



COMPUTER SCIENCE AND EDUCATION IN COMPUTER SCIENCE (CSECS)

14th ANNUAL INTERNATIONAL CONFERENCE

June 29th - 30th, 2018

Boston, USA



Hochschule Fulda
University of Applied Sciences



COMPUTER SCIENCE AND EDUCATION IN COMPUTER SCIENCE (CSECS)

14th ANNUAL INTERNATIONAL CONFERENCE

With financial support of the Central Strategic Development Found of
New Bulgarian University

With financial support of the EU ASPIres Project

June 29th - 30th , 2018

Boston, USA

Chairmen: Ivan Landjev (Bulgaria), Peter Peinl
(Germany), Lou Chitkushev (USA)

General Secretaries: Petya Asenova (Bulgaria), Guanglan
Zhang (USA)



Hochschule Fulda
University of Applied Sciences



Editors: Petya Assenova, Guanglan Zhang, Peter Peinl

Copyright © 2018

New Bulgarian University 21 Montevideo Str.,
1618 Sofia, Bulgaria

University of Applied Sciences Fulda 123 Leipziger Str.,
36037 Fulda, Germany

Boston University, MET
808 Commonwealth Avenue,
02215 Boston, USA

ISSN 2603-4794



Conference Schedule

June 29 – June 30, 2018

Location: Room 109, 808 Commonwealth Ave, Boston, MA, USA

All times are shown in Eastern Daylight Time.

Thursday, June 28, 2018

| | |
|-------------------|---|
| 2:00 to 5:00 p.m. | <i>Registration & Housing Information</i> <i>Location: 808 Commonwealth Avenue, Room 250</i> |
|-------------------|---|

Friday, June 29, 2018

| | |
|---------|---|
| 9:00am | <i>Breakfast & Registration</i> <i>Location: 808 Commonwealth Avenue, Room 109</i> |
| 9:30am | Opening by Tanya Zlateva, Dean, MET, BU |
| 9:50am | Title: The Nonexistence of Linear Codes with Parameters Presenter Asia Ruseva, Sofia University |
| 10:10am | Title: Teaching Ellipse with CAS Presenter: Petya Asenova, New Bulgarian University (Online) |
| 10:30am | Title: Foreign Direct Investment Net Inflows: Data-Driven Analysis Presenter: Dimitar Trajanov, Ss. Cyril and Methodius University, Macedonia (Online) |
| 10:50am | <i>Short Break</i> |
| 11:10am | Title: Software System Aesthetics and Readability Metrics Presenter: Latchezar Tomov, New Bulgarian University (Online) |
| 11:30am | Title: Aesthetic Metrics and the Evolution of C Language Family Presenter: Latchezar Tomov, New Bulgarian University (Online) |
| 11:50am | Title: Real-Time Comparison of Movement in Bulgarian Folk Dances |

| | |
|---------|---|
| | Presenter: Zlatka Uzunova, New Bulgarian University (Online) |
| 12:10pm | <i>Lunch Break</i> |
| 1:30pm | Title: Simulation and Evaluation of Scenarios in a Gas Station Using Simul8 Software Presenter: Denis Ramos de Oliveira, Universidade Federal de Viçosa |
| 1:50pm | Title: Geographically Weighted Regression in the Analysis of the Development of Information and Communication Technology in Indonesia Presenter: Dwi Puspita Sari, Boston University |
| 2:10pm | Title: Case Study: Power BI Visualizations Applied to Digital Publisher's Advertising Inventory Presenter: Joseph Chomski, Boston University |
| 2:30pm | Title: An Online Platform for Study of Effects of Violence and Traumatic Events Presenter: Guanglan Zhang, Boston University |
| 2:50pm | <i>Afternoon Break</i> |
| 3:10pm | Title: Striking the Balance: Teaching Data Mining with the Right Mixture of Depth and Breadth Presenter: Vladimir Zlatev, Boston University |
| 3:30pm | Title: A Framework for Modeling in Scale: an Introduction Presenter: Eric Braude, Boston University |
| 3:50pm | Title: On Sperner's Theorem for Modules over Finite Chain Rings Presenter: Ivan Landjev, New Bulgarian University |
| 4:10pm | Title: Introducing Fundamental Agile Concepts in Project Management and Software Development Courses Presenter: Vijay Kanabar, Boston University |
| 4:30pm | <i>End of Day 1</i> |

Saturday, June 30, 2018

| | |
|---------|--|
| 9:00am | <i>Breakfast</i> <i>Location: 808 Commonwealth Avenue, Room 109</i> |
| 9:30am | Title: Blockchain Solution for Annual Evaluation in Bulgarian Schools Presenter: Delyan Keremedchiev, New Bulgarian University (Online) |
| 9:50am | Title: Test Design Process Improvement by Six Sigma (Dmaic) and R Presenter: Dobromir Dinev, New Bulgarian University (Online) |
| 10:10am | Title: Introduction of Bell-Lancaster Method and Learning by Doing into the Practical Curriculum at Undergraduate and Graduate Levels Presenter: Valentina Ivanova, New Bulgarian University (Online) |
| 10:30am | Title: Representative Sample as a LP Problem Presenter: Dimitar Atanasov, New Bulgarian University (Online) |
| 10:50am | <i>Short Break</i> |
| 11:10am | Title: Generation of Virtual Annotated Corpora Presenter: Mariyana Raykova, New Bulgarian University (Online) |
| 11:30am | Title: How to Improve Teaching in Discrete Mathematics Via Programming Presenter: Mariyana Raykova, New Bulgarian University (Online) |
| 11:50am | Title: Early Detection Of Forest Fires - Standard Interfaces and Protocols at Sensor Network and Cloud Level Definition Presenter: Katerina Zlatanovska, Ministry of Defence, Macedonia (Online) |
| 12:10pm | <i>Lunch Break</i> |
| 1:30pm | Title: Don't be afraid to commit Presenter: Andrew Wolfe, Boston University (Online) |
| 1:50pm | Title: Seeing the Staircase: Reflections on a First Semester Teaching Data Mining to Business Students Presenter: Greg Page, Boston University |
| 2:10pm | Title: Teaching Data Mining Techniques to Applied Business Analytics Students with the Help of Interactive Hands-on Tutorials |

| | |
|---------|---|
| | Presenter: Penko Ivanov, , New Bulgarian University |
| 2:30pm | Title: Data Analytics for Devops Effectiveness Presenter: Penko Ivanov, New Bulgarian University |
| 2:50 pm | Title: End-User Application for Early Forest Fire Detection and Prevention Presenter: Peter Peinl, University of Applied Science |
| 3:10pm | <i>Afternoon Break</i> |
| 3:30pm | Panel Discussions |
| 4:30pm | <i>End of Conference</i> |

