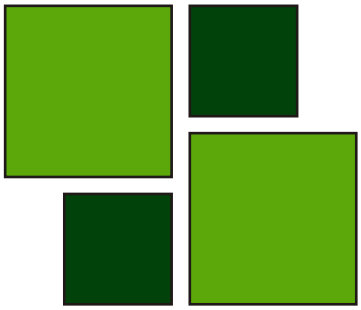

WWW.JAQM.RO

# JOURNAL OF APPLIED QUANTITATIVE METHODS

Quantitative Methods Inquires

Vol. 7
No. 2
Summer
2012

ISSN 1842–4562

# JAQM Editorial Board

## JAQM Advisory Board

JOURNAL OF APPLIED QUANTITATIVE METHODS

# Contents

JOURNAL OF APPLIED QUANTITATIVE METHODS

Quantitative Methods Inquires

# ESTIMATION METHODS FOR THE STRUCTURAL EQUATION MODELS: MAXIMUM LIKELIHOOD, PARTIAL LEAST SQUARES AND GENERALIZED MAXIMUM ENTROPY

**Anna CRISCI**
Second University of Naples

**E-mail:** crisci.anna@virgilio.it

## ABSTRACT

The concept of Latent Variables (LVs or latent constructs) is, probably, one of the most charming and discussed of the last fifty years, although, even today, it is only possible to give a negative definition of it: what is not observable, lacking both of origin and of measurement unit. One of the difficulties for a researcher in the economic-social sciences in the specification of a statistical model describing the casual-effect relationships between the variables derives from the fact that the variables which are object of the analysis are not directly observable (i.e. latent), for example, the performance, the customer satisfaction, the social status etc. Although such latent variables cannot be directly observable, the use of proper indicators (i.e. manifest variables, MVs) can make the measurement of such constructs easy. Thanks to the SEM, it is possible to analyze simultaneously, both the relations of dependence between the LVs (i.e, Structural Model), and the links between the LVs and their indicators, that is, between the corresponding observed variables (i.e, Measurement Model). The different and proper methodologies of estimate of the dependence are topics of this work. In particular, the aim of this work is to analyze Structural Equation Models (SEM) and, in particular, some of the different estimation methods mostly adopted: the Maximum Likelihood-ML, the Partial Least Squares- PLS and the Generalized Maximum Entropy - GME, by illustrating their main differences and similarities.

**Keywords:** Structural Equation Models, Maximum Likelihood, Partial Least Squares, Generalized Maximum Entropy.

## 1. Introduction

The growing availability of the data in the present information- based- society has underlined the need to have at our disposal the proper tools for their analysis. The "data mining" and the applied statistics are suggested as privileged tools to get knowledge from big volumes of data.

In particular, the non- homogenous and extremely complex vision of reality has urged the researchers to make use of techniques of multivariate analysis in order to analyze the relationships existing between more variables, simultaneously.

Among the different methods of multivariate analysis Structural Equation Models-SEM largely satisfy this requirement. The SEM are tools elaborated at the beginning of 1970's, and they obtained, in that decade, a lot of appreciation, and more and more spread use of them. They are the reinterpretation, arrangement and- above all- generalization of those that, in the 1970's, were called casual models and that, in the first half of the same decade, had met a remarkable popularity thanks to the technique of the path analysis.

Thanks to the SEM, it is possible to analyze, simultaneously, both the relations of dependence between the LVs (i.e., Structural Model), and the links between the LVs and their indicators, that is, between the corresponding manifest variables, MVs (i.e., Measurement Model).

The LISREL (Jöreskog, 1970; Jöreskog & Sorbom, 1989; Byrne, Barbara, 2001) or Covariance Structural Analysis (CSA) is at the bottom of such models. The Lisrel was born at the beginning as a name of software and used to estimate the structural parameters of the factorial analysis by adopting the maximum likelihood method. For many years, the Maximum Likelihood method (SEM-ML) has been the only estimation method for SEM, while, today, different estimation techniques can be used for the estimation of the SEM.

In fact, in 1975 Wold developed a *soft* modeling approach, making it different from the *hard* modeling approach of Lisrel, in order to analyze the relationships among different blocks observed variables on the same statistics units.

The method, known as PLS for SEM (SEM-PLS) or as PLS-Path Modeling (PLS-PM), is distribution free, and it was developed as a flexible technique aimed at the casual predictive analysis when the high complexity and the low theoretical information are present.

A new technique for the estimation of the Structural Equation Models has been introduced recently. In 2003 Al Nasser suggested to extend the knowledge of the information theory to the SEM context by means of a new approach called *Generalized Maximum Entropy (SEM-GME)*. This new method is still present in the PLS- approach since no distribution hypothesis is required.

These different and proper methodologies of estimate of the dependence are topics of this work.

The paper is organized as follows: in sections 2 the SEM- Maximum Likelihood is shown; in section 3 and section 4 the SEM-PLS and SEM-GME are shown. Finally, in section 5 a table illustrating the main different/similarities among the three estimation methods is shown.

## 2. The LISREL Approach (SEM- Maximum Likelihood, ML)

As mentioned before, on the basis of Structural Equation Models, the Covariance Structure Analysis (and, thus, LISREL modeling) can be found. The CSA is a " second generation" multivariate technique (Fornell, 1987) combining methodological contributions from two disciplines: the (confirmatory) factor analysis model from psychometric theory and structural equation model typically associated with econometrics. Its aim is to explain the

structure or pattern among a set of latent variables, each measured by one or more manifest and typically fallible indicators.

There are two parts into a covariance structure model (like other approaches analyzed later): the *structural model* and *measurement model***.**

The *structural model* specifies the relationships between the latent variables themselves (reflecting substantive hypothesis based on theoretical consideration). The analysis is predominantly confirmative in nature, that is, it seeks to determine the extent to which the postulated structure is actually consistent with the empirical data at hand. This is carried out by computing the implied covariance matrix by means of the specified model, and comparing it to the (actual) covariance matrix based on the empirical data.

It follows that the first equation of the Lisrel model is:

$$\underset{(mx1)}{\eta} = \underset{(mxm)}{B}\ \underset{(mx1)}{\eta} + \underset{(mxn)}{\Gamma}\ \underset{(nx1)}{\xi} + \underset{(mx1)}{\zeta} \qquad (1)$$

where $\eta\,(eta), \xi(ksi) e\ \zeta\ (zeta)$ are three vectors of the endogenous (variables external to the model which always perform only as independent ) and exogenous (variables internal to the model that at least in a one relation perform as a dependent variable) variables, and errors, respectively. $B\,(beta)$ and $\Gamma\,(gamma)$ are two matrix of structural coefficients between the endogenous variables, and between the exogenous and endogenous variables, respectively. The matrix $B$ has *mxm* element, that is a square matrix whose size is equal to number of the endogenous variables $\eta$; moreover, its diagonal is always composed of all zeros, since they concern with the coefficient regression of each variables with itself. The matrix $\Gamma$, instead, is *mxn order*. In order to be completely specified ,this part of the model needs other two matrices:

1. $\Phi(phi)$ a matrix containing the variances-covariances between the exogenous latent variables $\xi$;
2. $\Psi\,(psi)$ a matrix containing variances- covariances between the errors $\zeta$.

These matrices are squared and symmetric.

The *measurement model* describes how each latent variables is operationalized via the manifest variables, and provides information about the validities and reliabilities of the latter

The *measurement model* for endogenous variables is:

$$\underset{(p,1)}{y} = \underset{(p,m)}{\Lambda_y}\ \underset{(m,1)}{\eta} + \underset{(p,1)}{\varepsilon} \qquad (2)$$

where *y*, $\eta(eta)$ e $\varepsilon\,(epsilon)$ are three vectors of the observed endogenous variables, latent endogenous and errors, respectively. $\Lambda_y$ (lambda *y*) is the matrix of the structural coefficients between the observed variables and the latent variables; this matrix contains *pxm* elements.

The matrix of variance- covariance between errors $\varepsilon$, is indicated with $\theta_\varepsilon$ (theta-epsilon).

This matrix is a squared and symmetric, of *pxp* order (*p* is the number of errors $\varepsilon$, which is equal to that of the observed variables *y*); it must be, in most cases, specified as a

diagonal, that is, the variances of the error are estimated, but the covariance between the errors are set equal to zero.

The *measurement model* for exogenous variables is:

$$x = \Lambda_x \xi + \delta \qquad (3)$$
$$(q,1)\quad (q,n)(n,1)(q,1)$$

where *x*,$\xi$ *e* $\delta$ (delta) are three vectors of the observed exogenous, latent exogenous and errors, respectively. $\Lambda_x$ (*lambda x*) is the matrix of the structural coefficients between the observed variables and latent variables. This matrix contains *qxn* elements.

The matrix of variance-covariance between errors $\delta$, is indicated with $\Theta_\delta$(theta-delta), it is a matrix squared and symmetric, of *qxq* order, and is also specified, in most cases, diagonal.

### 2.1 *The Maximum Likelihood Estimation*

Since the half of the 1960's Maximum Estimation-ML (Jöreskog 1970) has been the predominant estimation method. The ML is an estimation technique referred to Lisrel approach defined *Covariance-Based*, whose objective is to reproduce the covariance matrix of the MVs by means the model parameters. The ML estimation implies that the MVs follow a Multivariate Normal distribution. The analysis is predominantly confirmative in nature, that is, it seeks to determine the extent to which the postulated structure is actually consistent with the empirical data at hand. This is carried out by computing the implied covariance matrix produced by the specified model starting from parameter estimation ($\hat{\Sigma}$), and by comparing it to the (actual) covariance matrix based on the empirical data (S).

