The mathematical model of estimation the multidimensional steganalytical methods reliability

I Atlasov$^1$, R Solodukha$^2$

$^1$ Moscow University of the Ministry of Internal Affairs of Russia, Moscow, Russia
$^2$ Voronezh Institute of the Ministry of Internal Affairs of Russia, Voronezh, Russia

E-mail: standartal@list.ru

Abstract. The most effective approach in steganalysis is adaptive multidimensional models. In order to apply multidimensional steganalytical methods in computer forensic expertise, it is necessary to evaluate their reliability in terms of mathematical statistics. The article is devoted to development of a mathematical model for estimating the reliability of the multidimensional stegananalytical methods. Wherein such parameters as sampling representativeness, significance level, and error are taken into account. The article describes the relevance of the topic, provides mathematical premises of the model development. The model based on evaluation of regression. The relationship between sample representativeness and significance level is determined. An example of application of the proposed model in the computer forensics expertise is considered.

1. Introduction

The progress in digital steganography and the massive development of freeware software products that implement steganographic methods have led to an increase of interest in methods and tools of detecting hidden information by law enforcement. At the same time the problem of providing forensic by a steganalytical system has been not solved. Despite the existence of a number of steganalytical methods their operability scope is significantly limited and the results can be contradictory. There is no unified approach to assessing the reliability of methods. Thus, there is a task of systematization of the steganalysis methodology with the final result as steganalytical system for forensics.

The most effective approach in steganalysis is adaptive multidimensional models (SRM [1], SPAM [2], JRM [3], CC-C300 [4], etc.). Here it is necessary to have a set of containers with a priori known embedding values, then extract the feature vectors, then carry out the classification procedure. As we can conclude from the publications the topic of feature extraction for steganalysis is beginning to exhaust itself. The overwhelming majority of publications are devoted to the classification technology (deep neural networks [5], ensemble classification [6], [7], SVM [8]) and their applications for analyzing containers obtained by various steganographic techniques (LSB, HUGO, WOW, YASS etc.) At the same time the majority of steganalytical-theme publications containing the experimental part are limited by the presentation of the ROC curves or the percentage of false negative and false positive errors for a standard sample set such as BOSSBase [9] or for author’s set. Herewith the results, by default, are extrapolated
to an arbitrary sample set where the reliability may be not the same [6], [1]. The question of sample representativeness remains open and strongly depends on a number of subjective reasons, for example, the researcher’s experience. However, in the computer forensic the expert must point the reliability degree of obtained results. In our opinion the reliability degree can be formalized using the mathematical statistics methods. It is necessary and sufficient to have two components. The first is the statistically significant results of applying steganoanalytical algorithms to various containers in a suitable for subsequent calculations form. The second is the corresponding mathematical model.

In [10] the conception of creating a steganalytic system (SAS) for computer forensic is presented. Within the framework of SAS functioning, the automation of the processes of obtaining trasological data (TD) is an integral task. The generalized scheme of TD-segment of SAS is shown in figure 1. It consists of three modules, DBMS and Data Storage (DS).

The potential containers searching module (PCSM) is a search-analysis script allows to download image files that satisfy the required parameters from the Internet resources to DS. The stego-embedding automation module (SEAM) is a console for managing the operation of scripts set (for example, AutoIt) that create the effect of user presence when working with steganography software. At the input of SEAM there are a lot of empty containers, at the output there are embedded containers with various payload size. The feature extraction module (FEM) implements the steganoanalytical algorithms to container sets, recording the result in a database. Each module is a DBMS client writing own information in the trasological data database (TDDDB). PCSM writes file name, size, type, required file metadata, SEAM writes embedded file name, payload size, stego-program name, FEM writes extracted features and image characteristics values.

One of the tasks of SAS creating is the development of mathematical models and algorithms for estimating the reliability of work and the criteria for comparing the effectiveness of steganoanalytical methods [11]. Within the framework of this concept, a database and several SAS modules have been developed, for example, an embedding automation module and a tracological characteristics calculating module. The advantage of using a large database of images is the possibility to select a set of [12] containers with characteristics similar to analyzed container. Further training and classification is executed for this set. Thus, an infrastructure for obtaining and accumulating statistical data has been created and the proposed mathematical model fills a theoretical gap in the reliability estimating of the results of multidimensional steganoanalytical methods.

