# A Genetic Approach to Security Cost Modeling for Structured Content Validation

Dragos PALAGHITA
Academy of Economic Studies, Bucharest, Romania
mail@dragospalaghita.ro

*Structured content consists of sets of structured entities that are compared for validity. The validity characteristics of structured entities are presented. The AVIO application is presented and its security risks are analyzed. A genetic cost model is defined based on input from over one hundred individual users that accessed the application freely. The input data for the genetic algorithm has been obtained from the AVIO application use, AVIO risk analysis, AVIO recorded security costs, and structural characteristics of the AVIO application. The output of the genetic algorithm is a low error cost model validated using real data from the AVIO application after several improvements have been made to the implementation logic.*
***Keywords:*** *Cost, Security, Genetic Algorithm, Risk, Model*

## 1 Introduction

Structured content is constructed from entities that are defined based on similar principles and formats. The entities that make up the content base have identical structural characteristics and similar content representation formats. The structured content validation is analyzed for validity automatically based on predefined base rules set which forms the validation model. The structured content is validated by checking:

- representation format type which is valid if and only if it is identical to the base format specified in the predefined rules or any other format that can be converted to the base format without altering the content of the entity in any way;
- structural characteristics which are restricted to the specifications given in the base rules of the validation model; these characteristics include content size, memory storage needed to record the content or content homogeneity;
- orthogonality of entities added to the content base against the ones already included in it; structured entity orthogonality is centered on the degree of difference between two structured entities at a time; orthogonality measures for text and images are presented in [1] and [2] and implemented in the AVIO application.

The AVIO application is designed to compute the orthogonality of organizational identifiers by completing an organizational name orthogonality analysis and an organizational logo orthogonality analysis. An organizational name is a structured text entity that has the following properties:

- length represented by the number of characters in the inputted text including white spaces and accepted symbols;
- word count which is computed based on the number of words in the inputted text and it is computed by splitting the text according to white spaces.

An organizational logo is defined by a digital image which represents the identity of the organization having the following properties:

- size which is defined by the width and height of the image file representing the dimensions of the pixel matrix associated to the image which is used in establishing validity;
- color depth which represents the number of bits used to represent each pixel associated to the pixel matrix;
- color model used to represent the image colors as tuples of numbers [3] associated to each pixel;
- image file format which is defined by the method used to represent the image data in computer memory.

The AVIO application successfully validates the organizational identifiers by adhering to the base rule set defined in [1], [2] and for image quality referring to the algorithm described in [4].

## 2 User behavior based risk analysis in the AVIO application

In order to examine user behavior in the AVIO application twenty eight outcomes have been identified by analyzing the possible user actions form use cases of the AVIO application presented in Figure 1. The identified outcomes are divided into two groups, the NSG group containing outcomes that are not connected with any security issues and the SG group including security related outcomes. The outcomes included in the SG group number twenty items and the processing made in the AVIO application to reach them implies that the information flow is either sensitive in nature or is subject to security validation.

User behavior in the AVIO application is determined based on functionality access frequency. The access frequency was recorded while the application was accessed by over one hundred different users each of which have completed once the structured entity validation process of the AVIO application.

