overestimates the loss.

Table 4. Backtests for portfolio VaR

Method	P	Expected	Actual	Kupiec p-value	Cristoffersen p-value
Normal	0.05	29	14	0	0
	0.01	5	6	0.99	0.131
Historical	0.05	29	16	0.004	0
	0.01	5	2	0.057	-
Cornish-Fisher	0.05	29	15	0.002	0.001
	0.01	5	1	0.011	-
GHYD	0.05	29	17	0.008	0.007
	0.01	5	2	0.057	-
NIG	0.05	29	16	0.004	0.003
	0.01	5	2	0.057	-
T	0.05	29	18	0.015	0.002
	0.01	5	1	0.01	-

The interest of the investors in a portfolio increases when the loss minimizes. Based on VaR estimations, it can be created a weighted portfolio such that the loss to be minimum. Graph 3 presents the estimated VaR values based on the distribution of the returns at different levels of probability. Table 5 presents the weights of each equity in the portfolio such that the estimated VaR to be minimum at 1% and 5% levels of probability. Such a portfolio is an optimal one because it assumes the minimum risk of loss.



Graph 3. Minimum loss portfolio VaR estimation at different levels of probability

Table 5. Minimum risk portfolio weights

Distribution	p	VaR minimum	Weights SIF1	SIF2	SIF3	SIF4	SIF5
GHYD	0.05	-0.027	0.16	0.18	0.19	0.27	0.20
	0.01	-0.049	0.16	0.17	0.19	0.28	0.20
NIG	0.05	-0.027	0.16	0.18	0.19	0.27	0.2
	0.01	-0.05	0.16	0.17	0.19	0.27	0.21
T	0.05	-0.027	0.16	0.18	0.19	0.27	0.2
	0.01	-0.058	0.17	0.16	0.19	0.27	0.21

6. Conclusions

Value- at- Risk is one of the most used measures of risk management. This measure is widely accepted both in practice and theory because of the benefits it brings compared to the ease of the computation. In this study is compared the performance of six methods used to compute VaR, where three of them take into consideration the distribution of the returns, other than normal distribution. The analysis is performed on five of the most important equities traded at Bucharest Stock Exchange and on their weighted portfolio.

All methods used overestimated the loss by considering more days with higher looses than they actual were. This overreaction can be explained by the fact that the analysis period covered the years 2007-2008, when the market went through moments of severe instability and important looses.

At a 5% probability, the backtests reject all the methods for estimating correctly the number of days with looses higher than VaR. Contrary to the expectation, at 1% level of probability, the normal method works the best, as shown by the backtests, followed by the historical one and the methods based on the generalized hyperbolic distributions. The performance of the normal method under the considered period is because the events that take place in that period are the ones that influence VaR estimation and these years are characterized by frequent extreme values. A distribution like the generalized hyperbolic one, Normal- Inverse Gaussian or asymmetric t- Student catch in its tails these extreme values while the normal one does not and underestimates the loss. This is in the advantage of the normal method because the test period is characterized by less severe looses.

When it was considered the portfolio of the five equally weighted Investment Funds, the best performance was of the historical method, normal and based on generalized hyperbolic distribution and Normal- Inverse Gaussian distribution at 1% level of probability. At a probability of 5%, all methods overestimate risk.

Although the results revealed the performance of the normal distribution in VaR estimation, it brings into attention the sensibility of VaR to the events that take place in the analyzed period, which leads to over or underestimated looses.

The analysis should be extended to a longer period of time and a wider portfolio in order to confirm the performance of the generalized hyperbolic distribution and the weakness of VaR to the chosen time period.

References

1. Baciu, O. A. **Generalized hyperbolic distributions: Empirical evidence on Bucharest Stock Exchange**, working paper, 2014
2. Brauer, C. **Value at Risk Using Hyperbolic Distributions**, Journal of Economics and Business, vol. 52, 2000, pp. 455-467
3. Brown, A. **Counterpoint: Capital inadequacy**, Global Association of Risk Professionals, vol. 42, 2008, pp. 19-26
4. Christoffersen, P. E. **Elements of financial risk management,** Second Edition, Elsevier, 2012
5. Christoffersen, P. E. **Evaluating interval forecasts,** International Economic Review, vol. 39, 1998, pp. 841- 862
6. Dempster, A. P., Laird, N. M. and Rubin, D.B. **Maximum likelihood for incomplete data via the EM algorithm,** Journal of Royal Statistical Society, Series B, vol. 39, no. 1, 1977, pp. 1-38
7. Dumitrana, M., Jianu, I. and Lapteş, R. **Panoptical on the financial statements-from international to national.** Accounting & Management Information Systems/Contabilitate si Informatica de Gestiune, vol. 9, no. 1, 2010, pp. 72-91
8. Einhorn, D. **Private profits and socialized risk,** GARP Risk Review, vol. 42, 2008, pp. 11-18
9. Favre, L. and Galeano, J. A. **Mean-modified Value-at-Risk optimization with hedge funds,** Journal of Alternative Investments, vol. 6, 2002, pp. 21-25
10. Kupiec, P. **Techniques for Verifying the Accuracy of Risk Management Models,** Journal of Derivatives, vol. 3, 1995, pp. 73-84
11. Morgan, J. P. **Risk Metrics- Technical Documente,** Reuters, Fourth Edition, 1996

12. Necula, C. **Modeling heavy-tailed stock index,** Romanian Journal of Economic Forecasting, vol. 6, no. 2, 2009, pp. 118-131
13. Niepolla, O. **Backtesting Value-at-Risk models, Helsinki School of Economics,** Master's Thesis, available at http://epub.lib.aalto.fi/fi/ethesis/pdf/12049/hse_ ethesis_ 12049 .pdf
14. Perignon, C. and Smith, D. **The level and quality of Value-at-Risk disclosure by commercial banks,** Manuscript, Simon Fraser University, 2006
15. Prause, K. **Modelling financial data using generalized hyperbolic distributions,** Preprint 48, Center for data analysis and modeling, University of Freiburg, 1997
16. Prause, K. **The Generalized Hyperboloc Model: estimation, financial derivatives and risk measures,** PhD dissertation, University of Freiburg, 1999
17. Rege, S. and de Menezes, A.G. **Comparing the Generalized Hyperbolic and the Normal Inverse Gaussian Distributions for the daily returns of the PS120,** Working paper series, available at http://www.ceeapla.uac.pt/uploads /pms/attachments/443_paper05-2012.pdf, 2012
18. Ristea, M., Jianu, I. and Jianu, I.**Experienta Romaniei in aplicarea standardelor internationale de raportare financiara si a standardelor internatțonale de contabilitate pentru sectorul public,** Revista Transilvana de Stiinte Administrative, vol. 1, no. 25, 2010, pp. 169-192
19. Socgnia, V. K. and Wilkox, D. **A comparison of Generalized Hyperbolic Distribution models for equity returns,** Journal of Applied Mathematics, 2014, pp. 1-26
20. Vee, D., Gonpot, P. and Sookia, N. **Assessing the performance of generalized autoregressive conditional heteroskedasticity-based value-at-risk models: A case of frontier markets,** Journal of Risk Model Validation, vol. 6, no. 4, 2012, pp. 95-111

KNOWLEDGE BASED ECONOMY IN ROMANIA: COMPARATIVE APPROACH

Sebastian Ion CEPTUREANU
PhD, Assistan Professor, Bucharest University of Economic Studies

E-mail: ceptureanu@yahoo.com

Abstract:
The knowledge-based economy places great importance on the diffusion and use of information and knowledge as well as its creation. The determinants of success of enterprises, and of national economies as a whole, is ever more reliant upon their effectiveness in gathering and utilising knowledge.
This paper is based on 2 different surveys, 4 years apart, on Romanian companies, addressing perception of knowledge based economy by local CEOs or entrepreneurs. It emphasize the changes in perception of this topic and the trends in this matter.