Content

| | | |
|-----|--|-----|
| 1. | Asia Ruseva (BG), Ivan Landjev (BG), The Nonexistence of Linear Codes with Parameters [2024,4,162] over GF (5) | 1 |
| 2. | Marin Marinov (BG), Petya Asenova (BG), Teaching Ellipse with CAS | 11 |
| 3. | Ana Gjorgjevikj (MK), Kostadin Mishev (MK), Irena Vodenska (USA), Lubomir Chitkushev (USA), Dimitar Trajanov (MK), Foreign Direct Investment Net Inflows: Data-Driven Analysis | 23 |
| 4. | Latchezar Tomov (BG) Software System Readability Metrics | 61 |
| 5. | Latchezar Tomov (BG), Aesthetic Metrics and the Evolution of C Language Family | 77 |
| 6. | Zlatka Uzunova (BG), Real-Time Comparison of Movement in Bulgarian Folk Dances | 93 |
| 7. | Denis Ramos de Oliveira (BR), João Pedro Fonseca de Barcelos (BR), Thiago Henrique Nogueira (BR), Simulation and Evaluation of Scenarios in a Gas Station Using Simul8 Software | 109 |
| 8. | Dwi Puspita Sari (USA), Jamilatuzzahro (ID), Vijay Kanabar (USA), Geographically Weighted Regression in Analysis of of Information and Communication Technology Development in Indonesia | 129 |
| 9. | Joseph Chomski (USA), Case Study: Power BI Visualizations Applied to Digital Video Publisher's Advertising Inventory | 139 |
| 10. | Yuting Zhang (USA), An Online Platform for Project Based Learning - a Proposal | 147 |
| 11. | Gregory Page (USA), Slav Angelov (BG), Penko Ivanov (BG), Vladimir Zlatev (USA), Striking the Balance: Teaching Data Mining with the Right Mixture of Depth and Breadth | 153 |
| 12. | Eric Braude (USA), A Framework for Modeling in Scale: an Introduction | 163 |
| 13. | Ivan Landjev (BG), On Sperner's Theorem | 167 |
| 14. | Vijay Kanabar (USA), Kalinka Kalaoyanova (BG), Introducing Agile Concepts in Project Management and Software Development Courses | 173 |

| | | |
|-----|---|-----|
| 15. | Dobromir Dinev (BG), Test design process improvement by Six Sigma (DMAIC) and R | 181 |
| 16. | Mariyana Raykova (BG), Valentina Ivanova (BG), Hristina Kostadinova (BG), Generation of Virtual Annotated Corpora | 193 |
| 17. | Mariyana Raykova (BG), Stoyan Boev (BG), How to Improve Teaching in Discrete Mathematics Via Programming and Vice Versa | 211 |
| 18. | Jugoslav Achkoski (MK), Nikola Kletnikov (MK), Nevena Serafimova (MK), Igorce Karafilovski (MK), Rossitza Goleva (BG), Katerina Zlatanovska (MK), Early Detection of Forest Fires - Standard Interfaces and Protocols at Sensor Network and Cloud Level Definition | 229 |
| 19. | Penko Ivanov (BG), Teaching Data Mining Techniques to Applied Business Analytics Students with the Help of Interactive Hands-on Tutorials | 251 |
| 20. | Alexandrina Ivanova (BG), Penko Ivanov (BG), Data Analytics for Devops Effectiveness | 271 |
| 21. | Peter Peinl (DE), Micha Heiderich (DE), Ivan Christov (DE), Jugoslav Achkoski (MK), Nikola Kletnikov (MK), Igorche Karafilovski (MK), Nikola Manev (MK), Rossitsa Goleva (BG), Alexander Savov (BG), Ivelin Andreev (BG), End-user Application for Early Forest Fire Detection and Prevention | 299 |
| 22. | Delyan Keremedchiev (BG), Juliana Peneva (BG), Blockchain Solution for Annual Evaluation in Bulgarian Schools | 319 |
| 23. | Valentina Ivanova (BG), Mariyana Raykova (BG), Introduction of Bell-Lancaster Method and Learning by Doing into the Practical Curriculum at Undergraduate and Graduate Levels | 321 |
| 24. | Dimitar Atanasov (BG), Representative Sample as a LP Problem | 323 |
| 25. | Andrew Wolfe (USA), Don't be afraid to Commit Experiences Using GitHub Classroom for Teaching CS | 325 |
| 26. | Greg Page, Seeing the Staircase: Reflections on a First Semester Teaching Data Mining to Business Students | 327 |
| 27. | Dimitar Trajanov, Ivana Trajanovska, Lubomir Chitkushev, Irena Vodenska, Using Google BigQuery for Data Analytics in Research and Education | 329 |

CSECS 2018, pp. 001 - 009

Computer Science and Education in Computer Science
14th Annual International Conference
ISSN 2603-4794

June 29 – 30, 2018, Boston, USA

THE NONEXISTENCE OF LINEAR CODES WITH PARAMETERS [204,4,162] OVER GF(5)

Assia Rousseva (1), Ivan Landjev (2)

(1) Sofia University, Department of Geometry

(2) New Bulgarian University, Department of Informatics

***Abstract** The main problem of coding theory for four dimensional codes over the field with five elements is solved for all but three values of d : $d=81,161,162$. In this talk we announce the nonexistence of linear codes over $GF(5)$ with parameters $[204,4,162]$. This problem is tackled from its geometric side. The existence of a code with these parameters is equivalent to that of a $(204,42)$ -arc in $PG(3,5)$. In order to rule out the existence of such arcs we consider a special dual arc which exhibits very strong divisibility properties. We prove that such an arc must have a hyperplane without 0-points which in turn implies the extendability of every $(204,42)$ -arc to an arc with parameters $(205,42)$. The nonexistence of the latter rules out the existence of $(204,42)$ -arcs in $PG(3,5)$. This result implies the exact value $n_5(4,162)=205$, where $n_q(k,d)$ denotes the shortest length of a linear code of fixed dimension k and fixed minimum distance d over the field with q elements.*

***Keywords:** linear codes over finite fields, Griesmer bound, optimal codes, Griesmer arcs, finite projective geometries, dual arcs, $(t \bmod q)$ arcs, lifted arcs*

1. INTRODUCTION

One of the central problems in coding theory is to optimize one of the three main parameters of a linear code given the other two. The most popular version of this problem is to find the exact value of $n_q(k,d)$ defined as the optimal length of a linear code of dimension k and minimum distance d over $\text{GF}(q)$. It has been solved in the following cases: for $q=2$, $k \leq 8$ for all d , for $q=3$, $k \leq 5$ for all d , for $q=4$, $k \leq 4$ for all d , for $q=5,7,8,9$, $k \leq 3$ for all d . In the case of for $q=5$, $k=4$ there exist three values of d for which $n_q(k,d)$ is not known: $d=81,161,162$.