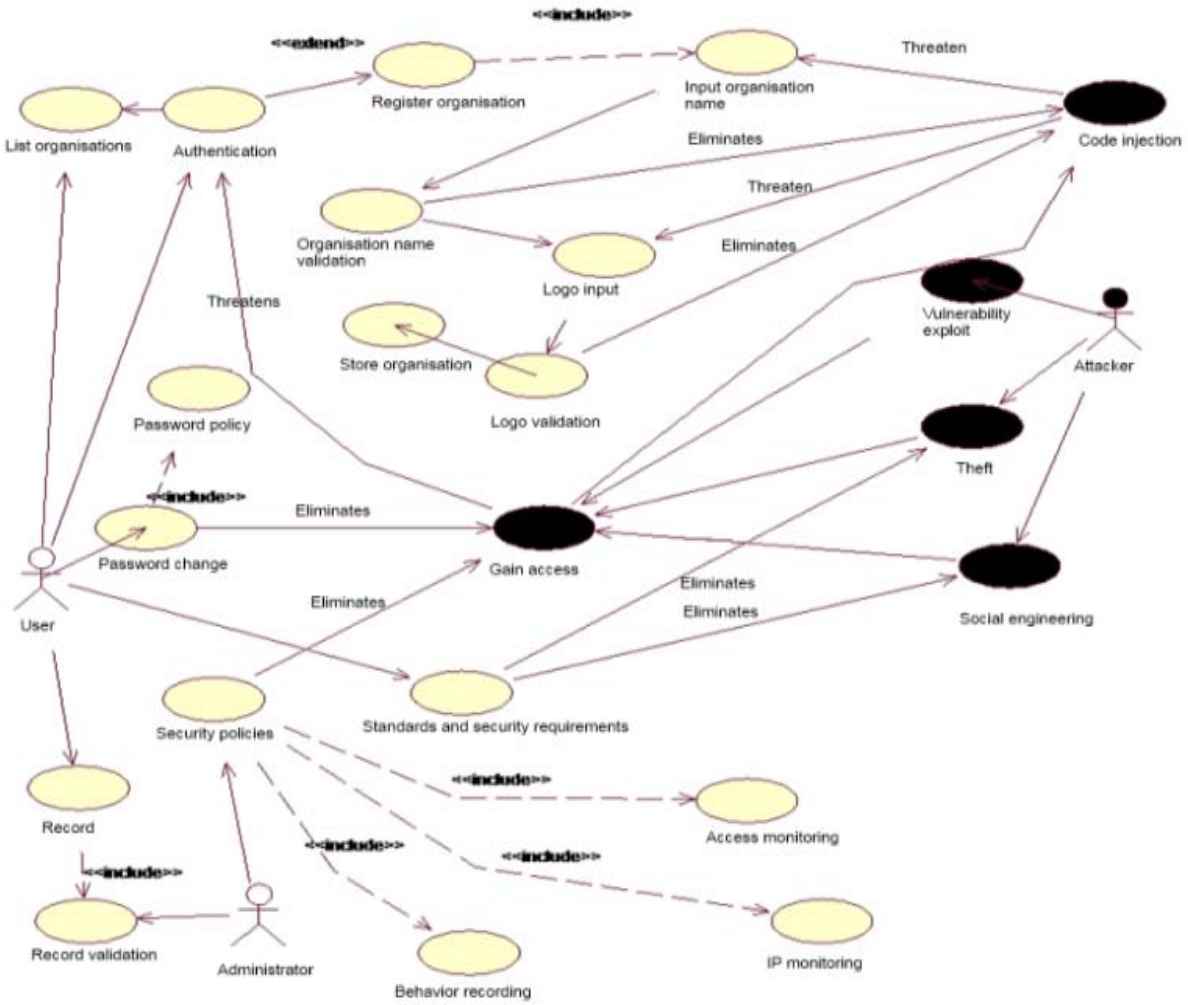


**Fig. 1.** Misuse case for the AVIO software

The outcome frequencies in both groups NSG and SG are presented in Figure 2. Overall 153 users have accessed the application, 112 users continued to the orthogonality computation page and completed 130 orthogonality computations and saved their content in the database 120 times.

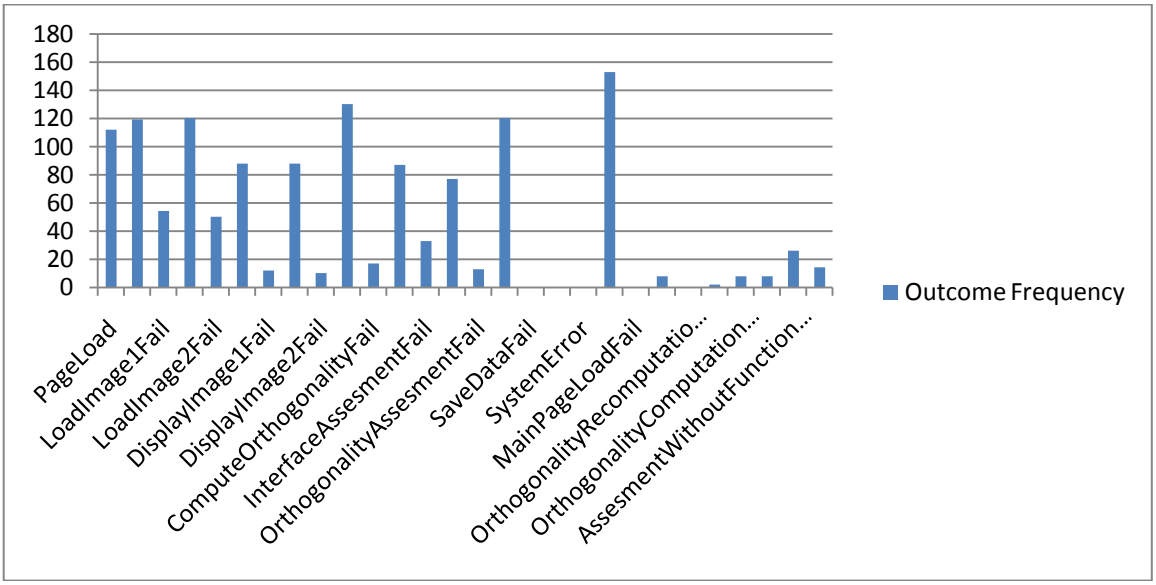


**Fig. 2.** Outcome frequencies in the AVIO application

For the outcomes in the SG group a risk rating on a scale of 1 to 10 is given based on the exposed functionality and exploit possibilities. The result of the qualitative risk evaluation is presented in Table 1.