![Generalized scheme of SAS TD-segment.](image-url)
The work is based on a study of the changes in the distance between the feature vectors of containers set. The accurate statistical estimates of the normal distribution are used. The representativeness of the set is researching for understanding whether it is necessary to continue the calculations or to stop at the achieved results during computer forensic examination.

2. Mathematical premises

Let us designate random quantities as letters of the Greek alphabet, and constants as letters of the Latin alphabet. We use the following notation: in large Latin letters $A = \{a_{ij}\}^{m}_{i,j=1}$, $X = \{x_{ij}\}^{m}_{i,j=1}$, $W = \{w_{ij}\}^{m}_{i,j=1}$ are denoted the matrices, $|A|$, $|X|$, $|W|$ are denoted determinants of these matrices, respectively.

Let us consider some known definitions [13], which are necessary for further work. The matrix $A = \{a_{ij}\}^{m}_{i,j=1}$ to be symmetric, if $a_{ij} = a_{ji}$ for every $i, j = 1, \ldots, m$. The matrix $A = \{a_{ij}\}^{m}_{i,j=1}$ to be positive-definite if for every $x = (x_1, \ldots, x_m) \neq 0$ the inequality $\langle Ax, x \rangle = \sum_{i,j=1}^{m} a_{ij} x_i x_j > 0$ is true.

We also introduce $m$-dimensional independent random vectors

$$\zeta_k = (\zeta_{1k}, \ldots, \zeta_{mk}), \quad k = 1, \ldots, n > m + 1,$$

with probability distribution density

$$f_{\zeta_k}(\bar{x}) = \frac{\sqrt{|A|}}{(2\pi)^{n/2}} \exp \left( -\frac{1}{2} \left\langle A \left( \bar{x} - \bar{h} \right), \left( \bar{x} - \bar{h} \right) \right\rangle \right),$$

where $\zeta = (\zeta_1, \ldots, \zeta_m)$, $A = \{a_{ij}\}^{m}_{i,j=1}$ is positive-definite symmetric matrix and

$$\bar{x} = (x_1, \ldots, x_m), \quad \bar{h} = (h_1, \ldots, h_m).$$

This distribution is called the $m$-dimensional normal distribution. Consider the example of construct the distribution of the random variable $\zeta_{ab}$. Let us consider $w = \sum_{i=1}^{s} w_i$ of files which consist of $v = \sum_{i=1}^{s} v_i$ pixels. Suppose $w_i$ ($i = 1, \ldots, s$) files contain $v_i$ specific color pixels. In this case we can construct a discrete distribution law of the random variable $\zeta_{ab}$, which takes on values $w_i$ ($i = 1, \ldots, s$) with $\frac{w_i}{w}$ probability.

Consider the meaning of distribution parameters. It is known [14] that

$$M(\zeta) = h_k, \quad K(\zeta_k) = M(\zeta_k \zeta_l) = M(\zeta_k) M(\zeta_l) = a_{kl}^{-1},$$

i.e. $h_k = M(\zeta_k)$ is mathematical expectation of random variable $\zeta_k$, and matrix's $A^{-1}$ element in row $k$ and column $l$ is $a_{kl}^{-1} = K(\zeta_k \zeta_l)$ is the correlation coefficient of random variables $\zeta_k$ and $\zeta_l$.