In this note we present a sketch of proof for the nonexistence of arcs with parameters $(204,42)$ in $\text{PG}(3,5)$, or, equivalently, of the nonexistence of linear $[204,4,162]$ -codes over $\text{GF}(5)$. This implies the exact value $n_5(4,162)=205$.

2. BASIC FACTS

Let P be the set of points of the projective geometry $\text{PG}(k-1,q)$. Every mapping $K:P \rightarrow N_0$ is called a multiset in $\text{PG}(k-1,q)$. This mapping is extended additively to the subsets of P . The integer $n=K(P)$ is called the cardinality of K . The support of K is the set of all points of positive multiplicity. In order to save space we refer for all the basic notions not defined here to [LR16].

Let K be an (n,w) -arc in $\text{PG}(k-1,q)$ with spectrum (a_i) . We denote by λ_j the number of points in the geometry of multiplicity j . Simple counting arguments yield the following identities that are equivalent to the first three MacWilliams identities:

$$(1) \quad \sum_{i=0}^w a_i = \frac{q^k - 1}{q - 1},$$

$$(2) \quad \sum_{i=0}^w i a_i = n \frac{q^{k-1}-1}{q-1},$$

$$(3) \quad \sum_{i=0}^w \binom{i}{2} a_i = \binom{n}{2} \frac{q^{k-2}-1}{q-1} + q^{k-2} \cdot \sum_{i \geq 2} \binom{i}{2} \lambda_j.$$

Set $v_k = (q^k - 1)/(q - 1)$. From the above three identities we can deduce that

$$(4) \quad \sum_{i=0}^w \binom{w-i}{2} a_i = \binom{w}{2} v_k - n(w-1)v_{k-1} + \binom{n}{2} v_{k-2} + q^{k-2} \sum_{i \geq 2} \binom{i}{2} \lambda_i.$$

Let us note that the sum on the left can be written as $\sum_H \binom{w-K(H)}{2}$, where the sum is taken over all hyperplanes. Fix a hyperplane H_0 . For a fixed hyperline S denote by H_1, \dots, H_q the remaining hyperplanes through S . Set

$$\eta_i = \max_{S:K(S)=i} \sum_{j=1}^q \binom{w-K(H_j)}{2}.$$

Here the maximum is taken over all hyperplanes S of multiplicity i contained in H_0 . Assume that the spectrum (b_i) of the restriction of K to H_0 is known. Thus we obtain the estimate

$$(5) \quad \sum_H \binom{w-K(H)}{2} \leq \sum_j b_j \eta_j + \binom{w-K(H_0)}{2}.$$

Now from (4) and (5) and plugging in our parameters $n = 204, w = 42$ we get the inequality

$$(6) \quad -732 + 25\lambda_2 \leq \sum_j b_j \eta_j + \binom{42-K(H_0)}{2}.$$

This inequality will be used repeatedly in our nonexistence proof.

3. SOME STRUCTURE RESULTS

In this section we state some structure results on arcs in the projective geometries $\text{PG}(2,5)$ and $\text{PG}(3,5)$ that are used in our nonexistence proof. They are proved by a bit lengthy but straightforward arguments. If K is a (n,w) arc in $\text{PG}(k-1,q)$, we denote by γ_i the maximal multiplicity of an i -dimensional subspace of $\text{PG}(k-1,q)$.

Lemma 1. Assume there exists a $(204,42)$ -arc K in $\text{PG}(3,5)$. Then we have:

- a) $\gamma_0 = 2, \gamma_1 = 9, \gamma_2 = 42$;
- b) the possible multiplicities of the planes with respect to K are: $29,30,31; 34,35,36, 39,40,41,42$;
- c) if H is a plane the the restriction of K to H is one of the following:
 - the complement of a triangle and two further points; in this case $\lambda_0 = 3,4,5, \lambda_2 = 14,15,16$;
 - arcs with $\lambda_0 = 1, \lambda_2 = 12$;
 - the sum of an $(11,3)$ -arc and the plane H which are arcs with $\lambda_0 = 0, \lambda_2 = 11$.

Let us note that the arcs in the first class are extendable while the arcs in the other two classes are not. It is also important to note that in the arcs in the second class have four collinear points that meet in the single 0-point. An easy consequence of Lemma 1 is that the arc K is 3-quasidivisible (cf. [LRS16]). Hence the dual arc K' defined by $K'(H)=3+K(H) \pmod{5}$ is $(3 \pmod{5})$ arc with point multiplicities $01,2,3$. Since K is not extendable the dual arc K' does not contain a full plane in its support.

Generally speaking the size of K' cannot be determined from the parameters of K . However we can get some information on the structure of the $(3 \pmod{5})$ -arcs in $\text{PG}(3,5)$ from the existing classification of the small $(3 \pmod{5})$ -arcs in $\text{PG}(2,5)$ completed in [R15]. First we state the classification of the plane arcs.

Lemma 2. Let F be a $(3 \bmod 5)$ arc in $\text{PG}(2,5)$ of size not exceeding 33. Then F is one of the following:

- if $|F|=18$, then F is the sum of three not necessarily distinct lines;
- if $|F|=23$, then F is the dual of the projective triangle; in this case $\lambda_3 = 3, \lambda_2 = 4, \lambda_1 = 6$, ; the 2-points form a quadrangle, the 3-points are the diagonal points and the 1-points are the intersections of the sides of the diagonal triangle with the sides of the quadrangle;
- if $|F|=28$, then $\lambda_3 = 6, \lambda_1 = 10$, ; the 3-points form an oval and the 1-points are the internal points to this oval;
- if $|F|=33$, then it is one of the following:
 - the dual of the complement of an $(10,3)$ -arc,
 - the dual of the complement of an $(11,3)$ -arc with four external lines and one point doubled;
 - an oval of five 2-points and one 0-point, the tangent at the 0-point is a 13-line with four further 3-points;
 - the mod 5 sum of three lines of respective multiplicities 3, 3, and 2.

This classification enables us to prove the following important result for $(3 \bmod 5)$ arcs in $\text{PG}(3,5)$.

Lemma 3. Let F be a $(3 \bmod 5)$ arc in $\text{PG}(3,5)$ with maximal point multiplicity 3 and cardinality not exceeding 168. Then F is lifted from a single 3-point. In particular, a $(3 \bmod 5)$ -arc with $|F| \leq 168$ has cardinality 93, 118, 143, or 168 and is obtained by lifting one of the plane $(3 \bmod 5)$ -arcs in Lemma 2 from a single 3 point.