**Table 1.** Risk rating for outcomes in the SG group

| Outcome | Outcome Count | Weighted Outcome | Risk rating | rbc |
|---|---|---|---|---|
| LoadImage1 | 119 | 91.54% | 10 | 9.1538 |
| LoadImage2 | 120 | 92.31% | 10 | 9.2308 |
| DisplayImage1 | 88 | 67.69% | 5 | 3.3846 |
| DisplayImage2 | 88 | 67.69% | 5 | 3.3846 |
| ComputeOrthogonality | 130 | 100.00% | 3 | 3.0000 |
| InterfaceAssesment | 87 | 66.92% | 6 | 4.0154 |
| OrthogonalityAssesment | 77 | 59.23% | 6 | 3.5538 |
| SaveData | 120 | 92.31% | 10 | 9.2308 |
| SystemError | 0 | 0.00% | 10 | 0.0000 |
| OrthogonalityComputationWithoutImage1 | 2 | 1.54% | 4 | 0.0615 |
| OrthogonalityComputationWithoutImage2 | 8 | 6.15% | 4 | 0.2462 |
| OrthogonalityComputationWithoutImages | 8 | 6.15% | 4 | 0.2462 |
| AssesmentWithoutFunctionality | 26 | 20.00% | 6 | 1.2000 |

Behavior risk based rating is established based on access frequency which is determined as a percentage of the total number of actions. The risk based indicator is useful as a most probable security incident highlighter based on functionality access frequency.

## 3 Model data analysis

The security cost model is developed based on:

- the behavioral risk based assessment, brc, which determines security risk exposure based on a qualitative risk assessment and frequency of access within the application according to experimental results;
- cyclomatic module complexity, cxm, which is determined for each module involved in the structured content validation procedures; the cyclomatic complexity is determined automatically using source code analysis tools; the complexity is determined only for modules that are activated while performing actions leading to outcomes placed in the SG group;

The complexity of the modules is presented in Table 2 according to each activated outcome from the SG group together with the development effort required to develop them in man hours.

**Table 2.** AVIO module complexity

| Outcome | Module Name | cxm | Man hours |
|---|---|---|---|
| LoadImage1 | Image1Validation | 3 | 20 |
| LoadImage2 | Image2Validation | 3 | 21 |
| DisplayImage1 | DisplayPathValidation1 | 2 | 10 |
| DisplayImage2 | DisplayPathValidation2 | 2 | 10 |
| ComputeOrthogonality | ImageCohesionValidation | 5 | 20 |
| InterfaceAssesment | InputValidation | 3 | 15 |
| OrthogonalityAssesment | InputValidation | 3 | 15 |
| SaveData | UserDataValidation | 4 | 25 |
| SystemError | UserDataBackup | 2 | 15 |
| OrthogonalityComputationWithoutImage1 | ImageSecurityCheck | 1 | 5 |
| OrthogonalityComputationWithoutImage2 | ImageSecurityCheck | 1 | 5 |
| OrthogonalityComputationWithoutImages | ImageSecurityCheck | 1 | 5 |
| AssesmentWithoutFunctionality | InputValidation | 3 | 15 |

The collected data is representative for the security cost model and due to the high number of users that have already used the AVIO application the behavioral risk based analysis has a high degree of correctness. The needed man hours to develop the modules were recorded during the development of the security system and include requirement development, software development, unit tests and maintenance operations done.

## 4 Genetic security cost model development

The genetic programming algorithm is developed based on a number of 10000 generations with an initial population of 20, an 80% crossover rate, a 8% mutation rate and uses elitism for preserving the best individuals of each generation. The input of the genetic algorithm used is defined by the behavioral based risk rating data presented in Table 1 and the module complexity and man hours cost presented in Table 2. The algorithm uses two parameters a, b and constant belonging to the interval [-5, 5] as outputs. In order to maintain the validity of the results the final parameters, a and b, are included in the interval [0, 10].

The output parameters are defined as:

- the *a* parameter corresponds to the behavioral risk rating, denoted by the variable rbc, computed based on risk rating and access frequency as shown in Table 1;
- the *b* parameter corresponds to the cyclomatic complexity of the modules accessed during the AVIO application processing, denoted by the cxm variable, according to Table 2.

The fitness function computes the average error of the chromosome $CH_j$ in the analyzed generation against the man hour costs results obtained from Table 2 according to the recorded data from the AVIO development cycle. The fitness function is given by the FFUNC indicator defined below.

$$FFUNC(CH_j) = \frac{\sum_{i=1}^{N} |x_j - d_i|}{N}$$

where:

N – number of outcomes that pose a security risk;

$d_i$ – man hours recorded for the development of module $d_j$;

$x_j$ – the man hour cost result obtained after parsing the mathematical expression tree associated to the $CH_j$ chromosome.