Because of we must to use normal distribution to find the regression consider the various forms of normal distribution. Note that if $A$ is a symmetric positive-definite matrix, then $A^{-1}$ is a symmetric positive-definite matrix too. Indeed for $w \in R^m$ are $x = A^{-1} w \in R^m$ and $\langle A^{-1} w, w \rangle = \langle x, Ax \rangle \geq 0$. Therefore, we can assume that in the distribution function of a random variable $\zeta$ the matrix $A$ can be replaced by a matrix $A^{-1}$. In this case, the distribution function of a random variable $\zeta$ becomes such as follows

$$f(\bar{x}) = \frac{\sqrt{|A|}}{(2\pi)^{n/2}} \exp \left( -\frac{1}{2} \left\langle A^{-1} \left( \bar{x} - \bar{h} \right), \left( \bar{x} - \bar{h} \right) \right\rangle \right),$$

$$= \frac{1}{\sqrt{|A||(2\pi)^{n/2}}} \exp \left( -\frac{1}{2} \sum_{i,j=1}^{m} a_{ij}^{-1} \frac{|A|}{|A|} (x_i - h_i)(x_j - h_j) \right),$$

(2)
where $A_{ij}$ is cofactor of $a_{ij}$. The value $a_{kl} = K(\xi_k, \xi_l)$ is the covariance coefficient of random values $\xi_k$ and $\xi_l$. Therefore it makes sense to introduce for $1 \leq k, l \leq m$ the mean square deviation and correlation coefficient

$$\sigma_k = \sqrt{D(\xi_k^2)} = \sqrt{a_{kk}}, \quad \rho_{kl} = \frac{K(\xi_k, \xi_l)}{\sqrt{D(\xi_k^2)D(\xi_l^2)}} = \frac{a_{kl}}{\sigma_k \sigma_l}. \quad (3)$$

Let’s consider the matrixes

$$\Delta = \begin{pmatrix} \rho_{ij} \end{pmatrix}_{ij=1}^m; \quad \Sigma = \begin{pmatrix} \sigma_i \delta_{ij} \end{pmatrix}_{ij=1}^m \overset{def}{=} \begin{pmatrix} \sigma_{ij} \end{pmatrix}_{ij=1}^m,$$

where $\delta_{ij}$ is the Kronecker symbol. Next, for brevity, we denote by a symbol $\Delta$ both the matrix $\Delta$ and the determinant of the matrix $\Delta$. Denote by a symbols $A_{ij}$, $\Delta_{ij}$, $\Sigma_{ii}$ the cofactors of $a_{ij}$ of matrixes $A$, $\Delta$, $\Sigma$, respectively. Notice that

$$A = \Sigma \Delta \Sigma; \quad |A| = |\Sigma|^2 \Delta = \Delta \prod_{k=1}^m \sigma_k^2; \quad A_{ij} = \Sigma_{ii} \Delta_{ij} \Sigma_{jj} = \sigma_i \sigma_j \Delta_{ij} \prod_{k \neq i,j} \sigma_k^2.$$ 

The inverse matrix $A^{-1}$ is as follows

$$A^{-1} = \frac{1}{|A|} \left\{ A_{ij} \right\}_{i,j=1}^m = \frac{1}{\Delta \prod_{k=1}^m \sigma_k^2} \left\{ \sigma_i \sigma_j \Delta_{ij} \prod_{k \neq i,j} \sigma_k^2 \right\}_{i,j=1}^m = \frac{1}{\Delta} \left\{ \Delta_{ij} \sigma_i \sigma_j \right\}_{i,j=1}^m.$$

Then for vectors $\vec{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$ and $\zeta = (\zeta_1, \ldots, \zeta_m)$ the equal to (1) $\zeta$ becomes such as follows

$$f_{\zeta}(\vec{x}) = \frac{\exp \left( -\frac{1}{2|\Delta|} \sum_{i,j=1}^m A_{ij} (x_i - h_i)(x_j - h_j) \right)}{\sqrt{|A| (2\pi)^2 \Delta \prod_{i=1}^m \sigma_i}} = \frac{\exp \left( -\frac{1}{2\sigma^2} \sum_{i,j=1}^m \Delta_{ij} (x_i - h_i)(x_j - h_j) \right)}{(2\pi)^2 \sqrt{\Delta \prod_{i=1}^m \sigma_i}}. \quad (4)$$