Lemma 1 and Lemma 3 imply the important corollary that every $(3 \bmod 5)$ -arc K' , obtained by dualizing a $(204,42)$ -arc K , has size at least 173.

Corollary 4. Let K be a $(204,42)$ -arc in $\text{PG}(3,5)$ and let K' be its dual defined in this section. Then $|K'| \geq 173$.

Proof. Let us start by observing that some of the arcs K' obtained by Lemma 3 have a hyperplane without 0-points and hence come from extendable $(204,42)$ -arcs. Since there exist no $(205,42)$ -arcs, we get a contradiction. This argument rules out all arcs obtained by lifting a plane arc with a line without 0-points. These are the arcs of cardinality 18, all but two arcs of the first family of cardinality 33, and the second family of cardinality 33.

All the remaining arcs are ruled out by a more elaborate structure argument. Essential is the fact that a 0-point in the dual arc can come only from a maximal plane with respect to K (i.e. a plane of multiplicity 42). Thus if K' has a plane of multiplicity 18 which consists of a triple line (a line consisting of six 3-points) then a straightforward double counting gives that this plane is the image of a 2-point and the image of the 3-line is a 4-line incident with planes of multiplicities $39, 39, 39, 39, 39, 29$ or $39, 39, 39, 39, 34, 34$.

Now we give a detailed proof of the impossibility of the last arc in the list in Lemma 3. So assume $|K'| = 168$ and it is obtained by lifting the arc in question from a single 3-point P . Then the lifting point is the image of a 29, 34, or 39-plane. Now P is incident with a line of type $3, 0, 0, 0, 0, 0$, and hence is the image of a 39-plane H . (Recall that 0-points come from 42-planes.) The line of type $(3, 0, 0, 0, 0, 0)$ is obviously a 9-line. On the other hand, a line of type $(3, 3, 3, 3, 3, 3)$ through P is a 4-line (by our previous remark). So the 39-plane H has a point incident with one 9-line and five 4-lines. Now this is impossible since $39 > 9 + 4 + 4 + 4 + 4 + 4$.

4. THE PROOF

Lemma 5. Let K be a $(204,42)$ -arc in $\text{PG}(3,5)$ and let H be a maximal plane. Then the restriction of K to H is not extendable (i.e. it has $\lambda_0 = 0$ or 1).

Proof. Let H_0 be a plane of maximal multiplicity. So, K_H is a (42,9)-arc. By Lemma 1c such an arc is obtained by deleting a triangle and two points from two copies of the plane PG(2,5). The possible spectra of such arcs are:

- 1) $a_9=18, a_8=10, a_4=2, a_2=1$;
- 2) $a_9=19, a_8=8, a_7=1, a_4=1, a_3=2$;
- 3) $a_9=18, a_8=9, a_7=1, a_4=2, a_3=1$;
- 4) $a_9=18, a_8=10, a_7=1, a_4=3$;
- 5) $a_9=22, a_7=6, a_4=3$.

Now we fix a line L in H_0 and consider the possible contributions of the planes through L (different from H_0) to the right-hand side of (6). In the same time we take into account the multiplicity of the line corresponding to L in the dual plane. These contributions are given in the following table

| $K(L) \setminus K'(L)$ | 3 | 8 | 13 |
|------------------------|-----|-----|-----|
| 9 | 3 | No | No |
| 8 | 28 | 7 | No |
| 7 | 78 | 32 | 12 |
| 6 | 81 | 82 | 37 |
| 5 | 121 | 157 | 87 |
| 4 | No | 156 | 112 |
| 3 | No | 156 | 162 |
| 2 | No | 211 | 187 |

Consider the spectrum 1). From (6) we get

$$-732 + 25\lambda_2 \leq 18.3 + 10.7 + 1.112 + 2.187,$$

whence $\lambda_2=53$ and all 0-ponts are in H_0 . This possibility is easily ruled out by the fact that there is no (42,9)-arc with two double points. The remaining spectra are treated similarly.

In a similar way, we can rule out the impossibility of maximal planes with one 0-point. This result is stated without proof in the lemma below.

Lemma 6. Let K be a (204,42)-arc in PG(3,5) and let H be a maximal plane. Then the restriction of K to H is not a (42,9)-arc with $\lambda_0=1$.

Theorem 7. There exists no $(204,42)$ -arc in $\text{PG}(3,5)$. Equivalently, there exists no $[204,4,162]$ -code over $\text{GF}(5)$.

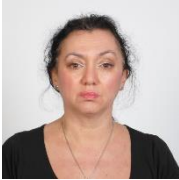
Proof. Assume K is a $(204,42)$ -arc in $\text{PG}(3,5)$. It has 48 double points. On the other hand it is well known that the maximal size of a 3-cap in $\text{PG}(3,5)$ is 43 [EL10]. Hence there exist four collinear double points and hence a line of multiplicity 10. This contradicts Lemma 1a.

Acknowledgements. The research of the first author was sponsored by the Scientific Research Fund of Sofia University under Contract 80-10-51/17.04.2018. The research of the second author was supported by the Strategic Development Fund of the New Bulgarian University.

REFERENCES

- [EL10] Y. Edel, I. Landjev, On multiple caps, Designs, Codes and Cryptography 56(2010), 163-175.
- [LR16] I. Landjev, A. Rousseva, The non-existence of $(104,22;3,5)$ -arcs, Advances in Math. Of Communication 10(3)(2016), 601-611.
- [LRS16] I. Landjev, A. Rousseva, L. Storme, On the extendability of quasidivisible Griesmer arcs, Designs Codes and Cryptogr. 79(2016), 535-547.
- [R16] A. Rousseva, On the structure of $(t \bmod q)$ -arcs in finite projective geometries, Annuaire de l'Universite de Sofia 103(2016), 5-22.

Author's Information



Assia Rousseva, PhD
Sofia University, Department of Geometry
Major Fields of Scientific Research: finite
geometries, coding theory
assia@fmi.uni-sofia.bg



Ivan Landjev, DSc
New Bulgarian University, Department of
Computer Science
Major Fields of Scientific Research: finite
geometries, coding theory, combinatorics
i.landjev@nbu.bg

CSECS 2018, pp. 011 - 021

Computer Science and Education in Computer Science
14th Annual International Conference
ISSN 2603-4794

June 29 – 30, 2018, Boston, USA

TEACHING ELLIPSE WITH CAS

Marin Marinov, Petya Asenova

***Abstract:** The paper presents an approach to teaching Mathematics at university level using the computer system for symbolic manipulations Wolfram Mathematica. It gives an environment for fast calculations, visualization of properties, and explorations leading to the hypotheses. These way students participate actively in constructing their knowledge, obtain deeper understanding and they are able to apply their knowledge and skills for solving problems. We illustrate the approach on the topic Ellipse and demonstrate how it facilitates the understanding of the main concepts.*

***Keywords:** teaching Mathematics at university level; Analytical Geometry; Ellipse; Computer Algebra Systems (CAS); conceptual understanding*

ACM Classification Keywords: *Computing Classification system, 2012 Revision, Computing education, Software creation and management*
(<http://www.acm.org/about/class/class/2012?pageIndex=0>)

Introduction

Computer technologies assist all forms of education, in all subjects and educational stages. They radically change the teaching process. This paper presents an approach to teaching Mathematics at university level using the computer system for symbolic manipulations *Wolfram Mathematica*.