Key words: knowledge based economy; Romanian companies; awareness of KE; transition to KE

1. Introduction

In the contemporary economy, learning and knowledge have become key success factors for both companies and national economies. Competition between firms and countries moved in large part from tangible resources to intangible ones. In terms of the latter, elements such as knowledge and ability to use it (knowledge based economy and knowledge based management) are crucial.

Knowledge becomes the basic resource of companies, the way they get power, prestige and wealth in the economy and modern society. Generation, acquisition and use of knowledge - to name just a few of the transformation of knowledge - are extremely important for sustainable economic, social and cultural development. This trend applies equally to individuals, organizations, institutions, companies, regions or states.

2. Literature survey

The concept of knowledge-based economy and its variants - "knowledge economy", "new economy" or "intangible economy" (Coyle, 1999) - is widely used and increasingly in a variety of contexts and with several meanings. We therefore consider it useful to present some considerations on the use of the term in the literature (Huang & Soete, 2007):

Knowledge-based economy is linked to an extent rooted in what came to be seen as the key role of high-tech industry growth and competitive advantage. It is also due to the increasing application of information and communication technologies and the spread of digital technologies in various different types of activity. So in other words, knowledge-based economy was initially addressed as the sum of high-tech and telecommunications industries. They remain an important component, but now knowledge-based economy is addressed more broadly and is seen as broader than simply overall high-tech and telecommunications industries.

Most of the literature points to the difference between "knowledge" and "information" or between explicit and tacit knowledge (Lundvall & Johnson, 1998). Without denying their importance, we believe that for knowledge based economy both types of knowledge are important, suffering various conversion processes.

In other works (Ordoñez & Serrat, 2009) is the difference between knowledge found in natural products and therefore can be used or applied by others to add value in the production and knowledge built form of human capital.

Process innovation, generation and in particular, application of knowledge to generate new products or services, also occupies a central place in the literature devoted to the knowledge economy. However, more recent work (Sissons, 2011) tends to address broader concept, addressing not only innovation.

Spatial geographic knowledge based economy is also important, authors considering that networks and clusters are vital in generating and sharing various types of knowledge and innovation. This was reflected in such concepts as "regional innovation systems" or "learning regions", found mainly in United Kingdom, as generators of wealth. Other recent works (REKENE, 2011) have emphasized the importance of a wider geographical approach than the regional knowledge based economy. This includes knowledge workers and / or knowledge based activities and the role they can play in driving innovation and economic development at national level. Florida (2002) emphasizes the key role of "social class creative" in generating competitive advantage.

A global knowledge-based economy creates simultaneously significant opportunities and threats for all countries, but especially for those who struggle to combat widespread poverty and create sustainable development, or those who are in transition from the centralized forms of economic organization to democratic forms.

To create these opportunities and face the risks, a country must simultaneously provide three premises (Jones, 2002):

- Set up a coherent, multi-dimensional national strategy, to build and support knowledge-based economy;
- Develop this strategy in a participatory manner, using a broad-based support to include all major sectors of society including the private sector, education, scientists, civil society, media and others;
- Implement a strategy to create knowledge-based economy in a sustained and persistent manner, carefully balancing priorities in the context of increasingly openness to the unpredictable and highly competitive global economy.

There are four essential and interrelated components of any strategy to create a knowledge based economy (Ásgeirsdóttir, 2005):

First: Creating a stimulating economic and institutional environment, to encourage widespread and efficient use of local and global knowledge in all sectors of the economy, fostering entrepreneurial spirit and enabling and supporting economic and social transformations generated by the knowledge revolution;

Two: Creating a society based on qualified, creative and flexible employees (Ceptureanu S., Ceptureanu E., 2010), offering opportunities for quality education and lifelong learning available to all, and a flexible and appropriate public and private funding;

Three: Building a dynamic ICT infrastructure and ICT sector has a competitive and innovative solutions and services to promote information and communication available to the economy and society (Verboncu et al., 2009). These services will include not only "high end" products such as internet and mobile telephony, but also a wide range of communications services and other elements of a developed information society, such as radio, television and other media, computers and other devices for storing, processing and use of information.

Four: Creating an efficient system of innovation including companies, research centers, universities, think tanks (Ceptureanu S., Ceptureanu E., 2010), facilitating access and use the growing stock of global knowledge, adapting it to local needs and using it to create new products and services.

3. Research

We followed two main topics during research: familiarity of the subject with KE and perception of KE.

I. Familiarity of the subjects with KE

According to the first survey, 45.37% of subjects responded that they are familiar with the concept, 35.39% said they were partly familiar and 19.24% have never heard of it. So, overall, the situation is favorable, more than 80% of investigated managers saying that they at least know the concept.

In the latter research, the figures improved even more, with more than 93% of interviewed managers being more or less familiarized with KE.

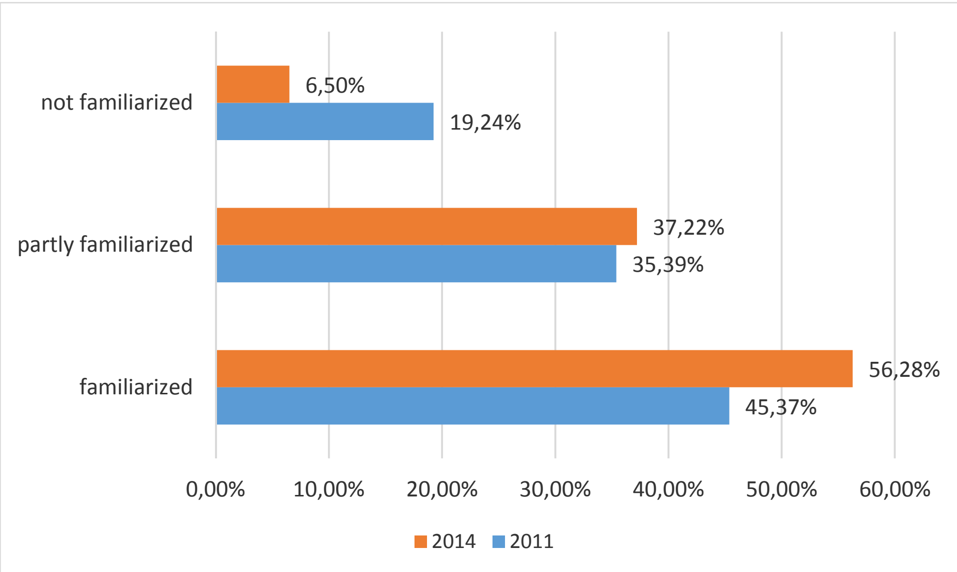


Figure 1. Familiarity with the concept of knowledge-based economy in Romanian companies