The model of the best individual is given by the CGS indicator with the formula:

$$CGS = ((((rcb*(rcb+4)/\sqrt{2})/(cxm+2+2*cxm))+cxm))+(cxm+\sqrt{(rcb+cxm)}/\sqrt{cxm}))+2*cxm$$

At an increase of one unit of complexity the value of the value of the CGS indicator increases with 3.479 units. At an increase of behavior based risk rating with one unit the value of the CGS indicator increases with 0.796 units. The value of the CGS indicator is significantly influenced by the module complexity. Table 3 presents the results obtained for the genetic information security cost indicator CGS.

**Table 3.** CGS indicator results

| **Result** | **rbc** | **cxm** | **CGS** |
|---|---|---|---|
| LoadImage1 | 9.1538 | 3 | 19.61237 |
| LoadImage2 | 9.2308 | 3 | 19.66925 |
| DisplayImage1 | 3.3846 | 2 | 11.83452 |
| DisplayImage2 | 3.3846 | 2 | 11.83452 |
| ComputeOrthogonality | 3.0000 | 5 | 22.88143 |
| InterfaceAssesment | 4.0154 | 3 | 15.90663 |
| OrthogonalityAssesment | 3.5538 | 3 | 15.58817 |
| SaveData | 9.2308 | 4 | 23.06271 |
| SystemError | 0.0000 | 2 | 9.189207 |
| OrthogonalityComputationWithoutImage1 | 0.0615 | 1 | 5.073824 |
| OrthogonalityComputationWithoutImage2 | 0.2462 | 1 | 5.29037 |
| OrthogonalityComputationWithoutImages | 0.2462 | 1 | 5.29037 |
| AssesmentWithoutFunctionality | 1.2000 | 3 | 14.0368 |

Statistics were obtained for each $100^{th}$ generation and the best individual was represented graphically to show the developed tree and operations. Figure 3 presents the best individual in the final population at generation 10000.