There is not enough research on the feasibility of computers for teaching Mathematics in Bulgaria. More data is available for elementary and secondary schools. The latest survey, held under the project financed by the Bulgarian National Scientific Fund (No DN-05/10, 2016), studies the use of computer games in primary and secondary education, including in Mathematics. [Tuparova, D., Tuparov G., Veleva V., Nikolova E., 2018]. There is insufficient scientific data about the state of the area in respect to the higher education.

This paper does not aim to analyze the use of computer technologies in Mathematics education in general. It aims to demonstrate some teaching strategies using computer algebra system (CAS) for teaching Analytical Geometry - the topic Ellipse in the plane.

We use the following instruments of the CAS (with Wolfram Mathematica) [M. Marinov, 2014], [P.Asenova, M. Marinov, 2018]:

- Fast calculations;
- Graphical representations of the properties of the studied objects;
- Animations, modelling and simulations ;

- Computer games.

The students use the same set of features of the CAS under the supervision of the teacher during the class sessions and as part of their study at home.

The main strategy of the teacher is to utilize the listed CAS instruments to visualize the properties of the ellipse and this way to help students reach better understanding of the basic concepts. [A.V. Usova, 2011], [A.V. Usova, 1986], [A.V. Usova, 1989] Students explore the properties of the object; they make hypotheses, transform expressions, and suggest proves. They are more active and engaged in the learning process and construct deeper mathematical knowledge.

The teacher role is to select the appropriate tasks, to supervise the students and to stimulate them to explore different cases. Some tasks sound like games, so that to provide a positive emotional background and motivation to learn.

Some examples from our teaching experience are discussed in the following section.

Discover the ellipse

The initial step in the process is to prepare the students for the introduction of the concept they are going to study. Using animation, we focus on the demonstration of specific property that determines the ellipse as a curve in the Euclidean plane – this curve is a set of points, such that for any point M of the set, the sum of the distances from M to two fixed points F_1 and F_2 is a constant. To develop further the concept, we use the steps formulated by A.V. Usova. [A.V. Usova, 2011], [A.V. Usova, 1986], [A.V. Usova, 1989].

Example 1. The starting point is the notion *circle* mentioned above – every point M of the circle in the plane is equidistant from a given point O_1 (center). Let the radius of the circle is a than $|O_1M| + |MO_1| = 2a$. Using the Mathematica function `ContourPlot` students draw a circle - for example a circle k with the center $O_1(1, \frac{1}{2})$ and diameter 5.6:

```

In[2]= ContourPlot[ $\sqrt{\left(x - \frac{1}{2}\right)^2 + (y - 1)^2} + \sqrt{\left(x - \frac{1}{2}\right)^2 + (y - 1)^2} == 5.6, \{x, -3, 4\},$ 
   $\{y, -3.5, 4.5\}, \text{Axes} \rightarrow \text{True}, \text{Frame} \rightarrow \text{False}, \text{PlotRange} \rightarrow \text{Automatic}]$ 

```

What happens when the point O_1 becomes double? Students explore the answer using animation. They choose any diameter and two points F_1 and F_2 belonging to the diameter and symmetric to O_1 , and they draw a curve with the same property $|F_1M| + |MF_2| = 5.6$. They move F_1 and F_2 on the diameter (on the line t) starting from O_1 and notice the circle change its form becoming flattened circle related to the diameter (*Figure 1.*)

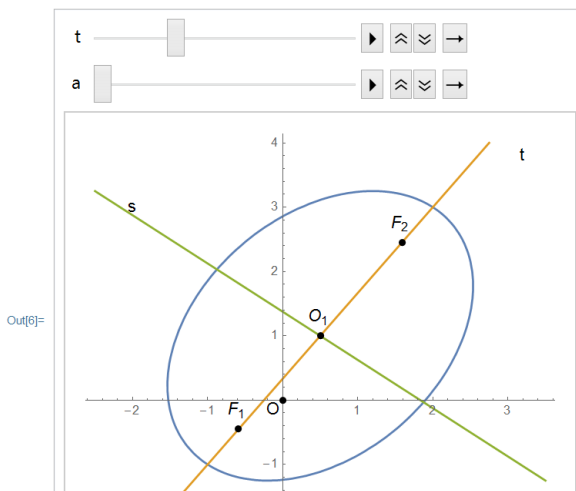


Figure 1. Discovering ellipse

The line s (Figure 1) is perpendicular to t and $O_1 \in s$. When $|O_1F_1| = |O_1F_2|$ then the curve is symmetric related t and s .

The animation above shows as well as the following statements:

- If $|F_1F_2| = 0$, i.e. $F_1 = F_2 = O_1$, then the curve is the circle k .
- If $|F_1F_2| = 5.6$, then the curve is transformed to the segment F_1F_2 .
- For every point M from the disk with boundary circle k there is a number z , such that: if $|O_1F_1| = |O_1F_2| = z$, then point M belongs to the curve.

The replacement of the diameter length 5.6 by $2a$ in the program code allow the students to explore the change of the curve.

The exploration described above leads to the definition of ellipse:

Definition. An ellipse is a curve in a plane such that the sum of the distance from any of its points to the two points F_1 и F_2 , called foci, is a constant.

Example 2. Canonic equation of ellipse

The possibilities for symbolic manipulations of Mathematica make easy to obtain the ellipse canonic equation.