Source: own research

Considering **the age of companies**, in the first research we found out that managers of companies established in the last 10 years are more familiar with the concept of knowledge based economy than those of older firms. By category, the percentage is highest among young (62.5%) and very young companies (59.48%), while the proportion of managers who have not heard of the knowledge economy is highest among mature ones (57.63%). This finding is not surprising, in that it was expected that young firms to be more connected and more willing to use opportunities generated by the new economy. The trend was observed in the latter research. The most important change was for mature companies (10 to 15 years old), where the percentage of familiarity with KE increased by 11,63% and partly familiarized by 17,67%. So, older companies become more aware of KE.

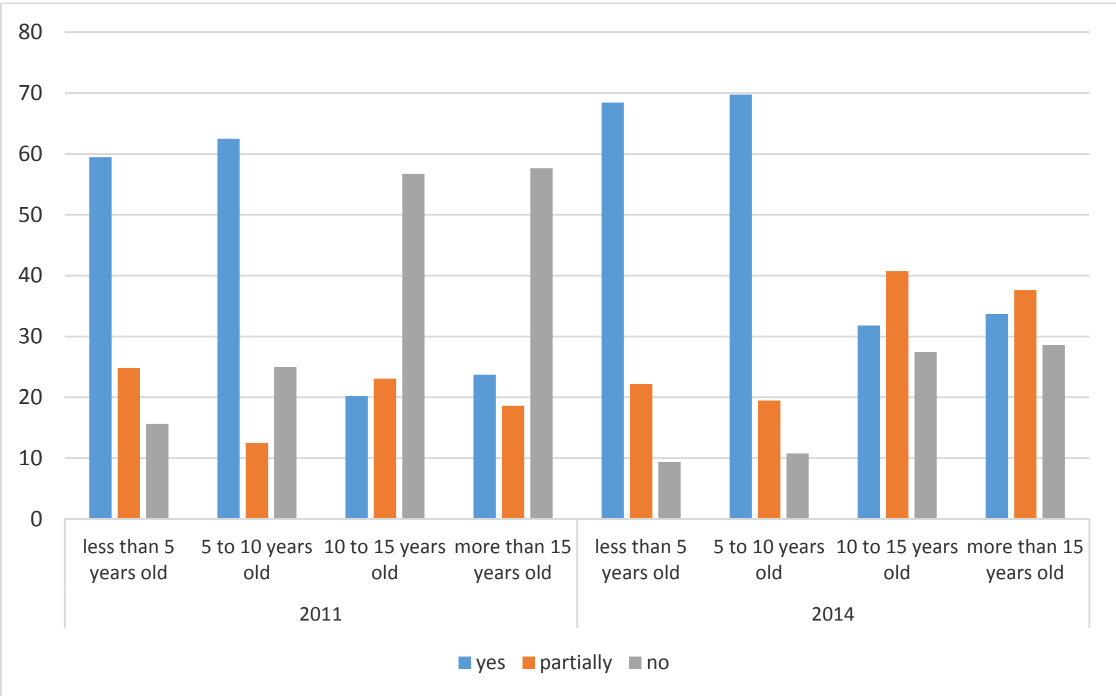


Figure 2. Familiarity with the concept of knowledge-based economy in Romanian companies, by age

Considering Romania's **development regions** we found out during first research that firms localized in Bucharest are more familiar with knowledge-based economy (30.8%), followed by companies form Central Region and the North West Region, and while at the opposite side were companies from North East (15.9%) and South East Regions. The heterogeneity based on geographical situation is still present during the second research. There were regions like Bucharest and South were the percentage of companies familiarized to KE increased, while the percentage decreased in North East and West Regions. However, considering respondents partly familiarized to KE, overall the situation improved, South and South East regions performing best.

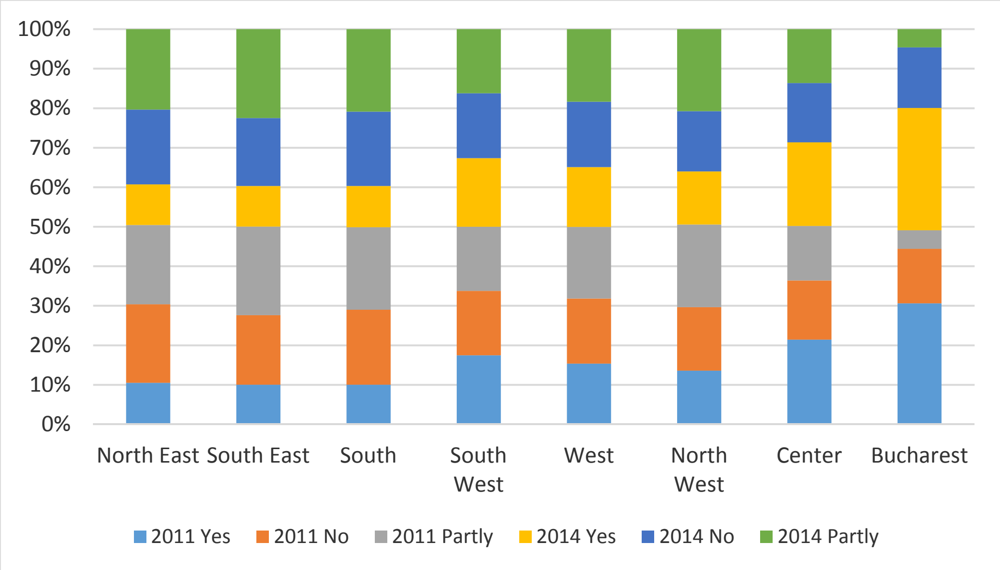


Figure 3. Familiarity with the concept of knowledge-based economy in Romanian companies, by Development Regions

Source: own research

By size of companies investigated, we concluded in the first research that knowledge based economy is known predominantly in small (53%) and medium companies (41.74%), while the most unfavorable situation is among micro companies (40.91% of surveyed managers stated that they did not know the concept). In large companies there is the highest percentage of respondents stating the notion of knowledge-based economy is known in part. These findings are validated by the fact that in knowledge based economy SMEs are advantaged by their organizational flexibility and adaptability, enabling them to cope better their customers' requirements, on the one hand, and the type of business pursued by managers (predominantly service or trade), on the other hand.