For example, in the particular case $m = 2$ the formula (4) for vectors $\vec{x} = (x_1, x_2) \in \mathbb{R}^2$ and $\zeta = (\zeta_1, \zeta_2)$ is as follows

$$f_{\zeta}(\vec{x}) = \frac{\exp \left( -\frac{\sigma_2^2(x_1 - h_1)^2 - 2\sigma_1 \sigma_2 (x_1 - h_1)(x_2 - h_2) + \sigma_1^2(x_2 - h_2)^2}{2\sigma_1^2 \sigma_2^2 (1 - \rho_{12}^2)} \right)}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho_{12}^2}} = \frac{\exp \left( -\frac{1}{2(1 - \rho_{12}^2)} \left( \frac{(x_1 - h_1)^2}{\sigma_1^2} - \frac{\rho_{12}(x_1 - h_1)(x_2 - h_2)}{\sigma_1 \sigma_2} + \frac{(x_2 - h_2)^2}{\sigma_2^2} \right) \right)}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho_{12}^2}}. \quad (5)$$

Also consider the regression. The regression coefficients $\beta_{21}$, $\beta_{32.1}$ and $\beta_{m_i,s}$ are founded with considering the condition

$$M (|\xi_2 - M (\xi_2)| - \beta_{21} |\xi_1 - M (\xi_1)|)^2 = \min_{v \in \mathbb{R}} M (|\xi_2 - M (\xi_2)| - v |\xi_1 - M (\xi_1)|)^2,$$

$$M (|\xi_3 - M (\xi_3)| - \beta_{32.1} |\xi_2 - M (\xi_2)| - \beta_{31.2} |\xi_1 - M (\xi_1)|)^2 = \min_{v \in \mathbb{R}} M (|\xi_3 - M (\xi_3)| - v |\xi_2 - M (\xi_2)| - w |\xi_1 - M (\xi_1)|)^2,$$

$$M (|\xi_1 - M (\xi_1)| - \sum_{i=2}^m \beta_{1i,s} |\xi_i - M (\xi_i)|)^2 = \min_{v \in \mathbb{R} \cup \{0\}} M (|\xi_1 - M (\xi_1)| - \sum_{i=2}^m v_i |\xi_i - M (\xi_i)|)^2.$$ 

According to [14] the coefficients of the regression are written as follows. The first index is associated with the index of the random variable $\beta_{m_i,s}$, according which the remaining
coefficients are sought. The second is related to the index of a random variable, next to which it is located. Further after the comma the remaining coefficients are written in an arbitrary order.

Let us find the analytical expressions for these coefficients in terms of the partial derivatives

\[ M ([\zeta_1 - M (\zeta_1)] - \sum_{i=2}^{m} \beta_{1i,*} [\zeta_i - M (\zeta_i)])^2 \]

for all input variables equated to zero. We get a system of \(m - 1\) equations

\[ \sum_{i=2}^{m} \beta_{1i,*} M ((\zeta_i - M (\zeta_i)) (\zeta_k - M (\zeta_k))) = M ((\zeta_1 - M (\zeta_1)) (\zeta_k - M (\zeta_k))). \]

Denote the matrix

\[ \Lambda = \{ M ((\zeta_i - M (\zeta_i)) (\zeta_k - M (\zeta_k)))\}_{i,j=1}^{m}. \]

Because of \(\Lambda\) is positive-definite the system equations determinant is also positive. Moreover the determinant is equal to cofactor \(\Lambda_{11} > 0\) for element \(\lambda_{11}\). Hence the system has only decision

\[ \beta_{11,*} = -\frac{\lambda_{11}}{\Lambda_{11}}. \]

Thus the coefficients \(\beta_{11,*}\) are defined uniquely and can be calculated analytically. In particular,

\[ \beta_{21} = \rho_{12} \frac{\sigma_2}{\sigma_1}, \beta_{31,2} = \frac{\sigma_3 \rho_{13} - \rho_{23} \rho_{12}}{\sigma_1 (1 - \rho_{12}^2)}, \beta_{32,1} = \frac{\sigma_3 \rho_{23} - \rho_{13} \rho_{12}}{\sigma_1 (1 - \rho_{12}^2)}, \beta_{21,3} = \frac{\sigma_2 \rho_{12} - \rho_{23} \rho_{13}}{\sigma_1 (1 - \rho_{13}^2)}. \]