Let F_1 and F_2 are any given points and the number a is such that $|F_1F_2| < 2a$. We choose the Cartesian coordinate system O_{xy} such that the point O is the middle of the segment F_1F_2 and the points F_1 and F_2 belong to the axis O_x . We can assume the coordinates of the two foci are: $F_1(-c; 0)$ and $F_2(c; 0)$, where $0 < c < a$. Then:

$$e = \{(x; y) : \sqrt{(x+c)^2 + y^2} + \sqrt{(x-c)^2 + y^2} = 2a\}$$

$$(i) \text{ If } M(x; y) \in e, \text{ then } \sqrt{(x+c)^2 + y^2} = 2a - \sqrt{(x-c)^2 + y^2}.$$

So:

$$\text{In[8]} = \left(\sqrt{(x+c)^2 + y^2} \right)^2 = \left(2 \cdot a - \sqrt{(x-c)^2 + y^2} \right)^2 // \text{Simplify}$$

$$\text{Out[8]} = c x + a \sqrt{c^2 - 2 c x + x^2 + y^2} = a^2$$

$$\text{In[9]} = \left(a \sqrt{c^2 - 2 c x + x^2 + y^2} \right)^2 = (a^2 - c x)^2 // \text{Simplify}$$

$$\text{Out[9]} = a^4 + c^2 x^2 = a^2 (c^2 + x^2 + y^2)$$

We receive: $(a^2 - c^2)x^2 + a^2 y^2 = a^4 - a^2 c^2$

Let $b^2 = a^2 - c^2$ and if $M(x; y) \in e$, then:

$$(1) \quad \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

(ii) If the coordinates of the point $M(x; y)$ satisfied the equation (1), then $y^2 = \frac{(a^2 - c^2)(a^2 - x^2)}{a^2}$. Then the sum of the lengths of the sections $|F_1M| + |F_2M|$ is equal to

$$\text{In[1]} = \sqrt{(x+c)^2 + y^2} + \sqrt{(x-c)^2 + y^2} /. \left\{ y^2 \rightarrow \frac{(a^2 - c^2)(a^2 - x^2)}{a^2} \right\} // \text{FullSimplify}$$

$$\text{Out[1]} = \sqrt{\frac{(a^2 - c x)^2}{a^2}} + \sqrt{\frac{(a^2 + c x)^2}{a^2}}$$

Moreover, $|x| \leq a$ follows from (1), then $a \pm \frac{c}{a} x \geq 0$. So:

$$|F_1M| + |F_2M| = \left(a - \frac{c}{x} \right) + \left(a + \frac{c}{x} \right) = 2a.$$

Example 3. Optic properties of ellipse

We suppose that the students are familiar with tangent to ellipse, as well as with lines in the plane. So they can use their knowledge to learn the optic properties of ellipse. There are three steps in the example 3:

(i) They are able to define the function $oM[n, A, M, z]$, depending on the normal vector n , the points A and M , and the number z . This function determines the coordinates of the point M_z with the following properties (Figure 2):

- Ray AM_z is a reflection of the ray MA by the line at the point A and perpendicular to the vector n .
- $|AM_z| = z$.

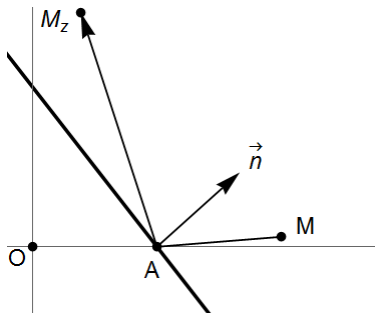


Figure 2. The function $oM[n, A, M, z]$ determines the point M_z

The code of the function is given below:

```
ln[22]= oM[n_, A_, M_, z_] := Module[{n1 = n, a = A, m = M, z1 = z, t, t0, r},
  t0 = t /. Solve[(a - m + n * t) . n == 0, t][[1]];
  r =  $\frac{a - m + (2 * t0) * n}{\text{Norm}[a - m + (2 * t0) * n]}$ ;
  r = a + z1 * r];
```

The students validate the function $oM[n, A, M, z]$ on simple examples. They apply later this function in example (iii) for drawing a reflection of a ray.

- (ii) Using their previous experience students create an animation of ellipse

$$e: \frac{x^2}{a^2} + \frac{y^2}{a^2 - c^2} = 1, \quad 0 < c < a$$

and point $A(a \cos(t); \sqrt{a^2 - c^2} \sin(t))$, $\forall t \in [0; 2\pi)$ of the ellipse.

When t is moving then A traces the ellipse e .

In[6]= `Animate[p[a, 2, t], {t, 0, 2 * π}, {a, 2, 10}, AnimationRunning -> False]`

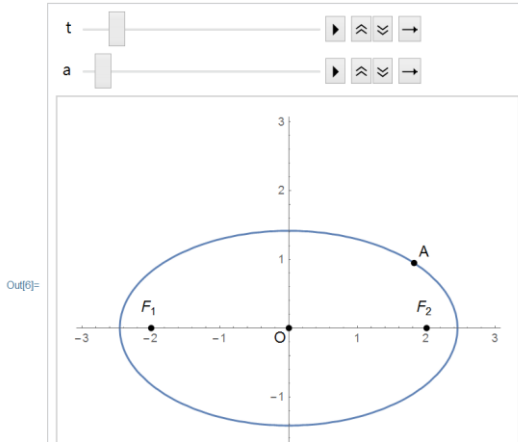


Figure 3. The point A traces the ellipse

- (iii) The students explore how the ray starting at the point F_2 reflects on the ellipse at the point A . They change the animation from (ii) adding the ray starting at the point F_2 . They use the validated function $oM[n, A, F_2, z]$ from (i) to draw the reflected ray.

```
Animate[Show[{p[a, 2, t], ol[a, 2, t, z]], {t, 0, 2 *  $\pi$ }, {a, 2, 10}, {z, 0, 10},
  AnimationRunning -> False]
```

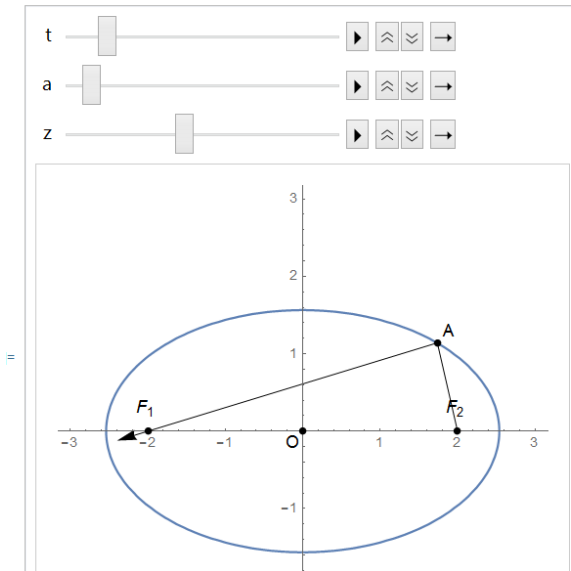


Figure 4. Optic properties of ellipse

This animation allows the students to observe what happens when the point A traces the ellipse. This way students discover the optic property of this curve - the rays from the first focus are reflected by the ellipse to the second focus. The next step is to prove this hypothesis.