Yet, in order to be able to continue, we need to compute the probability of obtaining *S* given $\hat{\Sigma}$. This is possible by means of the so-called Wishart distribution that defines such a probability. In the 1928 Wishart computed the probability function of the distribution *S*, hence called Wishart distribution. The sample covariance matrix *S* with general terms $(s_{ij})$ is :

$$S = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij}-\bar{x}_j)(x_{ij}-\bar{x}_j)' = \frac{1}{n}\sum_{i=1}^{N}(x_{ij}-\bar{x}_j)(x_{ij}-\bar{x}_j)' \qquad (4)$$

where: n=*N*-1, e $\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$. where: $n = N-1$, e $\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$

It can proved (Ghosh and Sinha,2002) that n*S* follows a Wishart distribution as:

$$W(\hat{\Sigma}, n) = \frac{esp\left\{-\frac{n}{2}tr(S\hat{\Sigma})\right\}|nS|^{1/2(n-p-1)}}{|\hat{\Sigma}|^{n/2}2^{np/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma(n+1-i)/2} \qquad (5)$$

If we join all the constant terms (which do not depend on $\hat{\Sigma}$) and we call this combined *T*, the equation 5 it may be re-written as follows:

$$W(S; \hat{\Sigma}, n) = \frac{esp\left\{-\frac{n}{2}tr(S\hat{\Sigma}^{-1}\right\}}{|\hat{\Sigma}|^{n/2}}\, T \qquad (6)$$

A statistical function is developed using the likelihood ratio (Neyman and Pearson, 1928) that compares any theoretical model with a perfect fitting model, that is, the distance between the hypothesized model and theoretically perfect model.

The likelihood ratio is defined by:

$$\text{Likelihood ratio} = \frac{likelihood\ \ for\ \ any\ given model\ under\ H_0}{likelihood\ with\ a\ perfectly\ fitting\ model\ under\ H_1} \qquad (7)$$

$$\text{Likelihood ratio} = \frac{esp\left\{-\frac{n}{2}tr(S\hat{\Sigma}^{-1})\right\}|\hat{\Sigma}|^{-n/2}T}{esp\left\{-\frac{n}{2}tr(SS^{-1})\right\}|S|^{n/2}T} \quad (8)$$

We can observed that $\hat{\Sigma}$ has been replaced by *S* in the denominator of equation (8), because this represent a perfect model.

$$\text{Likelihood ratio} = esp\left\{-\frac{n}{2}tr(S\hat{\Sigma}^{-1})\right\}|\hat{\Sigma}|^{-n/2}\, esp\left\{-\frac{n}{2}tr(SS^{-1})\right\}|S|^{n/2} \quad (9)$$

Taking the natural logarithm of both sides of this equation:

$$\text{log likelihood ratio} = \left\{-\frac{1}{2}tr(S\hat{\Sigma}^{-1})\right\} - \frac{n}{2}\log|\hat{\Sigma}| + \frac{1}{2}tr(SS^{-1}) + \frac{n}{2}log|S|$$

$$= -\frac{n}{2}\left\{tr(S\hat{\Sigma}^{-1})\right\} + \log|\hat{\Sigma}| - tr(SS^{-1}) - log|S| \quad (10)$$

where $SS^{-1}$ is identity matrix .

Therefore, following the ML approach for the estimation of the structural model of covariance- CSA o Lisrel (topic of the present study and that can be considered a generalization of the factorial analysis model inside a more complex and sophisticated system), the discrepancy function to be minimized (by means of derivation operation, that is by calculating the partial derivates compared to the single parameters to be estimated) is:

$$F_{ML} = \text{tr}\,(S\hat{\Sigma}^{-1}) + \ln|\hat{\Sigma}| - \ln|S| - (p+q) \quad (11)$$

where *q* is the number of *X* variables and *p* is the number of *Y* variables.

It is important to underline that, as $\hat{\Sigma}$ converges to S, the inverse $\hat{\Sigma}$ approaches S, and $S\hat{\Sigma}^{-1}$ approaches the identity matrix $SS^{-1}$. Since the identity matrix has the ones on the diagonal, the trace of $S\hat{\Sigma}^{-1}$ will be equal to width of the matrix , i.e a $(p+q)$. The ML function to be minimized is distributed as follows:

$$(N-1)*F_{ML} \sim \chi^2\,[(p+q)(p+q+1)-t] \quad (12)$$

where *t* is the number of free parameters (i.e to be estimated).

## 3. An approach based on Partial Least Squares-Path Modeling (PLS-PM or SEM-PLS)

The PLS Path Modeling is a statistical method which has been developed for the analysis Structural Models with latent variables. As opposed to the covariance-based approach (LISREL), the aim of PLS to obtain the scores of the latent variables for predicted purposes without using the model to explain the covariation of all the indicators. According to Chin (1988), the estimation of the parameters are obtained by basing on the ability of minimizing the residual variances of all dependent variables (both latent and observed).

PLS –Path Modeling aims to estimate the relationships among *M* blocks of variables, which are expression of unobservable constructs. Specifically, PLS-PM estimates the network of relations among the manifest variables and their own latent variables, and the latent variables inside the model through a system of interdependent equations based on simple and multiple regression.

Formally, let us usually assume *P* manifest variables on *N* units. The resulting data $x_{npm}$ are collected in a partitioned table of standardized data ***X*** :

$X = [X_1, \ldots..X_m, \ldots\ldots X_M]$,

where $X_m$ is the generic *m*-th block.

It is important to point out that in SEM literature there is no arrangement on the notation used to define the latent variables and all the other parameters of the model. As the matter of fact, as seem for Lisrel, the exogenous and endogenous latent variables, as the manifest variables and the parameters, are noted in different way, while in the PLS-PM all latent variables are expressed in the same way without considering their role in relation similar the regression. For this purpose, in this study no distinction un terms of notation between exogenous and endogenous construct is made, and all latent variables are defined as $\xi's$.

The path models in the PLS involve three sets of relations:

1. ***Inner Model or Model Structural,*** which refers to the structural model and specifies the relationships between the latent variables LVs. Latent variable can play both predictand role an predictor one; a latent variable which is never predicted is called an exogenous variable, otherwise, it is called endogenous variable. The structural model can be expressed as :

$$\xi_m = \mathrm{B}\,\xi_m + \zeta_m \qquad (13)$$

where $\mathrm{B}$ is the matrix of all the path coefficients in the model. This matrix indicates the structural relationship between LVs. $\zeta_m$ is the inner residual term, and the diagonal variance/covariance matrix among inner terms is indicated with $\Psi$.

2. ***Outer Model or Measurement Model***, which refers to the measurement model and specifies the relationships between the constructs and the associated indicators MVs. Two ways to establish these links can be distinguished as follows:

- ***Reflective way***: in which the indicators (manifest variables) are regarded to be reflections

or manifestations of their latent variables: a variation of the construct yields a variation in the measures. As a result, the direction of causality is from the construct to the indicator. Each manifest variables represents the corresponding latent variable, which is linked to the latent variable by means of a simple regression model. The reflective indicators of the latent construct should be internally consistent, and, as it is assumed that all the measures are indicators equally valid of a latent construct, they are interchangeable. The reflective measures are at the basis of the theory of the classical tests, of the reliability estimation, of and factorial analysis, each of them considers the manifest variable $x_{pm}$ being a linear combination of its latent variable $\xi_m$.

$$x_{pm} = \lambda_{pm}\xi_m + \varepsilon_{pm} \qquad (14)$$

where $\lambda_{pm}$ is the generic loading coefficient associated to the *p*-th manifest variable in the *m* block, and we indicate with the matrix containing all the loading coefficients in the block. $\varepsilon_{pm}$ represents the generic outer residual term associated to the generic manifest variable and the corresponding diagonal variance/ covariance matrix is indicated with $\Theta_\varepsilon$.

- ***Formative way***: in which the indicators are regarded as causes of their latent constructs: a

variation of the measures yields a variation in the construct. As a result, the direction of causality is from the indicator to the construct. The elimination of items that have low correlations compared with the overall indicators will compromise the construct validation, narrowing the domain.

This is one of the reasons by which the reliability measures of the internal consistency should not be used to estimate the fitting of the formative models. Moreover, the multi-collinearity between the indicators may be a serious problem for the parameter estimations of the measurement model when the indicators are formative, but it is a good point when the indicators are reflective. The latent variable $\xi_m$ is assumed to be a linear combination of its manifest variables $x_{pm}$:

$$\xi_m = \sum_p \pi_{pm} x_{pm} + \delta_m \qquad (15)$$

***3. Weight relations,*** the specification of the relations between LVs and their set of indicators is carried out at a conceptual level. In other words, the outer relations refer to the indicator and the "true" LV, which is unknown. As a result, the weight relations must be defined for completeness. The estimation of LVs are defined as follows:

$$\widetilde{\xi_m} = \sum_p \breve{w}_{pm}\ x_{pm} \qquad (16)$$

where, $\breve{w}_{pm}$ are the weights used to estimate the LV as a linear combination of their observed MVs.

In order to estimate the parameter, two double approximations for LVs are considered by PLS algorithm (Wold,1982; Tenenhaus,1999):

- ***the outer approximation or external estimation***, called $v_m$, is used for the measurement model. In this stage we find an initial proxy of each LV, $\xi_m$, as a linear combination of its MVs $x_{pm}$. The external estimation is obtained as the product of the block of MVs and the outer weights $\breve{w}_{pm}$;
- ***the inner approximation or internal estimation***, called $\vartheta_m$, is used for the structural model. The connections among LVs are taken into account in order to get a proxy of each LV worked out as weighted aggregate of its adjacent LVs. The internal estimation is obtained as the product of the external estimation $v_{m'}$ (of $\xi_{m'}$) and the so-called inner weights, $e_{mm'}$.

There are three ways to calculate the internal weights:

**- centroid scheme (Wold)**: II centroid scheme is the scheme of the original algorithm by Wold. This scheme considers only the direction of the sign among the latent variables $e_{mm'} = sign\{cor(v_m , v_{m'})\}$.

**- factorial scheme (Lohmöller)**: this scheme uses the correlation coefficients, $e_{mm'} = \{corr\ (v_m , v_{m'})\}$, as internal weights instead of using only the correlation sign and, therefore, it considers not only the direction of the sign but also the power of link of the paths in the structural model.