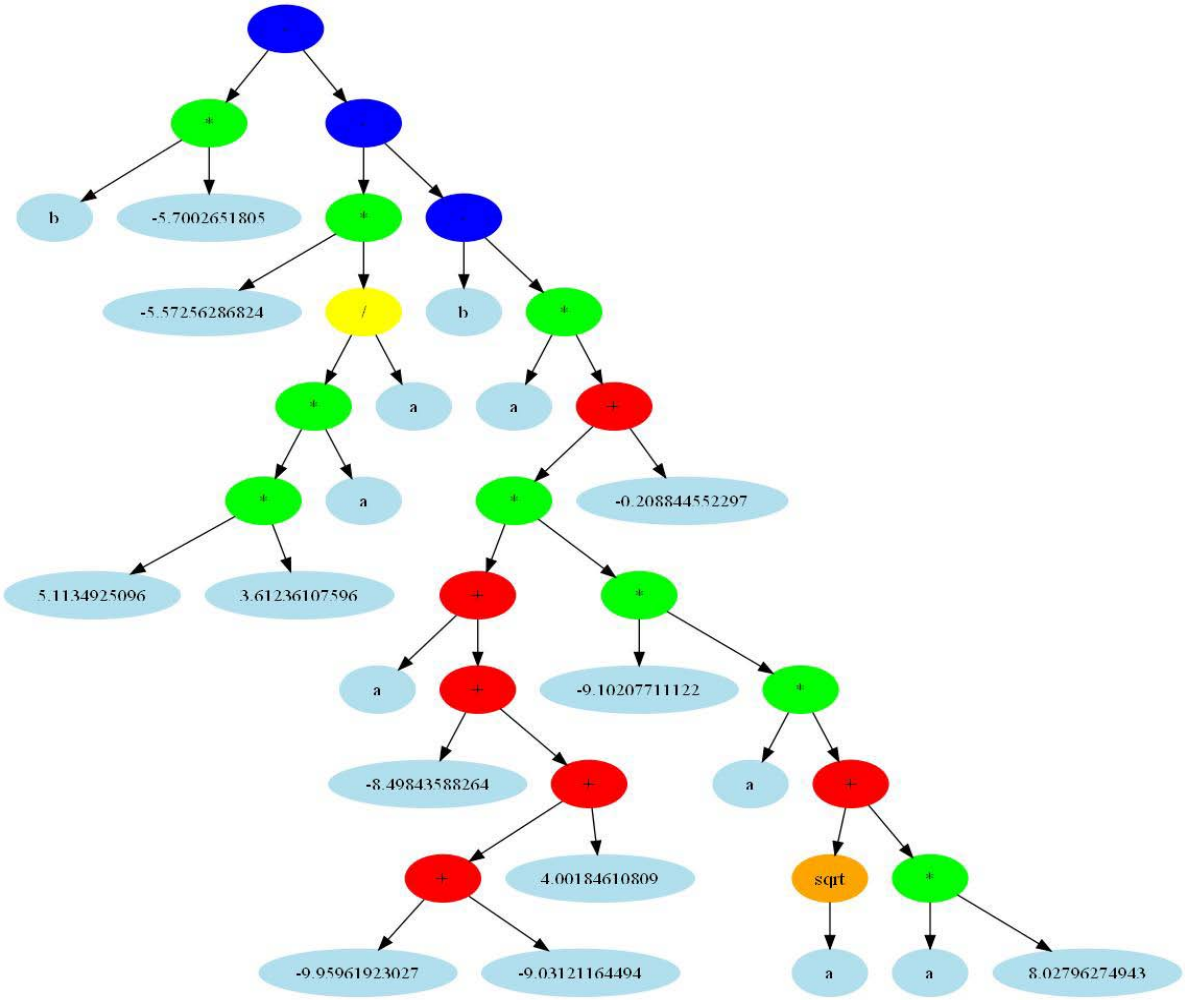


**Fig. 3.** Best individual

The genetic cost model has a complex form given by the mathematical operators and has a high flexibility level owed to the used evolution model. The model is open for improvement given that using parallel processing better results can be obtained in a shorter time period.

## 5 Genetic model testing

In order to test the genetic cost model the following elements are under analysis:

- sensitivity is represented by the presence of variations in the dependent variables as a result of small or high variations in the independent variables;
- the non-compensatory character is given by obtaining different results due to variation in the values of the values associated to the independent variables;
- the non-catastrophic character of the model is represented by the lack of situations that make it impossible to obtain a value for the resultant variable.

**Table 4.** Variation of the cxm variable in the genetic model

| rbc | cxm | CGS |
|---|---|---|
| 3 | 2 | 11.53022 |
| 3 | 1 | 8.12132 |
| 3 | 2 | 11.53022 |
| 3 | 9 | 38.64562 |
| 3 | 8 | 34.67918 |
| 3 | 2 | 11.53022 |
| 3 | 7 | 30.72567 |
| 3 | 8 | 34.67918 |
| 3 | 2 | 11.53022 |
| 3 | 3 | 15.21114 |
| 3 | 4 | 19.01308 |

The data associated to the testing of the genetic model are automatically generated.

The data for the b variable belong to the [1,15] interval and for the a parameter are located in the [0,10] interval.
The sensitivity of the genetic security cost model is defined through the mathematical expression of the GSC indicator is analyzed considering the variation of the dependent variable cxm according to the data in Table 4.

The dependent variable presents variations correlated with the ones of module complexity according to the graphical representation in Figure 4.
The sensitivity analysis of the genetic cost model based on the variation in the behavioral risk rating represented by the rbc variable is determined based on the data presented in Table 5.

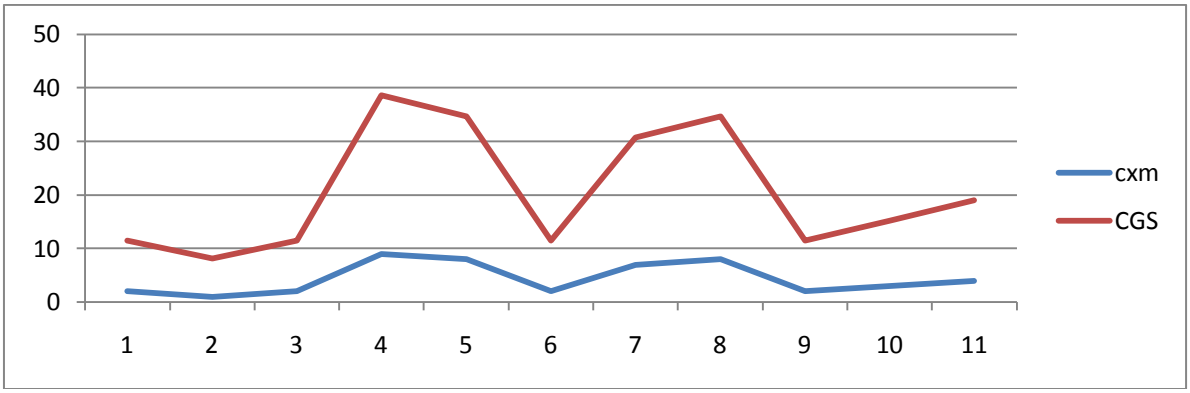


**Fig. 4.** The correlation of variations in complexity - genetic cost

**Table 5.** The variation of the rbc variable in the genetic cost model

| rbc | cxm | CGS |
|---|---|---|
| 2 | 7 | 30.31577 |
| 1 | 7 | 29.94686 |
| 9 | 7 | 33.76841 |
| 1 | 7 | 29.94686 |
| 1 | 7 | 29.94686 |
| 0 | 7 | 29.62658 |
| 9 | 7 | 33.76841 |
| 4 | 7 | 31.17039 |
| 8 | 7 | 33.20949 |
| 6 | 7 | 32.14512 |
| 2 | 7 | 30.31577 |

The values obtained for the resultant variable after applying the genetic cost model expression to the data in Table 5 are correlated with the behavioral risk levels as presented in the graphical representation in Figure 5.
The CGS indicator presents a sensitive and non compensatory behavior determined by the variations recorded compared to the modifications in the module complexity and behavioral risk rating.