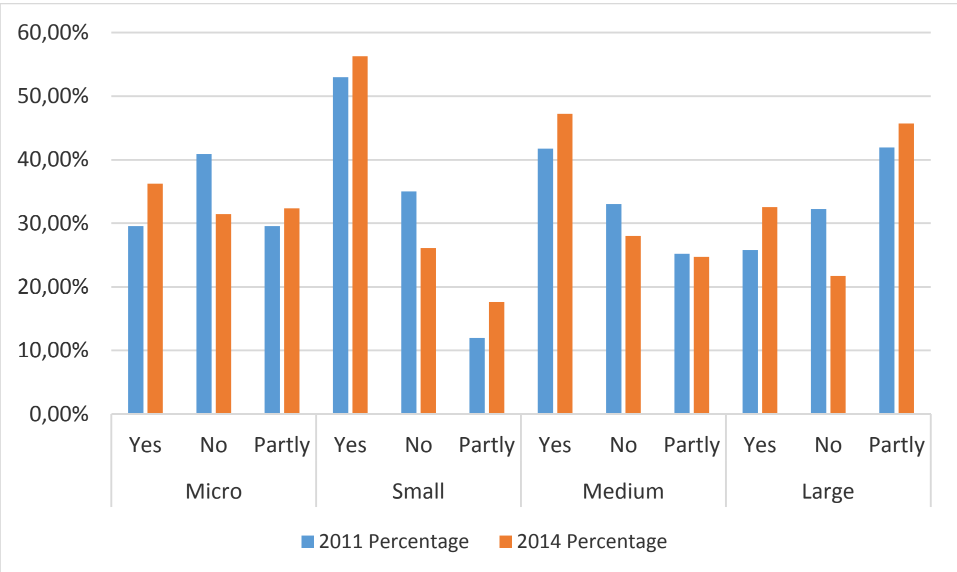


Figure 4. Familiarity with the concept of knowledge-based economy in Romanian companies, by size

Source: own research

In the second research the trend was present, overall the situation improved. The highest percentages in familiarity with KE were registered in large and micro companies, but the concept was still best known in small (56,28%) and medium (47,22%) sized companies.

Considering their **legal form**, in stock companies notion of knowledge-based economy is best known (67.39%), followed by other forms of organization with 43.33% and limited liability companies (40.6%). The situation is similar in the second research, with slightly different percentages among categories. Although it may seem a contradiction considering size criterion, where managers of small and micro firms were most aware of the concept, it is not the case because many small firms in sectors such as services and trade can be organized as stock companies. Weakest in terms of familiarity are limited liability companies with a share almost identical between those who know and have no idea (40.6% and 40.35%).

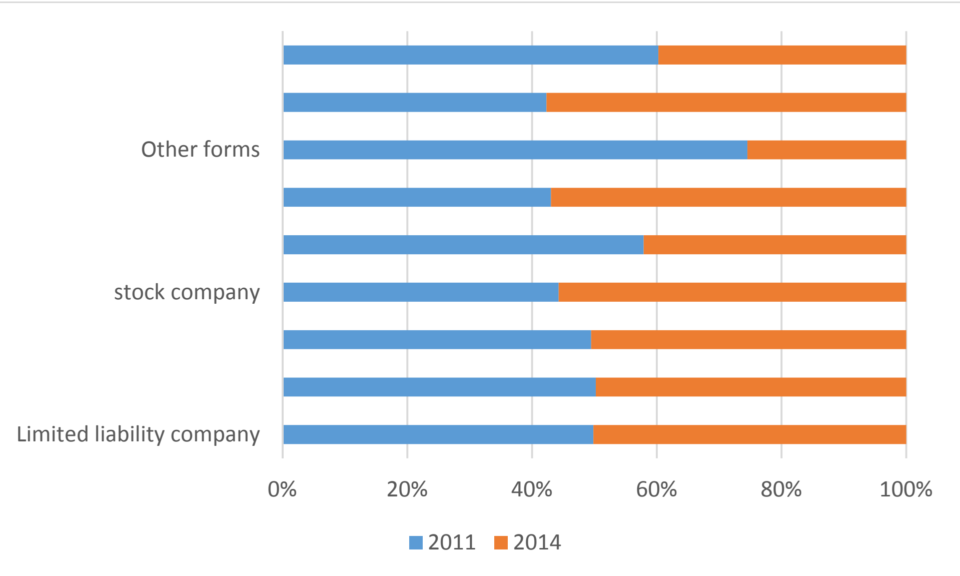


Figure 5. Awareness of knowledge based economy in the investigated companies, by legal status

Source: own research

By industry, knowledge based economy is well known in services and manufacturing, with percentages exceeding 50%, and trade, with almost 50%, while in tourism, construction and especially transport is worst in this respect. Biggest changes occur in manufacturing and transportation four years later (considering partially familiarization), but tourism recorded highest percentage in terms of managers familiarized to KE.

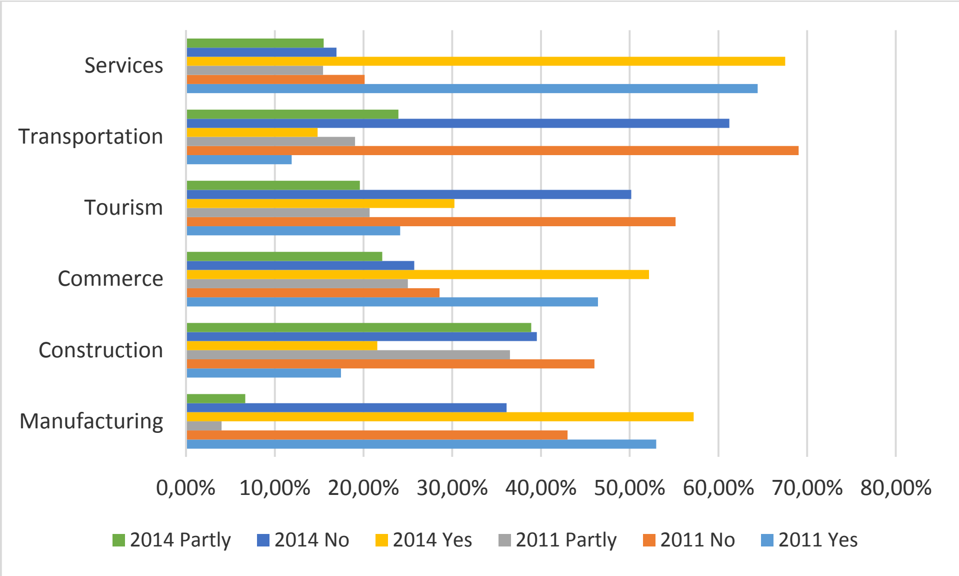


Figure 6. Percentage of knowledge based economy awareness in the investigated companies, by industry

Source: own research

Finally, **by ownership**, surprisingly, state owned companies are more familiar with the concept (50%), higher than private firms (45.27%), but the result is influenced by the share of small firms in the sample state (2.18% of total). Four years later, situation was better for state companies, were more than 60% of investigated managers are familiarized with KE, while in private companies more managers are aware of KE, even though to a lesser extent.