- **path weighting scheme**: in this case the latent variables are divided into predictors and followers according to the cause- effect relations between the two latent variables. A variable can be either a follower (if it is yielded by another a latent variable), or a predictor (if it is the cause of another latent variable).

$e_{mm'} =$ *simple/multiple OLS regression coefficient of* $v_m$ *on* $v_{m'}$ *if* $\xi_{m'}$ *is explaratory of* $\xi_m$;

$e_{mm'} = r_{mm'}$ *if* $\xi_m$ *is explaratory of* $\xi_{m'}$;

Once a first estimation of the latent variables is obtained, the algorithm goes on by updating the outer weight. There are two ways to calculate the outer weights:

***1. Mode A***: is preferred when the indicators are linked to their latent variables by means of the reflective way, in which each weight $w_{pm}$ is the coefficient regression of $\vartheta_m$ in the simple regression of $x_{pm}$ on $\vartheta_m$, that is the simple regression $x_{pm} = w_{pm}\vartheta_m$ in which:

$$w_{pm} = (\vartheta_m'\vartheta_m)^{-}\vartheta_m' x_{pm} = cor(x_{pm}, \vartheta_m) \qquad (17)$$

***2. Mode B***: is preferred when the indicators are linked to their latent variables by means of the formative way, in which $\vartheta_m$ is regressed on the block of indicators linked to the latent construct $\xi_m$, and the $w_m$of weight $w_{pm}$is the regression coefficients in the multiple regression:

$\vartheta_m = \sum_p w_{pm} x_{pm}$

and it is defined by means of:

$$\mathrm{w_m} = (X_m' X_m)^{-} X_m' \vartheta_m \qquad (18)$$

The algorithm is iterated till convergence, which is demonstrated to be reached for one and two-block models. However, for multi-block models, convergence is always verified in practice. After convergence, structural (or path) coefficients are estimated through an OLS multiple regression among the estimated latent variable scores.

**3.1 *Summary***

The PLS algorithm works on centered (or standardized) data, and it starts by choosing arbitrary weights (e.g *1,0..0*). Chin (1999) suggested starting with equal weights for all indicators (the loadings are set to 1) to get a first approximation of the LVs as a simple sum of its indicators starts with arbitrary initial weights used to calculate an external approximation of the LVs. The inner relations among LVs are considered to estimate the internal approximation by choosing three options: centroid, factoring and path scheme. After obtaining the internal approximation, the algorithm turns around the external relations with the estimate of outer weights obtained by means of Mode A (*Reflective*) or by Mode B (*Formative)*. The procedure is repeated until convergence of the weights is obtained. Once convergence of the weights is obtained and LVs are estimated, the parameters of the structural and measurement models are calculated by means of the Ordinary Least Squares (OLS).

## 4. An approach based on the Generalized Maximum Entropy or SEM-GME

Golan *et al.* (1996) suggested an alternative method to estimate the parameters for the regression models in case of ill- conditioned problem, as an extension of the measurement of entropy by Shannon and generalization of Maximum Entropy Principle (MEP) by Jaynes. The method is called Generalized Maximum Entropy (GME) and it is based on the re-parameterization and re-formulation of the generalized linear model $y = X\beta + \varepsilon$ with *n* units and *m* variables, in such a way to estimate the parameters by means of the MEP developed by Jaynes, according to the following equation:

$$y_{(n,1)} = X_{(n,m)} \cdot \beta_{(m,1)} + \varepsilon_{(n,1)} = X_{(n,m)} \cdot Z_{(m,m,M)} \cdot p_{(m,M,1)} + V_{(n,n,N)} \cdot w_{(n,N,1)} \qquad (19)$$

It is always possible to write the parameters $\beta_k$ as a convex combination of the variables of finite support variables, in this case five of them in, $\{z_{k1}, z_{k2}, z_{k3}, z_{k4}, z_{k5}\}$ (Paris,2001); this means:

$\beta_k = \{p_{k1} \cdot z_{k1} + p_{k2} \cdot z_{k2} + p_{k3} \cdot z_{k3} + p_{k4} \cdot z_{k4} + p_{k5} \cdot z_{k5}\}$, where the probabilities $0 \leq p_{kj} \leq 1$, $j = 1, \ldots .5$

e $\Sigma p_{kj} = 1$.

Likely, each error term is dealt with as a discrete random variables.

The matrices Z e V are diagonal matrices, whose diagonal elements are vectors of support variables:

$$\beta = Z \cdot p \begin{bmatrix} z_1 & 0 & \ldots & 0 \\ 0 & z_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & z_k \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{bmatrix}; \quad \varepsilon = V \cdot w \begin{bmatrix} v_1 & 0 & \ldots & 0 \\ 0 & v_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & v_T \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ v_T \end{bmatrix} \qquad (20)$$

The support variables are defined by the vectors *z* e *v*, whose dimension, usually from 2 to 7 elements, is identified by the numbers of fixed point *M* and *N*, respectively. The value "c" is a constant distributed in a symmetrical way around zero; in this application c=1 and M and N are equal to 5.

$$z_k = \left[-c \quad -\frac{c}{2} \quad 0 \quad \frac{c}{2} \quad c\right] v_k = \left[-c \quad -\frac{c}{2} \quad 0 \quad \frac{c}{2} \quad c\right] \qquad (21)$$

The vectors *p* and *w* are the probabilities associated with the regression coefficients $\beta$ and with the error terms $\varepsilon$, respectively. The aim is to estimate these probabilities in such a way to represent the coefficients and the error terms as expected value of a discrete random variable.

$$p_k' = [p_{k1} \quad p_{k2} \quad p_{k3} \quad p_{k4} \quad p_{k5}] \; w_k' = [w_{k1} \quad w_{k2} \quad w_{k3} \quad w_{k4} \quad w_{k5}] \qquad (22)$$

The estimation of unknown parameters *p* and *w* is obtained through the maximization of entropy function by Shannon:

$$H(p,w) = -p'_{1,m,M} \cdot \ln pi_{m,M,1} - w'_{1,n} \cdot N \cdot \ln w_{n,N,1} \qquad (23)$$

subjected to consistency and normalization constraints.

The former represent the information generated from the data, this implies a part of the model defined in the equation (17), while the latter identify the following conditions:

$0 \leq p_{kj} \leq 1, \{j = 1, \ldots \ldots, 5; \; k = 1, \ldots . m\}$,

$\Sigma p_{kj} = 1 \{k = 1, \ldots . . m\}$ e $0 \leq w_{kj} 1, \{j = 1, \ldots .5; \; k = 1, \ldots, n\}, \Sigma w_{kj} = 1 \{k = 1, \ldots, n\}$.

### 4.1 *Generalized Maximun Entropy for SEM*

The SEM based on the GME start from the classical formulation by Jöreskog, in which we can distinguish the equations of the structural and measurement models. In particular, this approach considers the re-parameterization of the unknown parameters and of the error terms, as a convex combination of the expected value of a discrete random variable. The equations (1), (2) and (3) can be re-formulated in just one function as:

$$Y_{(p,1)} = \Lambda^y_{(p,m)} \left(I_{(m,m)} - B_{(m,m)}\right)^{-1} \left\{\Gamma_{(m,n)} \Lambda^{x-1}_{(n,q)} \left(X_{(q,1)} - \delta_{(n,1)}\right) + \zeta_{(m,1)}\right\} + \varepsilon_{(p,1)} \qquad (24)$$

where *I* is the identity matrix, $\Lambda^{x-1}$ is the general inverse of $\Lambda^x$. The matrices of the coefficients $B, \Gamma, \Lambda_x, \Lambda_y$ and the matrices of variance-covariance $\Phi, \Psi, \Theta_\varepsilon, \Theta_\delta$, are re-parametrisated as an expected value of the random variables with *M* fixed points for the coefficients *j* and the errors.

GME provides a measurement of the normalized index that quantifies the level of the information generated from the model on the ground of the data collected, providing a

global measurement of the goodness of adaptation of the relation, assumed in the linear model of the simultaneous equations.

The normalized entropy measure can be expressed by means of the following expression:

$$S(\tilde{p}) = \frac{-\mathbf{p}' \cdot \ln \mathbf{p}}{K \ln M} \qquad (25)$$

This index of normalization is a measure of the reduction of the uncertain information, in which $-\mathbf{p}'\ln\mathbf{p}$ is the entropy function by Shannon, while $K$ the number of the predictors and $M$ is the number of fixed points (FPs). The quantity $K \cdot \ln M$ represents the maximum uncertainty. If $S(\tilde{p})=0$ there is n suitable to explain the data. The maximum entropy measure can be used as a method to select the explicative variables.

## 5. LISREL, PLS-Path Modeling e GME: differences and similarities

This section shows the main differences and similarities among the three estimation methods analyzed in the previous sections.

In particular, we have started out from the table suggested by Sanchez (2009) for Lisrel and PLS and, by keeping the same style by Sanchez, we have re-elaborated ( in the past authors, such as Al-Nasser and Ciavolino Enrico, developed some aspects of GME) the same features for GME.