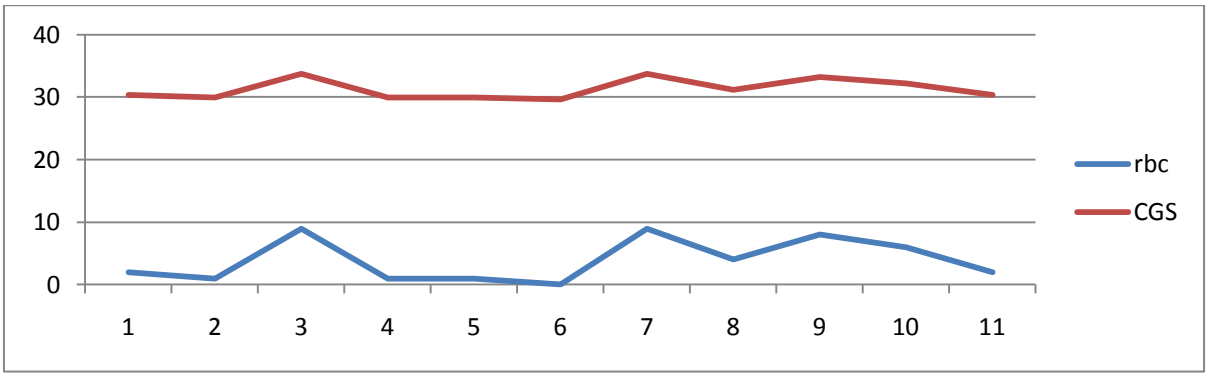


**Fig. 5.** The correlation of variations in behavioral risk - genetic cost

The non-catastrophic character is given by the lack of situations in which the cxm and rbc variables belonging to the specified intervals lead to the impossibility of obtaining a result using the mathematical expression of the genetic security cost model.

## 6 Genetic cost model validation

The data used for the genetic cost model validation are obtained through access frequency analysis for the AVIO application after optimizing the validation and orthogonality management classes that imply security measures. Table 6 presents the data obtained after monitoring forty unique user activity logs in the AVIO application.

**Table 6.** AVIO results after the optimization process

| Result | No. results | Percentage results | Risk rating | rbc | cxm | CS |
|---|---|---|---|---|---|---|
| LoadImage1 | 37 | 92.50% | 10 | 9.25 | 4 | 21 |
| LoadImage2 | 37 | 92.50% | 10 | 9.25 | 4 | 22 |
| DisplayImage1 | 27 | 67.50% | 5 | 3.375 | 2 | 10 |
| DisplayImage2 | 27 | 67.50% | 5 | 3.375 | 2 | 10 |
| ComputeOrthogonality | 40 | 100.00% | 3 | 3 | 6 | 25 |
| InterfaceAssesment | 27 | 67.50% | 6 | 4.05 | 3 | 15 |
| OrthogonalityAssesment | 24 | 60.00% | 6 | 3.6 | 3 | 15 |
| SaveData | 37 | 92.50% | 10 | 9.25 | 3 | 20 |
| SystemError | 20 | 50.00% | 10 | 5 | 2 | 13 |
| OrthogonalityComputationWithoutImage1 | 1 | 2.50% | 4 | 0.1 | 1 | 5 |
| OrthogonalityComputationWithoutImage2 | 2 | 5.00% | 4 | 0.2 | 1 | 5 |
| OrthogonalityComputationWithoutImages | 2 | 5.00% | 4 | 0.2 | 1 | 5 |
| AssesmentWithoutFunctionality | 8 | 20.00% | 6 | 1.2 | 2 | 10 |

In Figure 6 the number of completed access attempts for each result associated to the security system of the AVIO product.