3. Regression estimation

Let us consider the estimation of regression coefficients in accordance with \([15]\). If \(m = 2\) then

\[ m_1 = \frac{1}{n} \sum_{k=1}^{n} \zeta_{1k}, m_2 = \frac{1}{n} \sum_{k=1}^{n} \zeta_{2k}, \]

\[ s_1 = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (\zeta_{1k} - m_1)^2}, s_2 = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (\zeta_{2k} - m_2)^2}, \]

\[ r_{12} = \frac{\sum_{k=1}^{n} (\zeta_{1k} - m_1) (\zeta_{2k} - m_2)}{\sqrt{\sum_{k=1}^{n} (\zeta_{1k} - m_1)^2} \sqrt{\sum_{k=1}^{n} (\zeta_{2k} - m_2)^2}}, b_{21} = \frac{r_{12} s_2}{s_1}. \]

For the random variables \(\zeta_k = (\zeta_{1k}, \zeta_{2k}), k = 1, ..., n > 3\), the random variables

\[ t = \frac{s_1 \sqrt{n - 2}}{s_2 \sqrt{1 - r_{12}^2}} (b_{21} - \beta_{21}) \]

has Student’s distribution with \(n - 2\) degree of freedom.

If \(m > 2\) then the matrix

\[ R = \begin{pmatrix}
    r_{11} & r_{12} & \cdots & r_{1(m-1)} & r_{1m} \\
    r_{21} & r_{22} & \cdots & r_{2(m-1)} & r_{2m} \\
    \vdots & \vdots & \ddots & \vdots & \vdots \\
    r_{(m-1)1} & r_{(m-1)2} & \cdots & r_{(m-1)(m-1)} & r_{(m-1)m} \\
    r_{m1} & r_{m2} & \cdots & r_{m(m-1)} & r_{mm}
\end{pmatrix}, \]

\[ \text{(9)} \]
is considered where
\[ r_{ij} = \frac{\sum_{k=1}^{n} (\zeta_{ik} - m_{i}) (\zeta_{jk} - m_{j})}{\sqrt{\sum_{k=1}^{n} (\zeta_{ik} - m_{i})^2} \sqrt{\sum_{k=1}^{n} (\zeta_{jk} - m_{j})^2}} \]  

and \( R_{ij} \) is cofactor of the element \( r_{ij} \) of the matrix \( R \). Denote
\[ m_{1} = \frac{1}{n} \sum_{k=1}^{n} \zeta_{1k}, m_{2} = \frac{1}{n} \sum_{k=1}^{n} \zeta_{2k}, \]
\[ s_{1} = \sqrt{\sum_{k=1}^{n} (\zeta_{1k} - m_{1})^2}, s_{2} = \sqrt{\sum_{k=1}^{n} (\zeta_{2k} - m_{2})^2}, \]
\[ b_{12.34...k} = \left( \frac{s_{1}}{s_{2}} R_{12}, \frac{s_{1}}{\sqrt{R_{11}}}, s_{1}, s_{2}, \frac{s_{2}}{\sqrt{R_{22}}} \right). \]

In this case the random variable
\[ t = \sqrt{n - k} \frac{s_{1}}{s_{2}} (b_{12.34...k} - \beta_{12.34...k}) \]
has Student’s distribution with \( n - m \) degree of freedom
\[ f_{n-m} (t) = \frac{1}{\sqrt{\pi (n - m)}} \frac{\Gamma \left( \frac{n-m+1}{2} \right)}{\Gamma \left( \frac{n-m}{2} \right)} \left( 1 + \frac{t^2}{n - m} \right)^{-\frac{n-m+1}{2}}. \]