**Table 1**: LISREL, PLS-Path Modeling e GME: differences and similarities

| | **Lisrel (Covariance Structure Analysis)** | **PLS Path Modeling** | **GME** |
|---|---|---|---|
| **Object** | ***Parameter Oriented***: objective is to reproduce the covariance matrix of the MVs by means the model parameters. | ***Description-Prediction Oriented***: obtain the scores of the latent variables for predicted purposes without using the model to explain the covariation of all the indicators | ***Estimation precision-prediction oriented***: maximize the "objective function = ***Shannon's entropy function",*** emphasizing both estimation precision and prediction |
| **Approach** | ***Covariance-based***: the residual covariances are minimized for optimal parameter accuracy. | ***Variance-based***: aims at explaining variances of dependent variables (observed and unobserved) in regression sense (i.e residual variances are minimized to enhance optimal predictive power). | ***Theoretic Informantion- based***: under Jaynes' maximum entropy (uncertainty) principle, out of all those distribution consistent with the data-evidence we choose the one that maximizes the entropy function and thus maximizes the missing information, in order to get models based on real data.. |
| **Optimality** | If the hypothesized model is correct in the sense of explaining the covariations of all indicator, CSA provides optimal estimates of the parameters (i.e offers statistical precision in the context of stringent assumptions). | PLS trades parameter efficiency for prediction accuracy, simplicity, and fewer assumptions. | - the GME provides the estimation in case of negative freedom degrees;<br>-uses all the information in the data;<br>- is robust relative to the underlying data generation process and to the limited-incomplete of economic data;<br>-performs well relative to competing estimators under a squared error measure performance; |

| | | | |
|---|---|---|---|
| **Type of fitting algorithm** | Simultaneous estimation of parameters by minimizing discrepancies between observed and predicted Covariance/correlation matrix. **Full information method** | Muti-stage iterative procedure using OLS. Subset of parameters estimated separately **A limited information method**. | The estimation of the parameters is obtained by the maximization of the Shannon's entropy function subject a consistency and normalization constraints. **Full information method.** |
| **Conception** | Used more as an auxiliary tool for theory testing. | Used more as a decision making tool, with emphasis on parsimonious prediction. | Used as a tool to solve problems called ill-conditioned, where the lack of information and / or specific data about the problem at hand requires the recruitment of general assumptions as possible with respect to the parameters of the system under study. |
| **LV scores** | Indeterminate. Indirect estimation computed with the whole set of MVs. | LVs explicitly estimated as linear combination of their indicators. | Each a LV is re-parameterized as a convex combination of a discrete random variable. |
| **Relationships between the LVs and MVs** | Typically only with reflective indicators. | Reflective and formative indicators. | Reflective and formative indicators. |
| **Treatment of measurament residuals** | Combines specific variance and measurament error into a single estimate. | Separates out irrelevant variance from the structural portion of the model. | The variance/covariance matrix $\Psi, \Theta_{\varepsilon}, \Theta_{\delta,}$ are re-parametrization as a expected value of a discrete random variable. |
| **Manifest Variables** | Continuous and interval scaling | Continuous , interval scaling, categorical. | Continuous , interval scaling, categorical. |
| **Assumed distributions** | Multivariate normal if estimation through Maximum Likelihood. | No distribution assumpttions | Semi-parametric |
| **Sample size** | High<br>>200 unit | Medium<br>40<unit< 200 | Low<br>10<unità<40 |
| **Model correctness** | To the extent that the theoretical model is correct it is able to explained the covariations of all indicators. | To the extent that the theoretical model is correct it is determined partly from the power of the relations of path between the LVs. | To the extent that the theoretical model is correct it is determined by the chance to obtain a set of consistent relations based on data. |
| **Consistency of stimators** | Consistent, given correctness of model and appropriateness of assumptions. | Bias estimators tend to manifest in higher loading coefficients and low path coefficients. The bias is reduced when both the size and the number of indicators for the LVs increase. (***consistency at large***). | Consistent and asymptotically normal under four mild conditions:<br>1. The error support spans a uniform and symmetrical around zero;<br>2. The parameter support space contains the true realization of the unknown parameters;<br>3. The errors are independently and identically distributed;<br>4. The design matrix is of full rank (Golan 2003:5). |

| **Missing Value** | | Nipals algthm | Maximum Likelihood method |
|---|---|---|---|
| **Evaluation Model** | Evaluation Model by means hypothesis testing: ***Chi-square:*** the H0 hypothesis is: S- $\hat{\Sigma} = 0$ | -$R^2$ for dependent LVs; -GoF(*Amato et.al 2004*) - resampling (jackknifing and bootstrapping) to examine the stability of estimation. | -Normalized index of entropy that quantified the level of information generated from the model on the bases of the collected data. -Pseudo $R^2$. |
| **Applicability** | • The phenomena analyzed are clear; • Low complexity of the model; • Presumes the use of reflective indicators; • Usually stringent assumptions about the distribution, independence, large sample size; • Treatment of hierarchical data, multi-group; • Comparison of models which come from different populations with a single objective function. | • Relatively new phenomena or mutant; • Relatively complex model with a large number of indicators and / or latent variables; • Epistemological need to model the relationship between LVs and indicators in different ways (formative and reflective); • Hypothesis normality, independence and the sample size is not met; • Multi-group. | • Complex model with incomplete data and small sample size; • Use both reflective and formative indicators; • It is easier to impose non – linear constraints; • Does not require distributional hypothesis; • Multi-group, hierarchical data; • Ability to insert a priori information on the model. |

## 6. Conclusion

This work has had the purpose of illustrating the structural equation models and, in particular, the three estimation methods mostly used in the econometric applications, showing the main differences and similarities of them.

The Covariance Structure Analysis (and, thus, Lisrel model) approach belonging to **Covariance-based approach .** The aim of Covariance-based techniques is to reproduce the sample covariance matrix by the model parameters. In other words, model coefficients are estimated in such a way to reproduce the sample covariance matrix. In the covariance based approach, the measurement model is typically considered as reflective, the multivariate normal must be respected if estimation is carried out by means of the ML and works on large sample.

The PLS approach is, instead, **Variance-based**, i.e strongly prevision oriented, whose aim is to obtain the scores of the latent variables for predicted purposes without using the model to explain the covariation of all the indicators. According to Chin (1988), the estimates of the parameters are obtained by basing on the ability of minimizing the residual variances of all dependent variables (both latent and observed). The PLS does not require items which follow a multivariate normal distribution and adopts both formative and reflective indicators and works on small samples properly.

Finally, the GME approach is **Theoretical Information-based** whose aim is to maximize the entropy function and, thus, maximizes the missing information, in order to

obtain model based on eal data. This estimate technique remains in the optic the PLS approach since it does not require any distribution assumption (Ciavolino & Al-Nasser, 2006 demonstrated that the GME approach for SEM seems to work better than the PLS-PM when outliers are present.

## Reference

1. Al-Nasser, A. D., **Customer Satisfaction Measurement Models: Generalized Maximum Entropy Approach**. *Pak Journal of Statistics*, 19(2), 2003 213–226.

2. Amato, S., Esposito Vinzi, V., and Tenenhaus, M., **A global Goodness-of-Fit for PLS structural equation modelling. Oral communication to PLS Club,** HEC School of Management, France, March 24, 2004.

3. Byrne, Barbara M., **"Structural Equation Modeling with AMOS,EQS, and LISREL: Comparative Approaches to testing for Factory Validity of a Measuring Instrument"** *International Journal of Testing*, 2001, p.55-86.

5. Bagozzi, R.P, **"Structural Equation Models in Experimental Research"**, *Journal of Marketing Resarch*,14, 1977, pp 209-226.

6. Ciavolino, E. and Al-Nasser, A., **Comparing generalized maximum entropy and partial least squares methods for structural equation models**. *Journal of Non parametric Statistic* Vol. 21, No.8, 2009, 1017-1036.

7. Ciavolino, E., Al Nasser, A.D., and D'Ambra, A., ***The Generalized Maximum Entropy Estimation method for the Structural Equation Models***, GFKL *2006*, Berlino, marzo 2006.

8. Chin, W.W., and Newsted, P.R., **Structural Equation Modeling Analysis with Small Samples using Partial Least Squares**. In: *Statistical Strategies for Small Sample Research*, 1999, 307-341. Hoyle R. (Ed). London: Sage Publications.

9. Chin, W.W, **Partial Least Squares for research: An overview and presentation of recent advance using the PLS approach**, 2000, p.68.

10. Edwards, Jeffrey R. and Bagozzi, P.R, **"On the Nature and Direction of Relationships between Construct an Measures"**, *Psycological Method,*5(2), 2000, 155-74.

11. Esposito Vinzi, V., Chin, W.W., Henseler,J., and Wang. H, **Handbook of Partial Least Squares, Concepts, Methods and Application**, Springer Handbooks of Computational Statistics, New York, 2010.

12. Fornell, C., **"A Second Generation of Multivariate Analysis: Classification of Method and Implications for Marketing Research"**. In: *Review of Marketing*. (Ed) Houston, M.J. (Chicago), American Marketing Association, 1987, pp. 407-450

13. Fornell, C., and Bookstein, F.L., **Two Structural equation model: LISREL and PLS applied to consumer exit-voice theory.** *Journal of Marketing Research*,19, 1982, 440-452.

14. Golan, Amos and Judge, G. and Miller, D., ***The Maximum Entropy Approach to Estimation and Inference*: An Overview**, Staff General Research Papers 1327, Iowa State University, Department of Economics, 1997.

15. Golan, A., Judge G., and Perloff J., **Estimation and inference with censored and ordered multinomial response data.** In: *Journal of Econometrics*, 79 (1997), 23-51.

16. Golan A, Judge G, Miller D, ***Maximum entropy econometrics: robust estimation with limited data***. Wiley, New York, 1996.

17. Jaynes, E.T, **Information Theory and Statistical Mechanics.** *The Physical Review* 106(4), 620-630, May 15,1957.

18. Jöreskog, K.G., ***Factor Analysis and Its Extensions. Factor Analysis at 100: Historical Developments and Future Decisions***. Department of Psychology, University of North Carolina. May 13-15, 2004.

19. Jöreskog, K.G., and Wold, H., **The ML and PLS Techniques for Modeling with Latent Variables: Historical and Comparative Aspects**. In: *Systems under indirect observation: Causality, structure, prediction*. Part I, 1982, 263-270. K.G.

20. Jöreskog, K.G., **"Testing Structural Equation Models".** In: *Testing Structural Equation Models*. (Ed) Bollen, K.A and Lang, J.S., Beverly Hills, Sage, 1993, pp. 294-316.

21. Jöreskog, K.G and Sorbon, D., ***LISREL 7: A Guide to the Program and Application***, Chicago, IL, SPSS, Inc., 1989

22. Jöreskog, K.G and Sorbon, D., ***LISREL 7: User's Reference Guide*.** In: Scientific Software, Inc., Mooresville, 1989.