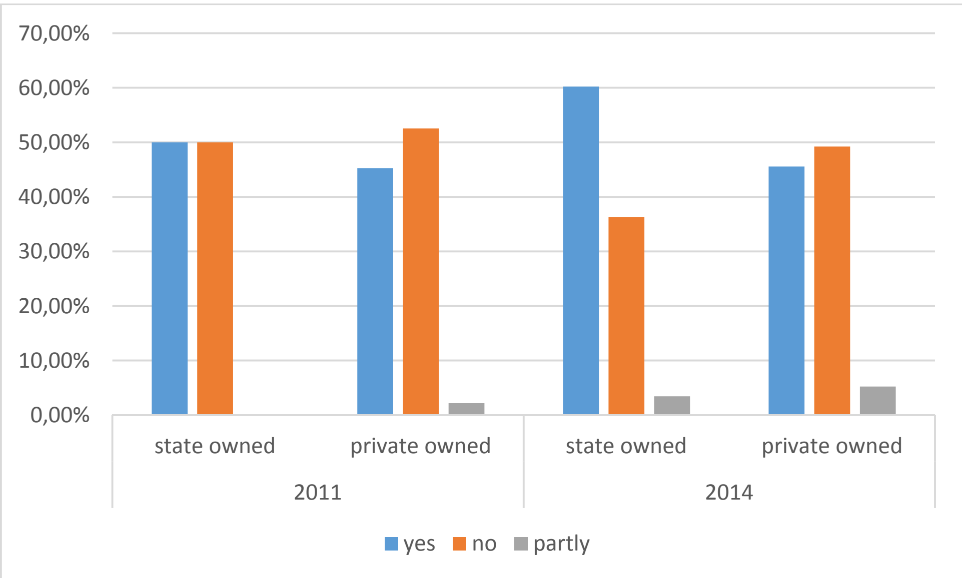


Figure 7. Awareness of the concept of knowledge-based economy in companies investigated, by ownership

Source: own research

II. Perception of Romania's transition to knowledge based economy among Romanian managers

Regarding Romanian companies' managers perception of country's transition to knowledge-based economy, managers of more than half of the companies analyzed (56.44%) believe that this is an opportunity, one third perceive as a threat (35.39%), while 8.17% have a neutral attitude. Four years later, more managers see KE as an opportunity, a clear evidence that the awareness level has increased and the advantages for companies are more evident.

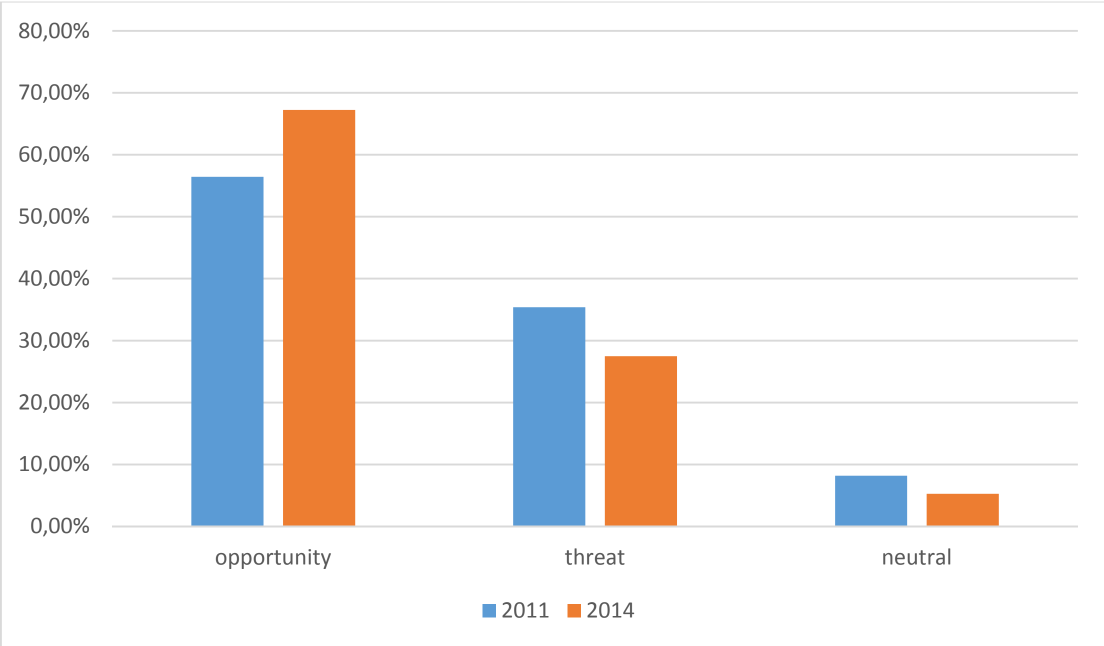


Figure 8. Perception of Romania's transition to knowledge economy
Source: own research

Considering **the age** of surveyed companies, the very young (79.74%) and young companies (65.91%) perceived the transition as an opportunity, while in mature and old firms is the other way around (56.73% for firms from 10 to 15 years old and 57.63% for firms older than 15 years). In the second research, the biggest change in this perception occurred for mature companies (10 to 15 years old), where percentage of those perceiving KE as an opportunity increased with more than 17%. Overall, for old companies KE is still perceived more as a threat, while in the others is seen as an opportunity.

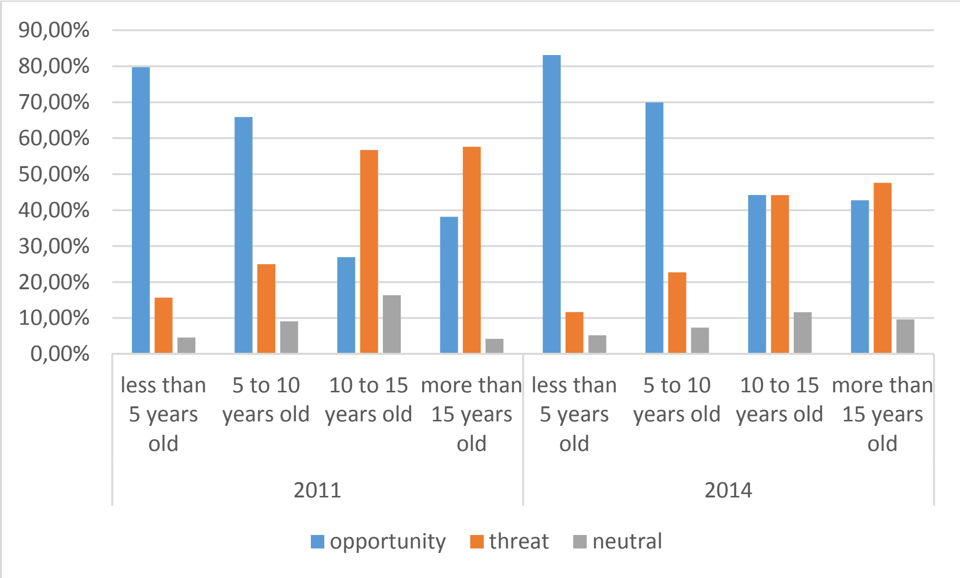


Figure 9. Perception of Romania's transition to knowledge economy, by companies age
Source: own research

Considering **development regions**, the transition of Romania to the knowledge economy is seen as an opportunity for companies in absolutely all developing regions, with higher percentages in the South (87.50%), Southeast (87.23%), West (86.79%) and South Western (86.15%) regions, while higher proportions of distrust were found in Bucharest (26.85%) and North East (20.83%) Regions. The situation remains the same 4 years later, with small scale changes among regions. For instance, in Bucharest and South more companies are willing to capitalize on opportunities of KE (+4,67% and +2,61%, respectively), while companies in West Region are slightly more reluctant (-1,31%).