The final result consists in calculating
\[ \min_{v=(v_{2},...,v_{m})} \left( \sum_{i=1}^{m} \beta_{i} \zeta_{i} - M (\zeta_{i}) \right)^{2} = \]
\[ = M (\sum_{i=1}^{m} \beta_{i} \zeta_{i} - M (\zeta_{i})) \]
\[ = M (\sum_{i=1}^{m} \beta_{i} \zeta_{i} - M (\zeta_{i})) + \sum_{i,j=1}^{m} \beta_{i} \beta_{j} (\zeta_{i} - M (\zeta_{i})) (\zeta_{j} - M (\zeta_{j})) \]
\[ = D (\zeta_{1}) + \sum_{i,j=1}^{m} \beta_{i} \beta_{j} (\zeta_{i} \zeta_{j}) + \sum_{i,j=1}^{m} \beta_{i} \beta_{j} (\zeta_{i} \zeta_{j}) \]

4. Determining the relationship between sample representativeness and significance level
Let \( m_{i} = \frac{1}{n} \sum_{k=1}^{n} \zeta_{ik} \). The approximate replacement is possible
\[ D (\zeta_{1}) = s_{1}^{2} = \sum_{k=1}^{n} (\zeta_{ik} - m_{i})^2, \]
\[ K (\zeta_{i}, \zeta_{j}) = s_{1} s_{j} r_{ij} = \sum_{k=1}^{n} (\zeta_{ik} - m_{i}) (\zeta_{jk} - m_{j}), \]
\[ \beta_{12.34...k} = b_{12.34...k} = \frac{1}{s_{2} R_{11}} \frac{s_{1}}{s_{2}} R_{12}. \]
So, it remains to estimate the coefficients $D(\zeta_1)$, $K(\zeta_1\zeta_1)$ and $K(\zeta_1\zeta_2)$ with determined degree of reliability. It is known [13] that random variable

$$n \left( \sum_{k=1}^{n} (\zeta_{ik} - h_i)^2 \right) / D(\zeta_i)$$

has probability density

$$f(x, n^{-1/2}, 1/2) = \begin{cases} \frac{x^{n-3/2}}{2^{n-1/2} \Gamma(n-1/2)} \exp\left(-\frac{1}{2}x\right) & \text{where } x > 0, \\ 0 & \text{where } x < 0, \end{cases}$$

where $\Gamma(k) = \int_0^\infty t^{k-1} \exp(-t) dt$.

Let’s sum up all the work described above. It is necessary to choose a significance level $P_0$ close to one (for example, $P_0 = 0.99$) that is required for the steganalytical method. Also it is necessary to choose a error $\varepsilon > 0$ for approximate calculating. The more close $\varepsilon$ to zero the more accurate the calculations. On the basis of $P_0$ and $\varepsilon$ the large natural number $n = n_0$ is chosen so as to satisfied the inequalities

$$P\left(|b_{12.34,...k} - \beta_{12.34,...k}| < \varepsilon\right) =$$

$$P\left(-\varepsilon \sqrt{n - k_{12.34,...k}^2} < \sqrt{n - k_{12.34,...k}^2} (b_{12.34,...k} - \beta_{12.34,...k}) < \sqrt{n - k_{12.34,...k}^2} \varepsilon\right) =$$

$$\int_{-\varepsilon \sqrt{n - k_{12.34,...k}^2}}^{\varepsilon \sqrt{n - k_{12.34,...k}^2}} f_{n-m}(x) \, dx > P_0$$

and

$$P\left(\left|\frac{n}{k_{12.34,...k}} - 1\right| < \varepsilon\right) = P\left(\frac{n-\varepsilon}{n} < \frac{D(\zeta_i)}{n \sum_{k=1}^{n} (\zeta_{ik} - m_i)^2} < \frac{n+\varepsilon}{n}\right) =$$

$$P\left(\frac{n}{1+\varepsilon} < \frac{n}{n \sum_{k=1}^{n} (\zeta_{ik} - m_i)^2} < \frac{n}{1-\varepsilon}\right) = \int_{1-\varepsilon}^{1+\varepsilon} f(x, n^{-1/2}, 1/2) \, dx > P_0.$$