23. Tenenhaus, M., ***L'Approche PLS***, *Statistquèe Applique* XLVII(2), 1999, 5-40.

24. Tenenhaus, M., ***Structural Equation Modeling for small samples***. HEC school of management (GRECHEC), 2008.

25. Tenenhaus, M., Esposito Vinzi, V., Chatelin,Y.M, and Lauro,C., **PLS Path Modeling**, *Computational Statistics & Data Analysis* 48, 2005, pp.159-205.

26. Wishart, J., **The Generalised Product Moment Distribution in Sample from Normal Multivariate Population**. *Biometrika,* Vol 20 A, No ½, Jul,1928, pp 32-52.

27. Wold, H., **Soft modeling: The Basic Design and Some Extensions**. In: *Systems under indirect observation: Causality, structure, prediction*. Part II, 1-54. K.G. Jöreskog and H. Wold (Eds). Amsterdam: North Holland, 1982.

28. Wold, H., ***Non Linear estimation by iterative least squares procedures*.** In: Researche Paper Statistics, Festchrft for J. Neyman, 411-444. F.David (Ed). New York: Wiley, 1966.

29. Wold, H., ***Estimation of Principal Components an Related Models by Iterative Least Squares***, In: Multivariate Analysis, Proceedings o fan International Symposium, 391-430, P.R. Krishmaiah (Ed). New York: Academic Press, 1966.

31. Sanchez, G., **Partial Least Squares**, Phd thesis, 2009.

32. Shannon, C.E., **A mathematical theory of communication**. *Bell system Technical Journal* 27, 1948, 379-423.

# EVALUATING SECURITY THREATS IN DISTRIBUTED APPLICATIONS LIFESTAGES

**Catalin Alexandru TANASIE** [1]

PhD Candidate,
Doctoral School of Economical Informatics
The Bucharest University of Economic Studies, Romania

Application Designer,
Raiffeisen Bank S.A., Bucharest

**E-mail:** catalin_tanasie@yahoo.com

**Eduard Emanuel HERTELIU** [2]

PhD Candidate,
Doctoral School of Economical Informatics
The Bucharest University of Economic Studies, Romania

**E-mail:** emanuel.herteliu@gmail.com

**Abstract:**

*The article starts with the classification of security threats as related to the context of operating distributed IT&C applications – DIAs, as concerning users, processes and the information exchanged. Security risks induced as part of the analysis, design and development phases of distributed application building are detailed alongside proposed countermeasures. A model addressing security element interaction is presented and details on its implementation as part of the authentication module of the model testing and refining application, MERICS, is shown.*

**Key words:** *distributed applications, information, security, risk, models, metrics.*

## 1. Classifying DIA security risks

DIA security relates to the interaction of actors, data messages, operations and contextual parameters in ensuring the privacy and operability of the system's informational content, operations and contextual parameters. The following constitute elements of risk in the context of the system's security:

- *information*, the content operated on by computing instruments in the processes that characterize operational DIA modules, alongside the output obtained from the underlying methods; risk sources include the loss of privacy for transferred or stored data, unauthorized acquisition of application and context of operation parameters and authentication credentials;

- *users*, which determine the quality and content of messaging and operations in DIA activities, impacting risks by the uncontrollable nature of their actions as viewed in the application's context – no security protocol is able to prevent the unauthorized disclosure of credentials; measures to prevent or minimize damage, include the updating of security tokens to levels which are difficult or impossible to replicate or know without context-dependent input – dynamic passwords, certificates that rely on third parties or cryptographic instruments; in situations where complex authentication is impractical – public information portals, virtual libraries, news networks – the prevention of incidents relies on auditing and preparing a set of runtime automatic assessment and threat prevention procedures – logging out suspect users, shutting down or refusing the initiating of new sessions past a predetermined threshold;
- *administrative processes* – relating to the set of maintenance tasks and actions done inside or outside the operating context of the application – assignment of user authentication credentials, monitoring of processes and communication channels, ensuring the operational status for the application's hardware and software platforms, logging and treating errors, managing and prioritizing tasks, interacting with databases and file systems in ensuring the proper functioning of querying and persistence-related functions; security risks relate to the implicit potential of damage resulting from the preferential operational and informational access that accompanies administrative roles and tasks;
- *communication* – the generating, transfer and reading of information through messaging tasks, over public and private networks; the encoding, encryption and decryption capabilities of the communicating parties determine the security potential of the packages of transferred data; risk sources include the number and operating context of communication channels, data encryption capabilities, asynchronous operation features, messaging or transfer paradigms – immediate or synchronous, queue-enabled as in Service-Oriented Architecture; the relevancy of the latter is due to the *time*-dependent nature of cryptographic tools – few, if any, of these are immune to dictionary-based attacks or exhaustive key searches over extended periods of time in the context of an exponential increase in processing power over the past decades, their security deriving from being able to prevent information leaks within relevant operational time frames – an attacker who deciphers dynamic, runtime-generated encryption, financial content months after the message was sent is not able to use this information in impersonating communication parties.

In considering the measures that the DIA interactions actors, involved in the design, development and usage of distributed applications, need to take in order to generate an accurate model for security threat pattern detection and identify targeted and improperly constructed components, the structuring of vulnerabilities and associated risks is required. The following constitute security risk classification criteria in assessing DIA vulnerability levels:

- *context of appearance*, with risks originating inside or outside the construction and usage domain of the system; when assessing a Web service, the incidents originating in database or file system information disclosure constitute *internal* causes for the associated costs, directly traceable to the improper development of cryptographic instruments; security errors due to loss or mismanagement of user credentials are

determined by both the improper design of crisis procedures and techniques and *external* factors including deployment and security platforms owner and maintenance crew;

- the *effect* on the application domain, with *operational* risks relating to incidents that target components, communication channels or repositories by preventing their proper functioning, either through the direct interaction and alteration of parameters, or indirectly by exploiting vulnerabilities in their design – brute force attacks, preventing synchronous messaging in interdependent components; the second subcategory is formed by *informational* risks, relating to the altering or unauthorized accessing of data structures, allowing the attacker gains by exploiting the discovered parameters and confidential content;
- *frequency and damage*, which together form the measure of costs to the operator and users of the system; in the generic economic context of a limited number of resources and infinite number of needs, the latter are organized and prioritized so that a maximum level of satisfaction is obtained; translating to risks, the losses due to the occurrence of unwanted operations are minimized by implementing supplementary controls and procedures.

Actors involved in the design and development of the distributed application, shown as they interact in *Figure 1*, contribute as factors to the constriction and evaluation of risk assessment models. *Table 1* shows the origin and relevant DIA lifecycle phase for each influencing operation, along with associated risk and nature of the model's input and/or results by description of countermeasures.

**Table 1**. Analysis, design and development actor-induced security risks

| Actor | Phase | Risk | Countermeasures |
|---|---|---|---|
| **User** | Pre-analysis | improper evaluation of security threat potential | assessing the potential losses by identifying and valuing operational and informational damage |
| **User** | Analysis | incomplete specification or knowledge of required activities and information sensitivity | security-oriented valuation of DIA content and operations by analysts |
| **Analyst** | Analysis | inaccurate understanding of security tasks as specified by the operational user, especially concerning large, interlinked activity sets | formalizing and documenting the requirements and acquiring cross-market information on relevant threats |
| **Analyst** | Design | improper specification of security constraints | inclusion of development and design parties in the evaluation of requirements and techniques |
| **Designer** | Design | technological choice limitations due to security costs | assessing risk frequency and potential damage in distributing security controls and tools between DIA components |
| **Designer** | Design | over extensive, interdependent security technique specifications | evaluation of cross-component impact of security protocol choices |
| **Developer** | Development | improper implementation of security controls | assessing threat levels for each DIA component type and communication channel |
| **Developer** | Development | insufficient auditing tools for operations and data structures | analyzing incident target area and supplementing information change monitoring |