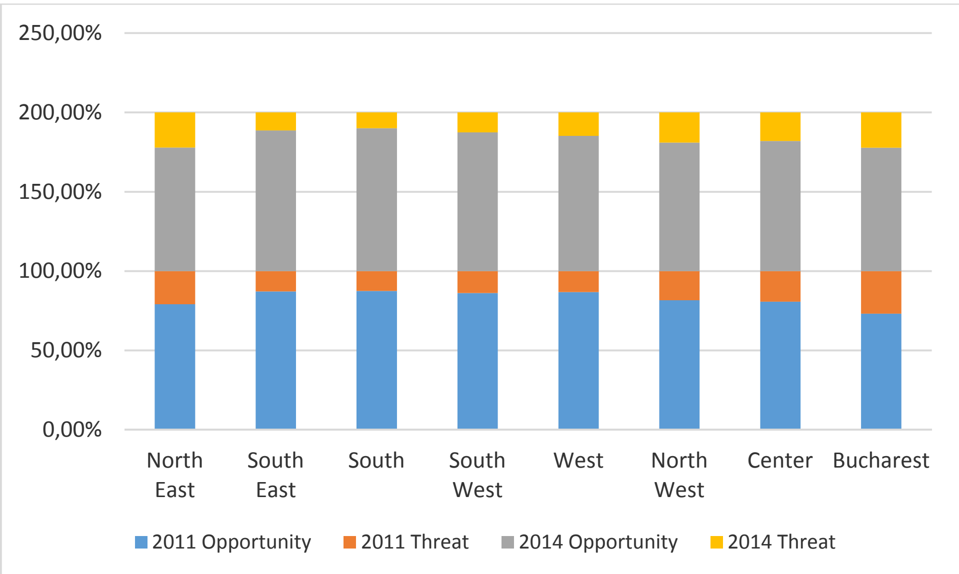


Figure 10. Perception of Romania's transition to knowledge-based economy, on development regions
Source: own research

By **size class**, managers of large companies see the transition as generating opportunities (80.65%), while medium-sized companies are facing the most striking negative attitude (22.61%). However, for the same medium sized firms our research reveals a strong segmentation between the two approaches - opportunity or threat, undecided percentage is below 1%, while for managers of small firms we find the highest percentage of neutral attitude (18.61%). The situation roughly remains the same four years later.

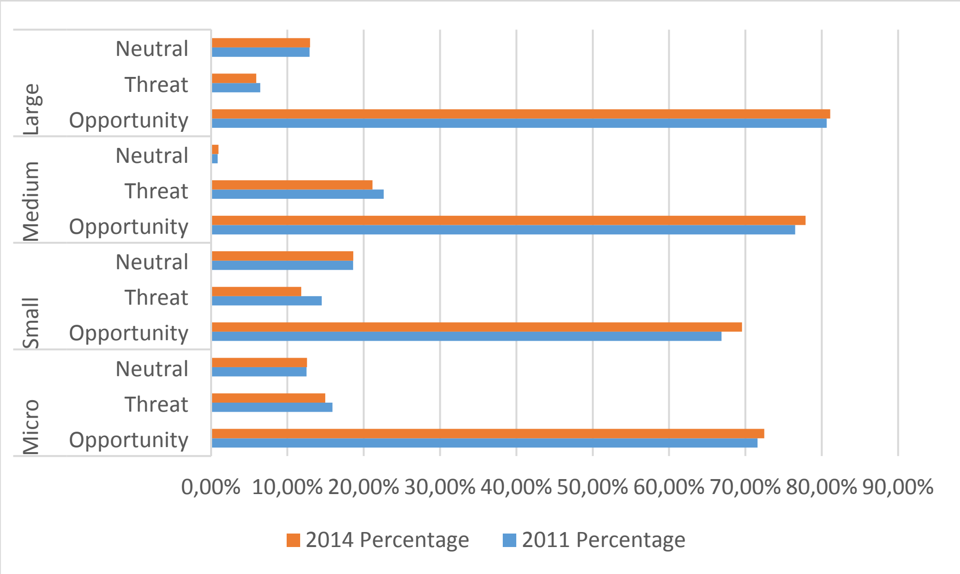


Figure 11. Perception of Romania's transition to knowledge economy, by size of surveyed companies

Source: own research

4. Conclusions

Knowledge based economy is a concept known for most of Romanian companies' managers or entrepreneurs, a trend intensified 4 years later. By the age of the surveyed companies we found that managers of companies established in the last 10 years are more familiar with the concept than those of older firms. In the Bucharest Development Region we found out that firms are most familiar with the concept of knowledge-based economy followed by the North Central Region and Western Region. In the last 4 years, South Region is increasingly catching up.By size, predominantly small (53%) and medium companies (41.74%) are familiar with the concept. By legal form, in joint stock companies' knowledge-based economy is best known, followed by other forms of legal form (GP, partnership, etc.) and limited liability companies.

Considering industry, the concept is best known in services and industry, but in the last years tourism and especially transport are improving their situation. By ownership, state owned companies are more familiar with the concept than private firms. In terms of perception of our country's transition to knowledge-based economy, managers of most of the investigated companies believe that this is an opportunity, and the trend is favorable in this respect. By age, in very young and young companies transition was perceived as an opportunity, but in the last four years mature companies are catching up at an increased pace. By Development regions, our country's transition to knowledge-based economy is seen as an opportunity to absolutely all developing regions. By **size class**, managers of large companies see the transition as generating opportunities while medium-sized companies are facing the most striking negative.