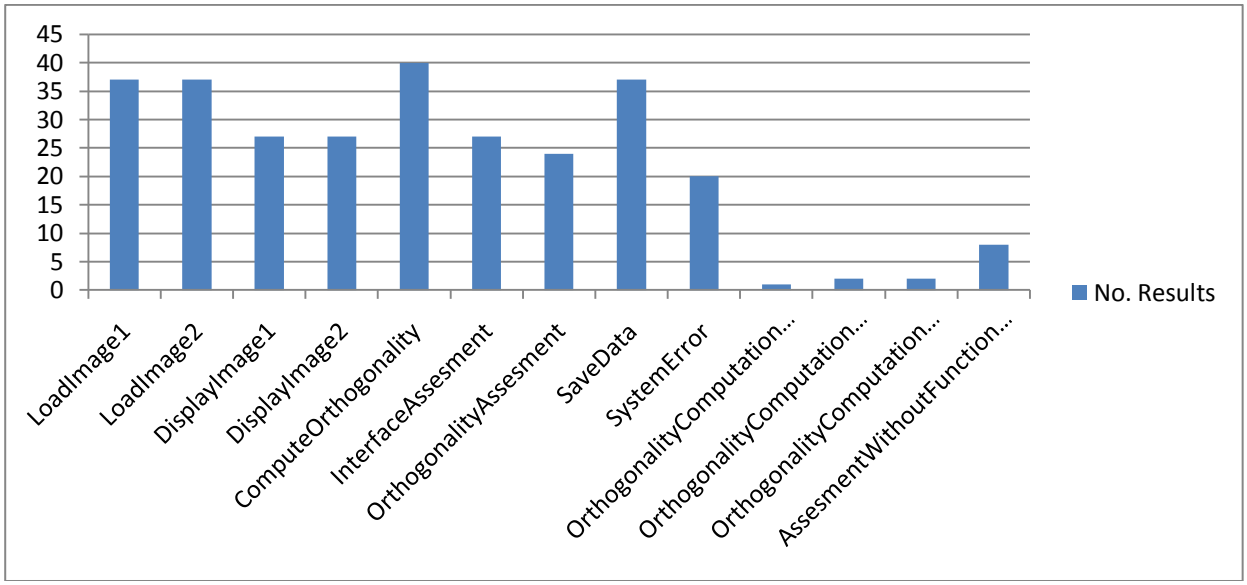


**Fig. 6.** Repartition of result activation frequency after optimization

The data obtained after finishing the improvement process of the AVIO software are suited for validating the genetic cost model because:

- it is not related to the model evolution phase of the genetic algorithm which evolved the security cost model;
- it originates from a different environment than the one that was used to determine the data which was the foundation of the genetic model.

The validation of the genetic model is based on determining the estimation error for the CGS indicator as presented in Table 7, according to expected results and those obtained using the genetic model for security cost.

**Table 7.** The results of the GCS indicator validation

| **Result** | **rbc** | **cxm** | **CS** | **CGS** | **Error** |
|---|---|---|---|---|---|
| LoadImage1 | 9.25 | 4 | 21 | 23.07597 | 2.07597 |
| LoadImage2 | 9.25 | 4 | 22 | 23.07597 | 1.07597 |
| DisplayImage1 | 3.375 | 2 | 10 | 11.82691 | 1.82691 |
| DisplayImage2 | 3.375 | 2 | 10 | 11.82691 | 1.82691 |
| ComputeOrthogonality | 3 | 6 | 25 | 26.79031 | 1.79031 |
| InterfaceAssesment | 4.05 | 3 | 15 | 15.93065 | 0.93065 |
| OrthogonalityAssesment | 3.6 | 3 | 15 | 15.61985 | 0.61985 |
| SaveData | 9.25 | 3 | 20 | 19.68347 | 0.31653 |
| SystemError | 5 | 2 | 13 | 13.11751 | 0.11751 |
| OrthogonalityComputationWithoutImage1 | 0.1 | 1 | 5 | 5.11952 | 0.11952 |
| OrthogonalityComputationWithoutImage2 | 0.2 | 1 | 5 | 5.236866 | 0.23687 |
| OrthogonalityComputationWithoutImages | 0.2 | 1 | 5 | 5.236866 | 0.23687 |
| AssesmentWithoutFunctionality | 1.2 | 2 | 10 | 10.11707 | 0.11707 |

For the results obtained after estimating the information security cost through genetic algorithms an average error of 0.869 man hours is obtained. The maximum error is 2.1 man hours and the minimum error is 0.117 man hours.

Figure 7 presents the error repartition after analyzing the results obtained from validating the genetic cost model.

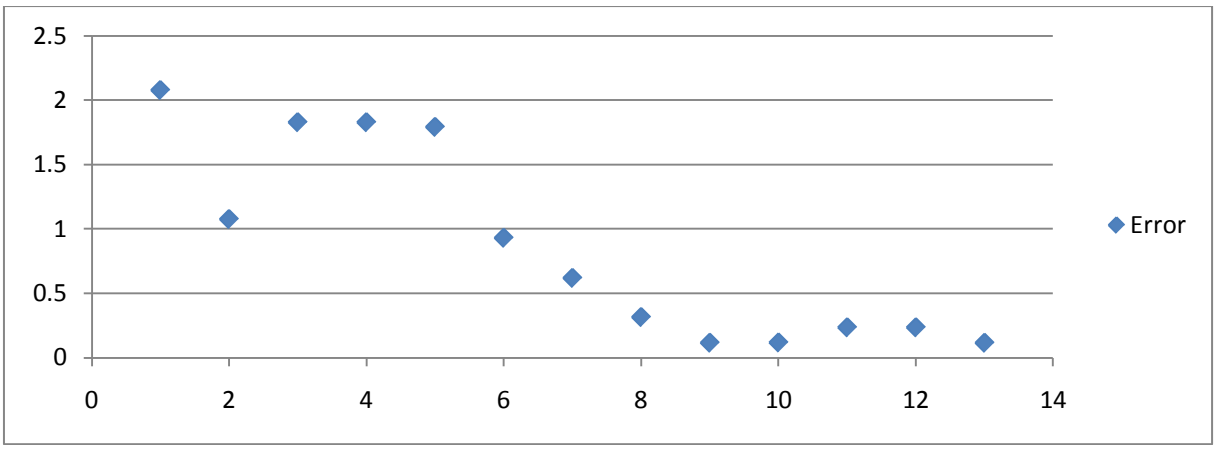


**Fig. 7.** Error repartition for the genetic model validation

The errors follow a balanced repartition without high deviations from the average value. This repartition is relevant for the good correlation between expected cost

levels and the obtained ones according to the graphical representation in Figure 8.