That is, by choosing sufficiently large $n$, we achieve the fulfillment of conditions

$$\int_{-\varepsilon \sqrt{n - k_{12.34,...k}^2}}^{\varepsilon \sqrt{n - k_{12.34,...k}^2}} f_{n-m}(x) \, dx > P_0 \int_{1-\varepsilon}^{1+\varepsilon} f(x, n^{-1/2}, 1/2) \, dx > P_0. \quad (15)$$

The obtained number $n$ answers the question about the sample representativeness. Further, substituting all the data, we calculate

$$W(\zeta_1) = s_1^2 - \sum_{i=2}^{m} b_{i1} s_i r_{1i} + \sum_{i,j=2}^{m} b_{ij} s_i s_j r_{ij}. \quad (16)$$

For every $1 < k \leq n$ by analogy with $W(\zeta_1)$ the $W(\zeta_k)$ are received. Then the

$$V(\zeta) = \frac{1}{n} \sum_{k=1}^{n} W(\zeta_k) \quad (17)$$

is obtained.

So, for random variables set $\zeta_k = (\zeta_{1k}, \ldots, \zeta_{mk}), \ k = 1, \ldots, n > m + 1$, which are corresponded to empty containers the value $V1(\zeta)$ is calculated. For payloaded container set the value $V2(\zeta)$ is calculated. If the inequality

$$V1(\zeta) < V2(\zeta) \quad (18)$$

is truly then it is possible to state the reliability of steganalytical method is $P_0$ with error $\varepsilon$. 


5. Example of application of the proposed model
Suppose that during the investigation a flash drive has been received. It contains several folders
with images in JPEG format approximately 100–200 files in every folder. Also the distribution
pack of steganography program StegHide is found on it. A forensic expert is asked whether any
files contain steganographic payloads embeded by the StegHide program. The expert applies a
method of digital steganalysis, for example, CC-CHEN (CHEN features enhanced by Cartesian
calibration) [17] with further classification using SVM. As a result of the analysis the several
files are determined as payloaded by StegHide. It remains unclear how accurate the result,
although the expert should point in the conclusion, not only the results, but also their reliability.
Suppose that for examinations of this class the level of significance should be at least 0.99, the
error — no more than 0.001. Now to verify whether the chosen steganalytical method meets
the requirements expert must perform the following actions.
1. Forming a set of payloaded and empty containers At this stage, the expert creates payloaded
containers from the presumably empty containers (by StegHide using). Then containers are
evenly distributed among the files. Now the expert must heuristically relate the size of the
sample and the computational costs. Let there are initially 10 folders with 100 files in each. The
first set is only empty containers, the another set is empty and payloaded containers.

2. The definition of a feature set of the payload presence The CC-CHEN model has a
dimension of 972. The calculation of characteristics is performed in the MatLab environment,
the corresponding m-file can be downloaded from the resource. A generalized scheme of process
is presented in figure 1. Analyzed files are located on storage system, the file paths are written in
a MS SQL Server database. Management is provided through the GUI Matlab. Images one by
one are read into MatLab, the features are written to the database. To calculate the reliability
of the method, the features are also read into MatLab and processed in accordance with the
following points.

3. Construction of random variables The range of characteristics value is divided in accordance
with the rules of constructing a discrete random variable. Since all the features are normalized
the interval $[0; 1]$ is divided into equal parts with a step of 0.1. Next, for each k-folder a discrete
random variable $\zeta_k = (\zeta_{1k}, \ldots, \zeta_{10k})$, $k = 1, \ldots, 10$ is constructed.

Figure 2. Generalized scheme of features extraction and storage.
4. **Calculation of coefficients** $s_{1\ast}$, $s_{2\ast}$ Using formulas (9)–(14) there are calculated for a set of empty containers $s_{1\ast}=3.15$ $s_{2\ast}=8.47$. For a set of empty and payloaded containers $s_{1\ast}=5.34$ $s_{2\ast}=9.58$.

5. **Verification of inequality (15)** If (15) is not fulfilled the steps 1–4 must be repeated for more number of n-folders with files. For our case (15) becomes fulfilled since $n=8$ for a set of empty containers, since $n=9$ for a set of empty and payloaded containers. Since $n=10$ was taken there is no need to recalculate the results of formulas (9)–(14) for more number of folders and files.