Acknowledgements

This work was cofinanced from the European Social Fund through Sectoral Operational Programme Human Resources Development 2007-2013, project number POSDRU/159/1.5/S/142115 „Performance and excellence in doctoral and postdoctoral research in Romanian economics science domain"

5. References

1. Ásgeirsdóttir, B. **OECD work on knowledge and the knowledge economy,** 2015 Retrieved from: http://www.flacso.edu.mx/openseminar/downloads/ocde_knowledge_speech.pdf
2. Ceptureanu, S. and Ceptureanu, E. **Knowledge creation / conversion process,** International Comparative Management Review, Vol 11. No. 1, 2010, pp. 150-157
3. Coyle, D. **The Weightless World: thriving in the digital age,** London: Capstone, 1999
4. Florida, R. **The Rise of the Creative Class,** Basic Books, New York, 2002
5. Huang, C. and Soete, L. **The global challenges of the knowledge economy: China and the European Union,** Working Paper Series, UNU-MERIT, Maastricht, The Netherlands, 2007
6. Jones, C. **Introduction to Economic Growth, 2nd edition,** New York: W.W.Norton & Co, 2002
7. Lundvall, B.-A. and Johnson, B. **The learning economy,** Journal of Industry Studies, Vol. 1, 1998, pp. 23-43
8. OECD, **Economic Surveys 1997-1998,** Paris: OECD, 1998
9. OECD, **Work on knowledge and the knowledge economy,** 2005, Retrieved from: http://www.flacso.edu.mx/openseminar/downloads/ocde_knowledge_speech.pdf
10. Ordoñez, M. and Serrat, O. **Disseminating Knowledge Products,** 2009, Retrieved from: http://www.adb.org/Documents/Information/Knowledge-Solutions/Disseminating-Knowledge-Products.pdf
11. Sissons, A. **The Knowledge Economy in 2011, the work foundation,** 2011, Retrieved from: http://www.workfoundation.com/Assets/Docs/The%20Knowledge%20Economy%20in%202011_shortened%20version.pdf
12. Verboncu, I. *et al.* **Romanian SMEs and Use of Information Technology,** The 2nd International Conference on Computer Science and Information Technology, Beijing, 2009
13. REKENE, **Regional trajectories to the knowledge economy - Nordic-European comparisons,** 2011, Retrieved from: http://www.regional-studies-assoc.ac.uk/events/2011/sept-ireland/programme.pdf

DISTRIBUTION CENTER OPTIMUM LOCALIZATION AND THE GRAVITATIONAL MODEL

Stefan Virgil IACOB
PhD Candidate, Bucharest University of Economic Studies

E-mail: stefaniacob79@yahoo.com

Abstract:
The present paper approaches the issue of identifying the most suitable position of a distribution point in order to make it attractive and accessible to most "power centers" (sources of potential buyers, donors, etc.), starting from the gravitational model used in Physics. Our study took into account Railly's formula into which an additional variable was introduced, namely, the land price in the area at a certain distance from the power centers. The results present deviations from the calculated distances according to Railly's formula, in the sense that they get closer to the minimum price area.

Key words: gravitational model; attraction; distribution center; distance; land price

1. Introduction

In order to start or develop a business, the current practice is to first take into account the potential of the space area where it is destined to lie by analyzing together the data related to the respective area and the type of commercial activity (the industry which the investment is destined to). Certainly, the analysis must deal with parameters related to population (unemployment, wage levels, etc.), weather conditions, infrastructure and so on. For the purpose of being more efficient we choose to focus on the starting condition, namely, the one related to the space area in which the business is to develop. Throughout this paper we aim to demonstrate that one way to reach optimum results regarding this issue is by appealing the experience demonstrated in Physics.

Following experiments based on bodies being attracted towards the earth surface, Physics demonstrated that all bodies attract each other with a certain force. In the case of small bodies usually used in experiments, the reciprocal attraction based on gravitation is so reduced that very sensitive instruments are necessary. On the contrary, the action of gravitation on big bodies is extremely high especially if one of the bodies is Earth, a particular case of the universal attraction phenomena (Newton 1687).

2. Variables and relationships in attraction process

The gravitational model proves to be a very productive approach in solving certain regional trade, industry and logistics problems. According to it, the interaction intensity between two entities is determined by their dimension or importance as well as by the distance between them. We consider that the dimension of population migration from place *i* to place *j* can be represented as a function of the population in each of the two places and the distance between them. Of course, other factors such as economic development, labor market, problems related to communication in each of the two locations, etc. are considered not to interfere. A general representation of the gravitational model is:

$$I_{ij} = \frac{A \cdot P_j^{\alpha} \cdot P_i^{\beta}}{D_{ij}^{\gamma}} \qquad (1)$$

where: I_{ij} = the size of the interaction between positions *i* and *j*
A = the constant
P = a variable such as population (number of inhabitants) or income
D = distance between *i* and *j*;
α, β, γ = parameters

A different variant of the gravitational model (Railly 1958) includes the same factors, namely, population and distance, having the following formula:

$$\frac{aA}{aB} = \frac{P_A}{P_B} \cdot \left(\frac{D_B}{D_A}\right)^2 \qquad (2)$$

where: aA = attraction to A
aB = attraction to B.
P_A = population of A
P_B = population of B
D_A = distance from the new store to A
D_B = distance from the new store to B

An equal attraction to the store of the population in A and of the B one respectively implies that the ratio on the left side of the equality (2) will be equal to 1, so that:

$$1 = \frac{P_A}{P_B} \cdot \left(\frac{D_B}{D_A}\right)^2 \qquad (3)$$

We make use of the gravitational model by introducing the attraction oriented towards the distribution point (store, retail shop, donation center, meeting place, etc.) at a distance which would make it accessible (attractive) to as more power centers as possible. All throughout the analysis we name *power centers* the sources of people likely to acquire, donate, show themselves, etc. In our following application we quantify attractiveness in relation to distance and to the number of potential clients (population).

The relation Railly put forward approaches attractiveness from the perspective of the distribution unit owner, focusing on his / her interest to gain access to several markets.

In our case, we particularly wish to attract as many clients as possible on two market represented by two urban centers different in terms of the level of population. Regarding this model, we notice that the presence of several variables is obvious, such as each city's attraction, the distance to the distribution point, the city size in terms of population. However, their number has been drastically reduced by introducing the condition of being equally attractive, therefore, in ratio equal to 1. Furthermore, in case the distance between the towns is known, it is sufficient to determine the distance to the distribution point from one of the two localities. Therefore, the existence of only one unknown (noted by x), namely the distance between one of the localities and *the new store*, has lead to reducing the model to only one equation. The fact that attraction is directly proportional with the population of the respective town and inversely proportional with the distance is according to reality.

3. Application

In the first variant of the application we choose to take into account the latter relation (3) in order to solve the concrete problem of the company PROFILO-METAL Prodcom Ltd., a buliding materials store located in Ploiesti. The owners plan to move and extend their activity in an optimum way to other urban areas, especially towards Bucharest. As follows, in such a location the so called *new store* will be situated.

Starting from the company's needs, which were communicated to us by the administration, we shall start our analysis by placing the *new store* somewhere between Ploiesti and Bucharest so that it is attractive to both centers.