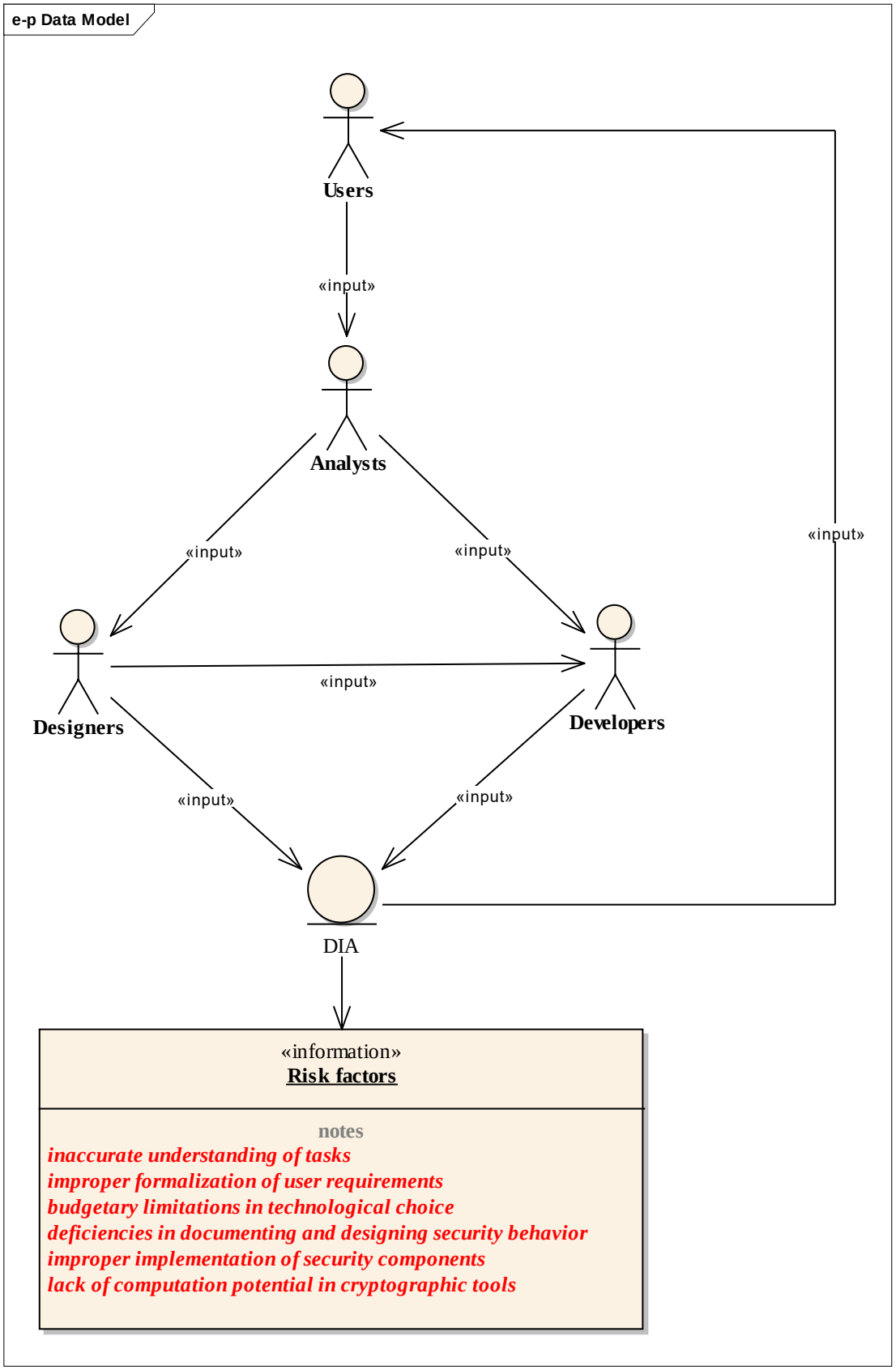


**Figure 1**. Analysis, design and development actors as risk factors

The influence of security incidents on the costs of maintaining and running DIAs is directly proportional to the exposure of the system's endpoints, dimensioning of sessions and user base, as well as the nature of operation. Sensitive processes and messaging require the addition of supplementary steps as part of the operational methods. The variance in distributed application span and coverage, as well as areas of usage, requires the establishment of universally valuable, cross-industry measuring units of cost. MERICS, the distributed software system used in factor analysis and risk assessment model evaluation, is designed with auditing controls that measure *timing* and *computing* resource strain. The first element is defined as the relation between the number of hours that users and developers, as well as automatic processes, require in order to prevent or remove effects as opposed to global indicators of value for the distributed application's owning organization.

## 2. Information security risk elements

Securing access to information as part of inter-component communication and persistence-related operations requires the formalization of interactions within DIA activities, as well as the development of procedural mechanisms that manage and audit authentication jobs.

Let $UP = \{U_1 \dots, U_E, \dots, U_R, \dots, U_n, P_1, \dots, P_E, \dots, P_R, \dots, P_m\}$ of $n$ users and $m$ processes, represented for simplicity purposes as a set of entities belonging to the two categories, $\{e_1, \dots e_{n+m}\}$, requiring authentication by means ranging from the simple providing of a password to dynamic, context-dependent token information and cryptographic operation-enabled credentials such as digital certificates.

Let the function $pass_j(e)$ describe the status of validity for user of process $e$ with regard to feature $j$ in the authentication criteria, as follows:

$$pass(e) = \begin{cases} 0, & authentication\ factor\ not\ present\ or\ complied \\ 1, & authentication\ factor\ present\ or\ complied \end{cases}$$

The granting of access to the presentation and service layers of the distributed application is described by function $acc()$ described as a product of $s$ conditions as follows:

$$acc(e_i) = \prod_{j=1}^{s} pass_j(e_i), \qquad i = \overline{1, n+m}, \qquad j = \overline{1, s}$$

where

- $e_i$ – element $i$ in set UP; authentication requester – user or process;
- $j$ – number of conditions that the authenticating entity must fulfill in order to be granted access to the functions of the distributed application;
- $s$ – Number of requirements applied to the authenticating entities.

Considering the previously presented format, the possible values of the $acc()$ function describe the same range as function $pass()$:

$$acc(e_i) = \begin{cases} 0, & one\ or\ more\ \ authentication\ conditons\ are\ not\ fulfilled\ by\ e_i \\ 1, & authentication\ factor\ present\ or\ complied\ by\ e_i. \end{cases}$$

The logic of authentication operations is described as a repetitive block of the form shown in *figure 2*. The activity diagram defines steps in evaluating entities as part of the MERICS.AUTHENTICATION module, with regard to factors as presented in *table 2*.

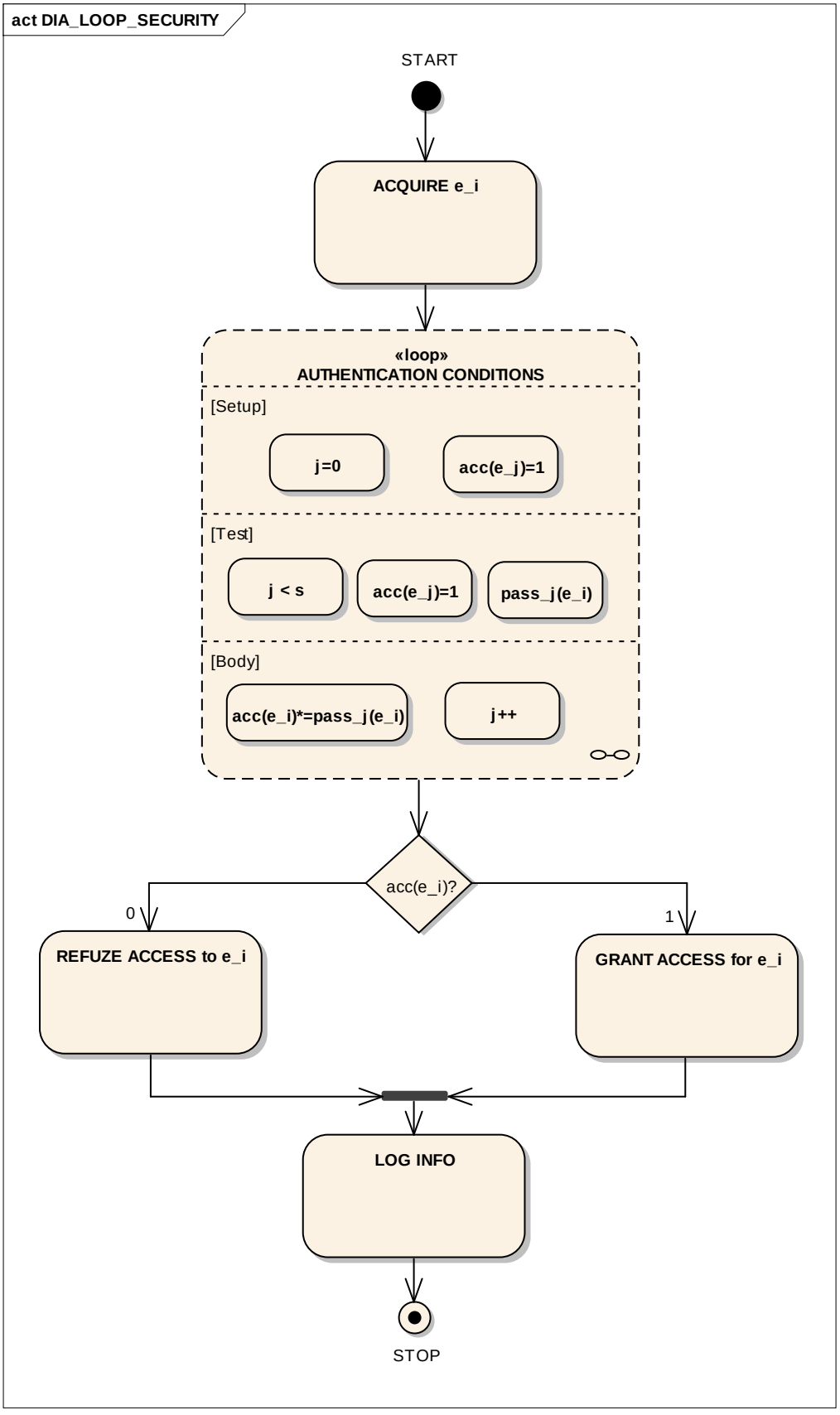


**Figure 2.** MERICS.AUTHENTICATION

**Table 2.** Authentication criteria as part of *MERICS.AUTHENTICATION*

| Criterion | Description | Test order |
|---|---|---|
| Credentials | validation of the presented security credentials by comparison to an internally administered user credential list | 1 |
| Black list | compare the specified credentials to a predefined list defining reported sources of incidents, dynamically updated based on observed behavior and effects of DIA interactions by the users and processes | 2 |
| Groups | the user group that the currently evaluated user belongs to | 3 |
| Roles | the roles that define permitted operations and component access for the assessed entity | 4 |
| Operations | the requested action as compared to allowed interactions | 5 |
| Time | the time of the request, used to determine the relevancy of the request | 6 |
| Location | the network location of the requester | 7 |
| History | derived from the blacklist, time and location criteria, the history of access by the evaluated entity is used to determine the authenticity of the request | 8 |

The following two cases constitute a hypothetical scenario for the *MERICS.WEBAPP* module communicating with MERICS.WCF and evaluated by *MERICS.AUTHENTICATION* – *table 3*.