Conclusion

A short enquiry about computer usage in Math is made by us with first year Computer Science students at New Bulgarian University. The results show they estimate very high the benefits of computer technologies as a valuable tool to support and promote obtaining

Math knowledge and skills. They notice computers contribute to deeper understanding of mathematical concepts and practically eliminate technical problems to calculate more complex expressions. All students strongly prefer learning Math courses with computers. This rise their interest and motivation to learn Math.

As our experience is based on CAS (Wolfram Mathematica) we underline the benefits of the CAS to develop mathematical concepts using visualization - calculations. The role of teachers is to select appropriate tasks and to stimulate exploration. CAS affords many opportunities for rich mathematical learning.

Acknowledgment

The study has been funded by the Bulgarian national scientific fund (N DN-05.10, 2016).

Bibliography

- A. V. Usova. Nekotore metodicheskie aspekti problem formirovaniya poniatii u uchastchih-sia i studentov vuzov. Mir nauki, kulturi, obrazovaniya, No 4(29), 2011 (in Russian).
- A. V. Usova. Psihologo-didakticheskie osnovi formirovaniya u uchastchih-sia nauchnih ponyatij: uchebnoe posobie. – Chelyabinsk: Chelyabinskiyj pedinstitut, 1986, Ch. 1 (in Russian).
- A. V. Usova. Psihologo-didakticheskie osnovi formirovaniya u uchastchih-sia nauchnih ponyatij: uchebnoe posobie. – Chelyabinsk: Chelyabinskiyj pedinstitut, 1989, Ch. 2 (in Russian).

D. Tuparova, G. Tuparov, V. Veleva, E. Nikolova., Educational Computer Games and Gamification in Informatics and Information Technology Education – Teachers’Point of View, in Proc. Of 41 Conference MiPr0, 2018, Opatija, 21.05-25.05.2018

Ellipse. mathworld.wolfram.com/Ellipse.html (12/05/2018)

M. Marinov, P. Asenova. Mathematical Proofs at University Level.// Computer Science and Education i Computer science, Fulda, Germany, 2013, 72-81, ISSN1313-624 (with P. Asenova)

M.Marinov. Obuchenie po matematika sas sistema za simvolno smiatane. // Matematika I matematichesko obrazovanie, UBM, V 44, 2014, pp.137 – 148.

P. Asenova, M. Marinov. Teaching Mathematics with Computer Systems.//Mathematics and Education in Mathematics. UBM. V. 47, 2018, pp.213-221. ISSN 1313-3330

Authors' Information



Prof. Marin Marinov, Ph.D., New Bulgarian University, Informatics Department, 21 Montevideo Str., 1618 Sofia, Bulgaria, mlmarinov@nbu.bg

Major Fields of Scientific Research: Differential Equations, Mathematical education, IT in education



Assoc. Prof. Petya Asenova, Ph.D., New Bulgarian University, Informatics Department, 21 Montevideo Str., 1618 Sofia, Bulgaria, pasenova@nbu.bg

Major Fields of Scientific Research: Multimedia technologies, ICT applications in education, Project based education

FOREIGN DIRECT INVESTMENT NET INFLOWS: DATA-DRIVEN ANALYSIS

**Ana GJORGJEVIKJ¹, Kostadin MISHEV¹, Irena
VODENSKA², Lubomir CHITKUSHEV², Dimitar
TRAJANOV¹**

¹⁾ ss “Cyril and Methodius” University, Faculty of Computer
Science and Engineering, Skopje, Macedonia


²⁾ Boston University, Metropolitan College, Boston, USA

***Abstract:** Foreign direct investments (FDIs) are considered an important driver of host countries economic growth, as they provide stable inflow of foreign capital, know-how and technology. FDI determinants have been studied extensively over the past decades and often identified among the different economic, regulatory and stability indicators of the host country. Studying FDI flows has not-trivial challenges too, because of the increased complexity of the multinational enterprises organizational structures and the difficulty this causes when compiling national FDI statistics. The purpose of this work is to study a large set of countries economic, governance and geopolitical indicators with regard to their correlation to the FDI net inflows, while taking into consideration the challenges*

suggested in the literature. Besides the detailed analysis of the correlations by indicator topic and countries income categories, the main contribution of this work is compilation of a set of geopolitical indicators from the world's largest events database, GDELT, and correlating them to the FDI net inflows statistics.

Keywords: *Foreign direct investment, FDI net inflows, GDELT, Data Science.*

ACM Classification Keywords: *Information systems - Information systems applications - Data mining*



Introduction

Over the last decades, foreign direct investments (FDIs) have become a major form of international capital transfer, as well as an important driver in creating stable and long-lasting relationships between the countries, mainly because of the liberalization of the exchange controls and market access, as well as of the technological advancement [OECD, 2008]. Foreign direct investments are defined as cross-border investments taken to gain control or significant degree of influence over the management of an enterprise resident in an economy that is different from the economy of the investor [IMF, 2009]. FDIs usually imply a lasting relationship between the investor and the affiliate, which may as well include supply of knowledge and technology to the affiliate, besides supply of capital [IMF, 2009]. In a proper policy framework, foreign direct investments usually promote host country's development and economic performance [OECD, 2008].

Identifying the factors that are closely related or lead to profitable foreign investments in certain country is important to both the countries interested in attracting new foreign investors and to the investors looking for the least risky country for their future investments. If these factors are unambiguously identified, countries would be able to become more attractive to investors by improving them, while the investors would have more precise methods for comparing different countries. Research on this topic has been done over the last decades and suggests that the problem is not easily solvable. The investors differ significantly by their main investing interest, countries differ by their stage of development and other aspects, so it is very difficult to reach a general consensus and model that covers all cases. However, the

shift of the scientific paradigms towards data-driven science and the advancement of the data processing methodologies give hope to gain some deeper understanding of the problem through analysis of all the data collected over the past decades.

FDI net inflows of one economy represent the net value of the inward direct investments taken by non-resident investors to that economy. Much of the available research related to FDI prediction, uses country's FDI net inflows as target indicator, a fact that motivated the comprehensive analysis of this indicator presented in this work. The main contributions consist of the analyses of countries annual FDI net inflows correlation to large number of economic, governance and geopolitical indicators, as well as identification of the trends these correlations follow for countries at different stage of development. The analyses can be considered as a first step in any subsequent development of a model for predicting future FDI inflows in one country. The relations between the different indicators do not have to be linear, so the presented analyses focus on monotonic relations. Besides using the publicly available economic and governance indicators, the set of analyzed indicators was extended with geopolitical indicators derived from one of the world's largest event database, GDELT¹. Throughout the paper, the known challenges related to FDI flows analyses are pointed out as well.