As follows, we take into account the statistics of the last census published on *www.recensamantromania.ro*: P_A = 197,522 inhabitants, P_B = 1,677,985 inhabitants; as well as the data recorded on the website *www.distantarutiera.eu* related to the distance between cities A and B: D = 60.8 km. In order to calculate the distance to Ploiesti necessary for determining the optimum location of the *new store*, we note: $D'_A = x$, respectively $D'_B = D - x$.

In equality (3), if we replace $D_A = x$ and $D_B = D - x$, the possiblility to obtain the optimum distance *x* results as follows:

$$x = \frac{D\sqrt{\frac{P_A}{P_B}}}{1+\sqrt{\frac{P_A}{P_B}}} \qquad (4)$$

Therefore, we have: $x = \frac{60.8\sqrt{\frac{197,522}{1,677,985}}}{1+\sqrt{\frac{197,522}{1,677,985}}} = 15.53$ km

Accordingly, the optimum location of the *new store* is 15.53 km from Ploieşti and 45.27 km from Bucharest.

In the same way, we apply the same method to the other important cities close to Ploieşti, namely, Targoviste, Brasov, Buzau si Slobozia. The data are gathered in the table below:

Table 1. Descriptive information

	City	Distance to Ploiesti (km)	Population	Optimum location from Ploiesti (km)
1	Bucharest	60.80	1,677,985	15.53
2	Targoviste	49.40	73,964	30.65
3	Brasov	110.00	227,961	53.03
4	Buzau	74.50	108,384	42.80
5	Slobozia	124.00	43,061	84.53

As a result of the analysis, it can be noticed that the optimum location of the *new store* is closer to the town with less population according as the ratio of the two towns population increases.

4. Land price and adjusted optimal distance

As follows, we aim to extend the analysis by including additional factors. Concretely, we shall introduce the *land price* in the area lying at determined distance (*x*) and the minimum price of the land between the two analyzed localities. This time too, attractiveness is obviously approached from the perspective of the distribution unit owner. The result is a certain distance (*x'*) which is sensitive to an additional important element, price. This can be the land price as such or rent, to which expenses (e.g. transport expenses) can also be added, an increasingly efficient solution being obtained in this way.

Hypotheses:

- Price (in the sense of rent) stands for a variable which increases the closer we get to each of the two towns
- Price can be known, being minimum at a certain distance between the two towns ($D_A^{(p_{\min})}$, $D_B^{(p_{\min})}$) and presenting increase rates calculated as opposed to the minimum price, so that a symmetrical evolution can be noticed as we get closer to each of the two localities in focus

For example, we can identify a minimum price area on the axis Ploiesti – Bucharest, at a distance $D_A^{(p_{\min})}$, not necessarily in central position. Our proposal is to include the variable into the calculus as a coefficient multiplying the determinate distance (*x*) from [4]. The coefficient is represented by the square root of the ratio between the land price situated at distance (*x*) to the minimum price if *x* is smaller than $D_A^{(p_{\min})}$, respectively the square root of such ratio if (*x*) is higher than $D_A^{(p_{\min})}$.

Therefore, the relation has the following form:

$$x' = x\sqrt{\frac{p_x}{p_{min}}}, \text{ if } x < D_A^{(p_{\min})} \tag{5}$$

$$x' = x\sqrt{\frac{1}{\frac{p_x}{p_{min}}}}, \text{ if } x > D_A^{(p_{\min})} \tag{6}$$

$$x' = x, \text{ if } x = D_A^{(p_{\min})} \tag{7}$$

where:

x = distance from [4]

p_x = price in the area at distance *x*;
p_{min} = minimum land price on the axis between the localities in focus;
x' = adjusted determined distance (in the sense of sensitivity to land price)

The results obtained by using formulae (5) – (7) are presented in Table 2.

Table 2. Computation outcomes

	Town	Distance to Ploiesti (km)	Population	Optimal positioning from Ploiesti (km) (x)	Land price at determined distance	Minimum price	Distance between Ploiesti and minimum price area	Adjusted determined distance (km) (x')
1	Bucharest	60,80	1.677.985	15,53	4	3	20	17,93
2	Targoviste	49,40	73.964	30,65	2,5	2,5	30	30,65
3	Brasov	110,00	227.961	53,03	7	6	40	49,10
4	Buzau	74,50	108.384	42,80	3	2	60	52,42
5	Slobozia	124,00	43.061	84,53	1,8	1,5	110	92,60

Adjusted determined distances (*x*') (Table 2) present deviations from the initially calculated distances (*x*) in the sense that they get closer to the minimum price area.

As expected, in the case in which the minimum price on the axis between the two localities in focus is the same with the price in the optimum area determined by calculation, then the adjusted determined distance (*x'*) is the same with the determined optimum distance (*x*). This is the case of the analysis in second position in the table, namely that for the axis Ploiesti-Targoviste.

5. Concluding remarks

Physics has been a continuous source of inspiration for economists in the last centuries and its recent developments open the door to new possible approaches in Economics. The formal side of Physics represents an example for Economics especially with regard to the search for constant values (coefficients) and the attempt to describe phenomena by means of ecquations, including model ellaboration. Taking into account the particularities of each subdomain, economists took over the concepts and laws of Physics in view to analyze economic processes as accurately as possible.

The analysis presented throughout the present article started from the gravitational model introducing attractiveness oriented towards the distribution unit (store, retail shop, donation center, meeting point, etc.) located at such a distance that would make it accessible (attractive) to as many *power centers* as possible. The results of the research point out that the optimum location of the store is closer to the town with less population according as the ratio of the two towns population increases.

Following the introduction of the price variable for the areas analyzed on each of the axes, deviations from the initial results were obtained. The maximum deviation is 22,4% in the case of Ploiesti–Buzau route and the highest deviation axis between the minimum price and the average land price is 50%.

References

1. Brenneke, R. and Schuster, G. **Fizica,** Ed. Didactica si pedagogica, Bucharest, 1973
2. Bulinski, M. **Econofizica si complexitate,** Ed. Univuniversitara, Bucharest, 2007
3. Gheorghiu, A. **Econofizica Investitionala,** Ed. Victor, 2007
4. Gheorghiu, A., Spanulescu, I. **Noi abordari si modele econofizice,** Ed. Economica, 2007
5. Gligor, M. **Perspective ale aplicarii metodelor fizicii statistice in problema optimizarii economice. In volumul statistica in cercetare economico-sociala,** Ed. Junimea, Iasi, 2001
6. Gligor, M. and Ignat, M. **Econofizica,** Ed. Economica, Bucharest, 2003
7. Isaic-Maniu, Al., Mitrut, C. and Voineagu, V. **Statistica,** Ed. Universitara, Bucharest, 2004
8. Paelinck, J. H. and Klassen, L. H. **Spatial Econometrics,** Saxon House, 1979
9. Reif, F. **Fizica statistica (Berkeley vol. V),** Ed. didactica si pedagogica, Bucharest, 1983
10. * * * www.recensamantromania.ro
11. * * * www.distantarutiera.eu
12. * * * www.imobiliare.ro