**Table 3**. Evaluation as part of *MERICS.AUTHENTICATION*

| Source | Target | Operation | $acc(e_i)$[1-8] | Resolution |
|---|---|---|---|---|
| **MERICS .OPERATIONAL** | ***MERICS .DataOperations*** | ***query*** | 1*1*1*1*1*1*1*1 | **1:ALLOW** |
| ***MERICS.WEBAPP*** | **MERICS.WCF** | ***query*** | 1*1*1*1*1*1*1*1 | **1:ALLOW** |
| ***MERICS.WEBAPP*** | **MERICS.WCF** | ***delete*** | 1*1*1*1*1*1*1*1 | **1:ALLOW** |
| ***MERICS .LOGICAL*** | **MERICS.WCF** | ***query*** | 1*1*1*1*1*1*1*1 | **1:ALLOW** |
| ***MERICS .LOGICAL*** | **MERICS.WCF** | ***delete*** | [1*1*1*1***0**]*1*1*0 | **0:DENY** |
| ***MERICS.WEBAPP*** | ***MERICS .DataOperations*** | ***query*** | [1*1*1***0**]*0*1*1***0** | **0:DENY** |

An additional source of vulnerabilities consists of the transfer of information between components as part of DIA activities. The specific separation of varying activities and role-based or geographical separation increase the incidence of cross-component communication as compared to other software application paradigms. *Figure 3* details on the buildup of risk augmenting factors, starting with the first stages of an application's lifecycle, as a graph detailing on dependencies as defined, in order of precedence, by:

- *requirements*, influencing the activity domain of DIA interactions, as well as the user roles that are defined for their management; identifying the vulnerabilities early on reduces costs; the lack of operational information in development environments tests and the restructuring of development tasks to account for the issues allow for

the validation and improvement of the application's components before incidents occur;

- *users*, whose number is a defining characteristic in the definition of operational computation and storage requirements, as well as frameworks chosen and extent of

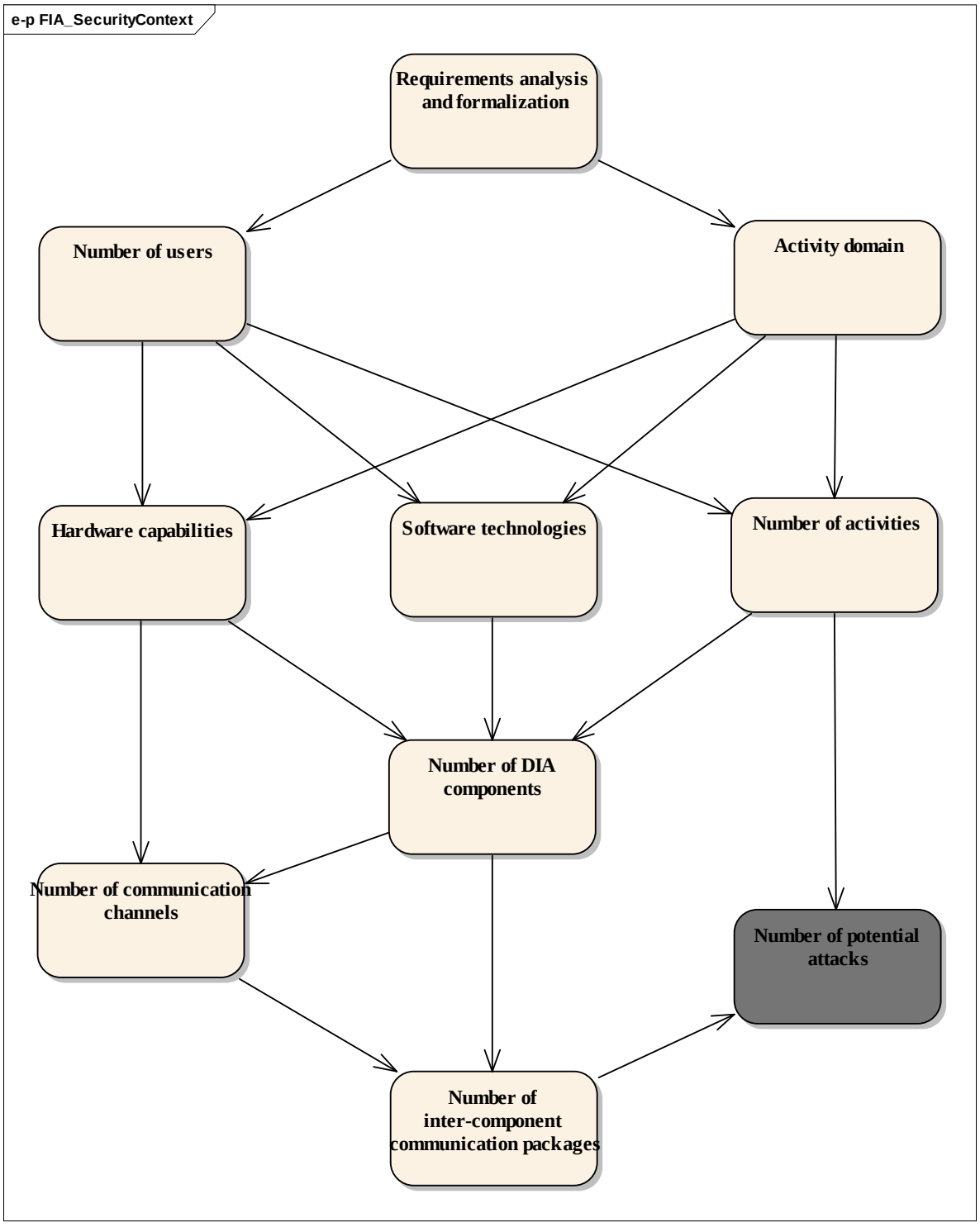


**Figure 3.** Information transfer threat potential factors

activities intermediated by the application; the nature of user activities identifies areas of increased risk incidence, based on the gained informational content – in financial transactions, user and account credentials;

- *hardware capabilities*, central to the identification of risks as the extent of resources determines the recovery times and error resistance; the budgetary constraints of the developers and users influence the future design and implementation of the components, as well as cryptographic tools in communication and authentication procedures;
- *software technologies*, deriving from the hardware capacity, as the efficiency and span of framework-implemented activities is dependent on the available hardware and projected activities; the extent of choices is determined by the deployment platform, communication capacity and environment owner as compared to the application owner;
- *component numbers*, directly controlling tolerance to incidents by allowing for activity autonomy and explicit redundancy as tools against overextended dependencies in components; there occurs a bidirectional effect on security, with increased component numbers allowing for better protection, yet increasing the occurrence probabilities for complexity-derived incidents;
- *channel numbers*, deriving from the number of components and influencing the risk susceptibility by exposing information and operational or analytical module endpoints to unauthorized interactions; the number of potential attacks depends on the number of inter-component messages, as it increases on chances to detect and impersonate authorized processes.

The criteria enumerated above, alongside factors ordered in Table 2, form the basis for the global assessor of security compliance, $SC$ in the distributed application, by evaluating the effects and origination of incidents. The $n$ factors in set $X = \{x_1, x_2, \dots, x_n\}$ are quantified by averaging their impact and including relative weight information, $w, w \in [0,1]$, as relating to a predefined system of measurement, where the comparison basis is formed by the optimum or total item number for the measured factor, represented by set $\overline{X} = \{\overline{x_1}, \overline{x_2}, \dots, \overline{x_n}\}$:

$$SC = \frac{\frac{\overline{x_1} - x_1}{\overline{x_1}} * (1 - w_1) + \frac{\overline{x_2} - x_2}{\overline{x_2}} * (1 - w_2) + \dots + \frac{\overline{x_n} - x_n}{\overline{x_n}} * (1 - w_n)}{w_1 + w_2 + \dots + w_n},$$

or in generic form as

$$SC = \frac{\sum_{i=1}^{n} \left( \frac{\overline{x_i} - x_i}{\overline{x_i}} * w_i \right)}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} \left[ \left( 1 - \frac{x_i}{\overline{x_i}} \right) * (1 - w_i) \right]}{\sum_{i=1}^{n} w_i},$$

where

- $x_i$ – factor *i* failure counter or costs;
- $\overline{x_i}$ – total number of factor I items or value;
- $w_i$ – relative weight associated to factor $i, i = \overline{1, n}$.

In Table 2, the 8 levels of application impact, starting with credential discovery or loss and ending with subtle variations in distributed application behavior observed at various moments infer on the severity of the loss. For the two observed incidents in Table 3,

considering compatible systems and units of measurement, the weights reflect the order of the impacted factor's performance as follows – the first one, impacting MERICS.LOGICAL, is a fifth level risk source, and therefore is assigned a weight of 5/8 or 0.625. The second one, in WEBAPP and level 4, is slightly lower in impact and therefore corresponds to weight 4/8 = 0.5. MERICS contains only one instance of MERICS.LOGICAL, yet two autonomous, self replacing interfaces. Therefore, the global security compliance over the period covering these two incidents is calculated as

$$SC = \frac{\frac{1-1}{1} * (1 - 0{,}625) + \frac{2-1}{2} * (1 - 0{,}5)}{0{,}625 + 0{,}5} = 0{,}22$$

The 22% security compliance reflects the impending need for updating the targeted components, especially the logical module, whose failure is augmented by its uniqueness.

## 3. Metrics validation

Validation is a process which assures that a result complies with the expectations and the desired standards. While developing distributed computer applications it is necessary to find ways for quantifying the level of compliance. Numeric values are assigned to specific features of the informatics products. This is done using metrics. In measuring, validation tells weather the usage of a metric will lead to a satisfactory result or not. Both validation of the model and validation of the result are done.

Each model has its own metrics. Validation of the model is done by verifying the properties for each of the model's metrics. Generally a metric has the following representation [7]:

$y = f(X)$ ,

with $X = \{x_1, x_2, \ldots, x_n\}$ the set of influence factors for the informatics product characteristic which is being measured, as detailed in the previous section.

The properties to be verified in the model's validation process are:

- metric's sensitivity
- the non-catastrophic character
- the non-compensatory character

The model is sensitive when the variation of the influence factors produces the variation of the measured value: $y$. For a small variation of the influence factors values the variation of the resulted value is small and for a big variation of the influence factors values the variation of the resulted value is big[8]. A model $y = f(x)$ is sensitive when the following is true:

$$f(x + h) = f(x) + f(h)$$

The model has a catastrophic character when there are values of the influence factors for which the measured value $y$ is impossible to calculate. This is the situation when dividing a number to a value which tends to be zero.