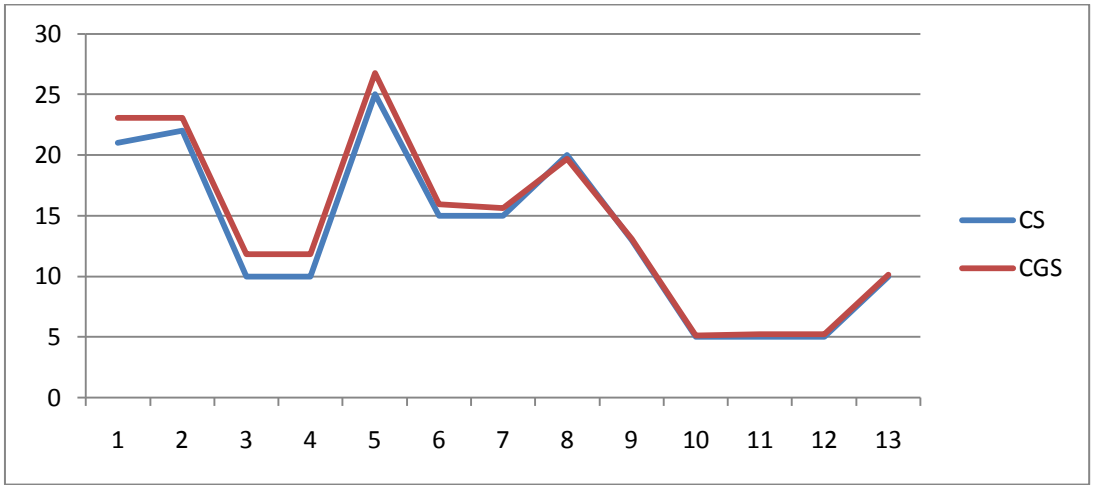


**Fig. 8.** Correlation of the expected and obtained levels for the genetic cost model

The average difference of 0.869 man hours obtained by applying the genetic cost model compared to the expected result is acceptable making the genetic cost model valid for use in estimating similar security costs.

## 7 Conclusions

The security for structured content validation is ensured through input validation, connection encryption, access control, file integrity, user activity monitoring and load balancing. The AVIO application implements all of the above resulting a balanced security system customized for the activities specific to it.

Information security cost knowledge is needed to determine the suitability of a security system investment. Information security cost analysis is determined by:

- finding the factors in the security systems' operation environment that have an impact on costs;
- determine the information security risks involved by the operation of the distributed system under analysis according to [5];
- correlate the cost factors with the risks to identify the cost hierarchy of the most cost vulnerable components.

The development of cost models using genetic algorithms provided an efficient cost computation model for future developments of security components according to their complexity, frequency of use and asset risk rating.

## Acknowledgements

This article is a result of the project „Doctoral Program and PhD Students in the education research and innovation triangle". This project is co funded by European Social Fund through The Sectorial Operational Programme for Human Resources Development 2007-2013, coordinated by The Bucharest Academy of Economic Studies.

## References

[1] I. Ivan and D. Palaghita, "The Informatics Security Cost of Distributed Applications," *Theoretical and Applied Economics*, pp. 49-68, 2010.

[2] D. Palaghita, "Open source procedures for image orthogonality analysis," *Open Source Science Journal*, pp. 82-105, 2010.

[3] Wikipedia. [Online]. Available at: http://en.wikipedia.org/wiki/Color_model#Color_systems

[4] Z. Wang, H. R. Sheikh, and B. C. A., "No-reference perceptual quality assessment of JPEG compressed images," in *IEEE International*

*Confonference Image on Processing*, 2002, pp. 477-480.

[5] L. Cotfas, D. Palaghita, and B. Vintila, "Audit Techniques for Service Oriented Architecture Applications," *Informatics Economics*, pp. 128-136, 2010.

**Dragos PALAGHITA** graduated from the Academy of Economic Studies of Bucharest, Cybernetics Statistics and Economic Informatics faculty, Economic Informatics section in 2008. He is programming in C++ and C# and his main areas of interest are Informatics Security, Software Quality Management, large data set analysis and graphical representation enhancements. Currently he is undergoing PhD studies at the Academy of Economic Studies of Bucharest, Cybernetics Statistics and Economic Informatics.