6. **Calculation of $V_1(\zeta)$ and $V_2(\zeta)$** Using formulas (16), (17) there are calculated $V_1(\zeta) = 0.2756$ and $V_2(\zeta) = 2.4358$.

7. **Verification of inequality (18)** Since $V_1(\zeta) > V_2(\zeta)$ it can be concluded on the acceptability of steganalytical method for the computer forensic expertise.

6. **Conclusion**

The article shows the relevance of the topic of estimating the effectiveness of adaptive multidimensional models (the question of models dimension reducing is considered in [19]) of stegananoanalysis in terms of mathematical statistics for computer forensics. The mathematical model allowing to estimate reliability of multidimensional steganalytical methods on the basis of calculation of regression coefficients is offered. The model makes it possible to determine the relationship between such parameters as sample representativeness, significance level, error. These parameters must be specified for computer forensic expertise, otherwise its results are not legitimate.

The advantage of the proposed model is simple algorithmization and application for various tasks changing the level of significance and error. The disadvantages of the model include the need for recalculation in the case of obtaining results of computer expertise that do not meet the required level of reliability. In other words this model estimates the reliability of the examination results of a posteriori.

The scope of the model is narrowed due to the inability to apply for a small number of files. This is due to the need to comply the requirement of normal distribution of the initial discrete random variables. Also, the analyst should be aware of the stego-program which performed the embedding. As a further development of this subject, it is proposed to establish a model in which it will be possible to determine the size of a representative sample at a fixed level of significance and error a priori based on the distribution parameters of the initial sample.

**References**

[1] Fridrich J and Kodovsky J 2011 *IEEE Transactions on Information Forensics and Security* 3 868

[2] Pevny T, Bas P and Fridrich J 2009 *Proceedings of the 11th ACM Multimedia & Security Workshop* 75

[3] Kodovsky J and Fridrich J 2012 *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security and Forensics* 8303

[4] Kodovsky J and Fridrich J 2011 *Proceedings SPIE, Electronic Imaging, Security and Forensics of Multimedia XIII* 7880

[5] Qian Y, Dong J, Wang W and Tan T 2015 *Proceeding SPIE, Media Watermarking, Security and Forensics* 9409

[6] Monarev V and Pestunov A 2016 *Lecture Notes in Computer Science* 9543 445

[7] Kodovsky J, Fridrich J and Holub V 2012 *IEEE Transactions on Information Forensics and Security* 2 432

[8] Deepa S and Umarani R 2016 *International Journal of Computer Science and Engineering (IJCSE)* 5 15

[9] BOSSBase http://agents.fel.cvut.cz/ boss/ index.php?mode=VIEW&tmpl=materials

[10] Solodukha R A 2016 *Bulletin of Voronezh Institute of the Ministry of Interior of Russia* 4 156
[11] Solodukha R A and Atlasov I V 2014 *Bulletin of Voronezh Institute of the Ministry of Interior of Russia* 2 139
[12] Eliseev A S and Tikidzhi-Khamburjan A R 2014 *Bulletin of South Federal University of Russia* 2 158
[13] Cramer H 1999 *Mathematical Methods of Statistics* (Princeton: University Press)
[14] Anderson T W 2003 *An Introduction to Multivariate Statistical Analysis* (Wiley: InterScience)
[15] Bartlett M S 1933. *On the Theory of Statistical Regression* (Edinburgh: Proc. Roy. Soc.)
[16] Kolmogorov A N 1974 *Basic Concepts of Probability Theory* (Moscow: Nauka)
[17] Kodovsky J and Fridrich J 2009 *Proceedings of the 11th ACM Multimedia and Security Workshop* 63
[18] Feature Extractors for Steganalysis http://dde.binghamton.edu/download/feature_extractors/
[19] Menshikh V, Atlasov I, Solodukha R 2017 *2nd International Ural Conference on Measurements (UralCon)* 227