The model has a compensatory character when for different values of the influence factors the result is the same. It is important in developing computer programs that for different input the output is different as well. The above model has a non-compensatory character when the next condition is true:

$$f(x_1) \neq f(x_2) \; \forall \, x_1 \neq x_2$$

Let $TD$ be a set of MERICS testing data used to validate t Security Compliance *SC* indicator.

$TD = \{td_1, td_2, \dots, td_{ntd}\}$

Where:

*ntd* – the number of test data sets used;

$td_1, td_2, \dots, td_{ntd}$ – the test data sets;

$pv_{ij}$ – the property value for the test data set $td_i$.

For validating the model properties [7], the Table 4 is populated with values corresponding to each property of a test data set: sensitivity, non-catastrophic character, non-compensatory character. At the intersection of the line *i* with the column *j* the value of $pv_{ij}$ is 1 when the *SC* property is verified. If the SC property is not verified the value of $pv_i$ is 0.

**Table 4** – Indicator properties validating [7]

| Test Data Set\Indicator Property | Sensitive | Non-Catastrophic | Non-Compensatory |
|---|---|---|---|
| $td_1$ | $pv_{11}$ | $pv_{12}$ | $pv_{13}$ |
| $td_2$ | $pv_{21}$ | $pv_{22}$ | $pv_{23}$ |
| ... | ... | ... | ... |
| $td_i$ | $pv_{i1}$ | $pv_{i2}$ | $pv_{i3}$ |
| ... | ... | ... | ... |
| $td_{ntd}$ | $pv_{ntd1}$ | $pv_{ntd2}$ | $pv_{ntd3}$ |
| | $TPV_1$ | $TPV_2$ | $TPV_3$ |

Where:

$td_i$ – the test data set *i* used as input in MERICS application for validating the *SC* indicator;

*ntd* – the total number of test data sets;

$pv_{ij}$ – the propriety value as 0 or 1 indicating whether the property is verified or not;

$TPV_i$ – total property value used to express the level of property verification by aggregation of $pv_{ij}$.

Knowing the above *ntd* and $pv_{ij}$, the aggregated property value, $TPV_j$ is given by:

$$TPV_j = \frac{\sum_{i=1}^{ntd} pv_{ij}}{ntd}, j=1..3$$

Knowing the above $TPV_j$, the indicator $I_{SC}$ is used to validate the security compliance model and is given by:

$$I_{SC} = \frac{\sum_{j=1}^{3} 0.33 * TPV_j}{ntd}$$

The value of $I_{SC}$ gives the validation of *SC* as following:

- If the value of $I_{SC} < 0{,}78$ the *SC* indicator is not validated
- If the value of $I_{SC} \geq 0{,}78$ *and* $I_{SC} < 0{,}92$ the *SC* indicator is validated as *good*
- If the value of $I_{SC} \geq 0{,}92$ the *SC* indicator is validated as *very good*

The model is being refined using MERICS and it is to be verified with every version of the informatics application.

## Conclusions

Developers and end-users communication-context threat awareness is required in order to provide the mechanisms of defense – disaster recovery documentation and procedural detailing, increased security in vulnerable sections, identified through the evaluation using automated modeling by analytical modules. The development and refining of risk assessment models and associated metrics enables the owners or developers of complex software applications to measure, quantify risk and evaluated individual and global behavior through successive stages of the application's lifecycle and associated versions of the assemblies and software structures that form the implemented content.

## References

1. Benaroch, M. and Appari, A., **Financial Pricing of Software Development Risk Factors**, IEEE Software, Volume 27, Issue 5, October 2010, pp. 65 - 73, ISSN 0740-7459.
2. Judea Pearl, **Causality: Models, Reasoning and Inference**, 2nd Edition, Cambridge University Press, 2009, 484 pp, ISBN 978-0521895606.
3. Keshlaf, A.A. and Riddle, S., **Risk Management for Web and Distributed Software Development Projects**, 2010 Fifth International Conference on Internet Monitoring and Protection (ICIMP), 9-15 May 2010, pp. 22 - 28, ISBN 978-1-4244-6726-6.
4. Munch, J, **Risk Management in Global Software Development Projects: Challenges, Solutions, and Experience**, 2011 Sixth IEEE International Conference on Global Software Engineering Workshop (ICGSEW), 15-18 August 2011, pp. 35 - 35, ISBN 978-1-4577-1839-7.
5. Sadiq, M., Rahmani, M.K.I., Ahmad, M.W. and Sher Jung, **Software risk assessment and evaluation process (SRAEP) using model based approach, Networking and Information Technology (ICNIT)**, 2010 International Conference, 11-12 June 2010, pp. 171 – 177, ISBN 978-1-4244-7579-7.
6. Saleem, M.Q., Jaafar, J. and Hassan, M.F., **Model driven security frameworks for addressing security problems of Service Oriented Architecture**, 2010 International Symposium in Information Technology (ITSim), 15-17 June 2010, pp. 1341 – 1346, ISBN 978-1-4244-6715-0.
7. Ion Ivan, Catalin Boja, **Metode Statistice in analiza software**, Editura ASE, Bucuresti, 2004, 482 pg, ISBN 973-594-498-7
8. Adrian Mihai Vişoiu, **Rafinarea Metricilor Software**, PhD Thesis

---

[1]Catalin Alexandru TANASIE is a graduate of the Faculty of Cybernetics, Statistics and Economic Informatics within the Bucharest University of Economic Studies, the Economic Informatics specialization, 2007 promotion. Starting the same year he attended the Informatics Security Master in the same institution, and is currently a PhD student at the Doctoral School within the Bucharest University of Economic Studies. His concerns lie in the fields of distributed applications programming, evolutionary algorithms development, part of the field of artificial intelligence - neural and genetic programming. Currently he works as an application designer in a financial institution, in areas concerning the development of commercial applications.

[2]Emanuel Eduard HERTELIU has graduated the Faculty of Economic Cybernetics, Statistics and Informatics, in 2009, as part of the Bucharest Academy of Economic Studies, followed by the Economic Informatics Master Program and is currently a PhD candidate as part of the same institution. He is interested in software development using computer programming languages as C++, C# and Java, as well as Web-oriented application programming, using Html, Php or the ADO.NET technology.

JOURNAL OF APPLIED QUANTITATIVE METHODS



# PIATA MUNCII INTRE FORMAL SI INFORMAL (LABOUR MARKET BETWEEN FORMAL AND INFORMAL) by Silvia PISICA, Valentina VASILE and Vergil VOINEAGU

**Alexandru ISAIC-MANIU**
PhD, University Professor,
Department of Statistics and Econometrics,
Bucharest University of Economic Studies, Romania

**Web page:** www.amaniu.ase.ro **E-mail:** alexandru.isaic@csie.ase.ro

**Abstract:**

National economy is functioning and generates results both in formal ("official") area and in the informal ("unregulated", "unofficial", "unstructured") area of the economy. The informal sector exists, to a higher or lower extent, in all modern economies. It generates results, creates jobs, entails population income and is conditioning a significant part of consumption expenditure, while its presence contradicts the experts' predictions, formulated few decades ago, on its predictable reduction and extinction as national economies develop themselves. The study on the informal sector of the economy presents both a theoretical interest for economics, and a practical one as well.

**Key words:** labor market; statistics; formal; informal


The paper, whose title has a journalistic nuance, is the outcome of the collaboration between well-known experts from research, public statistics and education fields, focuses upon the analysis of informal sector, with particular emphasis on labour force. It is, probably, the most important issue, for the following reasons:

- pointing out the peculiarities of the informal sector in Romania, as compared to other countries, taking into account the novelty of this topic for the transition countries, under the circumstances where the informal sector was practically inexistent in former socialist economies;

- accurate sizing of employment /unemployment;

- measuring the unobserved economy volume in the system of national accounts;

- estimating as accurately as possible the level of labour force budgetary income, etc.

- adjusting the social protection policies, taking into consideration the informal sector size.

The major significance this paper issuing is pointed out in the foreword signed by the Academician Gheorghe Zaman: "One of the major contributions of the paper envisages **the definition and the profile of a person employed in the informal area**, according to the main demo-socio-economic characteristics (such as sex, area of residence, age, ethnicity, employment, education, economic activity, occupational status etc.). The theoretical and practical virtues of the paper also reside in the competent research of the relationships between formal and informal labour market, realistically inquiring whether in a modern society the informal sector plays an exclusively harmful role or, on the contrary, it is also a propeller of progress. The paper structure serves to achieve the stated purpose of the authors undertaking and mainly develops the following topics:

- conceptual developments of formal and informal economy;

- informal sector approach in the labour market context;

- defining and identifying the data sources and setting up the methodologies for compiling the main indicators based on which employment in informal area could be measured;

- deepening the causality of amplified informal economy existence, during recent years, both in Romania and in Europe ;

- building up a regressive model for measuring the ratio between the number of those employed in the informal sector and the number of households;

- drawing up a consistent set of conclusions both at theoretical-methodological level (e.g. the necessity of ILO review for the algorithm of measuring the employment in the informal sector, as well as for the measurement of employment in informal economy and in households sector) and at the level of governmental strategies and policies meant to stimulate the activities shift from the informal towards the formal sector, actions meant to reduce illegal work etc.

The paper, a novelty in Romanian economic literature, is comparable to the most valuable international papers in this field, is based on a thorough theoretical documentation, characterised by scientific rigour and originality, is rich in concrete analyses of the informal sector and unobserved economy based on actual Romanian data and is finalised by wording pertinent conclusions at the practical level of macroeconomic management.

This remarkable editorial issuing, for which both the authors and the editor deserve congratulations, is launching a genuine challenge to theoreticians and to the official statistics experts, as well as to decision-makers from governmental policies field.

The paper addresses a wide audience, from experts in labour economy to the experts in human resources management, financiers specialised in budgetary resources, professors and researchers in economics, doctorate and master degree candidates, as well as students in economics.