Related Work

The motives for investing in a foreign country, as well as the factors that influence investors' choice of a particular country

¹ <https://www.gdeltproject.org/>

have been studied for long time, due to the host countries benefits associated with the foreign investments. According to the Eclectic (OLI) paradigm [Dunning and Lundan, 2008], investors undertake foreign investments for three reasons, i.e. ownership advantages, location advantages and market internalization advantages. The same authors classify multinational enterprises (MNEs) interested in foreign investments into several types, i.e. natural resource seekers (looking for resources of higher quality at lower cost compared to their home country), market seekers (investing in order to supply goods and services to the host or adjacent countries), efficiency seekers (rationalize the structure of already established investments) and strategic assets seekers (promoting long term objectives like global competitiveness). This indicates that the choice of a country can depend from the motive of the investor, as well as other aspects of the investment itself, such as its type, investing sector or MNE size [Botrić and Škuflić, 2006]. Additionally, depending on the motives and type of investment, investors take into consideration different types of countries determinants, e.g. market-seekers commonly consider market size, growth or per capita income, while the efficiency-seekers commonly consider the cost of production [Botrić and Škuflić, 2006].

Some research works show that the FDI net inflows do not provide a realistic view of the direct investments real value for the host country and should not be considered as indicator for the countries' attractiveness to foreign investors. [Leino and Ali-Yrkkö, 2014] find that foreign companies may use other means to found their investments and activities, for example, in case of Finland this may account about half of the financing of foreign-owned companies. Second, just few top transactions can heavily drive annual FDI inflow statistics, giving a biased picture of the

country's attractiveness to foreign investors. Additional challenge is the pass-through capital, where the capital just flows in and out of the country, and the financial round-tripping, where the ultimate investor is a domestic company, in both of which the host country does not have real gain. [Beugelsdijk et al., 2010] argue that FDI stocks are a biased measure of the value-adding activities of the MNE affiliates in the host countries, because they can overestimate or underestimate this value based on the host country characteristics. FDI in one country do not necessary add value to that country and locally raised external funds and labor productivity in the affiliate can contribute to value-adding MNE affiliate activity, not calculated in the FDI statistics. All these findings indicate that FDI inflow and stocks statistics should be used very carefully in any research work.

Datasets

The primary goal of this research is to correlate a large set of available economic, governance and geopolitical indicators of the world's countries to their annual FDI net inflows in order to identify the most important ones for countries at specific development stage. From the publicly available datasets, a collection of around 1600 indicators was created from three datasets, i.e. the World Development Indicators (WDI)², Worldwide Governance Indicators (WGI)³ and GDELT. More detailed description of each dataset follows.

² <https://data.worldbank.org/products/wdi>

³ <http://info.worldbank.org/governance/wgi/>

World Development Indicators (WDI) is the primary World Bank's collection of more than 1500 indicators on global development, including estimates on national, regional and global level, compiled from officially-recognized international sources like the Organization for Economic Cooperation and Development (OECD), International Monetary Fund (IMF), United Nations (UN) and others. The collection is published annually, but updated quarterly online. The indicators cover the time period from 1960 to 2017 and currently describe 217 economies. For easier understanding, the indicators are grouped into several topics like Economic Policy & Debt, Education, Environment, Financial Sector, Gender, Health, Infrastructure, Poverty, Private Sector & Trade, Public Sector, Social Protection & Labor. In this paper, the grouping was utilized for more clear reporting of the correlations between the set of indicators belonging to each group and the FDI net inflows. The whole collection of indicators was used in this research and retrieved in April 2018 from the official World Bank web site⁴.

Worldwide Governance Indicators (WGI) are a result of a long-standing research project aiming to develop cross-country indicators of governance, defined as traditions and institutions by which an authority in a country is exercised [Kaufmann et al., 2011]. The indicators are calculated based on several hundred variables obtained from more than 30 different sources like surveys, non-governmental organizations, private sector companies, public sector organizations, in order to capture their perception of the different aspects of the country's governance.

⁴ <https://datacatalog.worldbank.org/dataset/world-development-indicators>

Over 200 countries and territories are monitored since 1996 over six aspects of governance belonging to three areas, i.e. (a) the process by which the country's government is selected / monitored / replaced, (b) government's capacity to formulate and implement sound policies effectively and (c) the respect for the institutions that govern economic and social interactions by the citizens and the state. Two indicators are calculated for each of the three areas and they are: Voice and Accountability (perceptions of the extent to which the citizens can participate in the selection of their government, freedom of expression, free media), Political Stability and Absence of Violence (likelihood that the government will be destabilized by unconstitutional / violent means), Government Effectiveness (quality of the public services, civil services and independence from political pressure, the quality which the policies are formulated and implemented with), Regulatory Quality (government's ability to formulate and implement sound policies that promote the private sector development), Rule of Law (the extent to which the agents have confidence in the rules of the society and abide them, like the quality of the enforcement of contracts, property rights, police, courts and the likelihood of crime / violence), Control of Corruption (extent to which the public power is used for private gain) [Kaufmann and Kraay, 2008] [Kaufmann et al., 2011]. For our research all six governance indicators were used, as retrieved in April 2018 from the official World Bank web site⁵. Data for 215 economies was included.

⁵ <https://datacatalog.worldbank.org/dataset/worldwide-governance-indicators>

The Global Database of Events, Language, and Tone (GDELT) is a platform that monitors world's news media in print, broadcast and web formats in nearly every country, in more than 100 languages, in order to constantly update an extremely comprehensive database of a quarter billion georeferenced event records for the entire world. The data spans from January 1, 1979 up to present, with a main goal to create a catalog of the human societal-scale behavior and beliefs across the world countries, or in other words, a single massive network that contains everything that is happening in the world every day, the context of these events and the involved actors, as well as information on what the world feels about them. Since version 2.0, the Event dataset consists of two tables, i.e. the Mentions table and the Event table. The Mentions table contains record for each mention of an event in a news article with several indicators about the mention (location of the mention within the article, confidence that the event was identified correctly and etc.). The Event table stores details about each recognized event in the news articles. Several types of data are stored for each event and those are the event identification data, the data related to the first and second actor involved in the event, as well as several indices capturing the context of the event. Of interest for our research is the data related to the geography of the event, or more precisely the georeferenced location of the two actors showing the country the actors were in when the event happened. Using the georeferenced locations, we are able to associate the indices of the event and the geographic locations. The Goldstein Scale is a numeric score between -10 and 10 giving the likely impact which a specific type of event has on a country. This index is specific to the event type. The AvgTone index shows the average tone over all news articles that mention the event at least once. It ranges between -100,

extremely negative, and 100, extremely positive, but is normally between -10 and 10 [Leetaru and Schrodt, 2013].

Methodology

The presented analyses include all countries for which statistical data was available in the selected datasets. No missing values were estimated and imputed, but the analyses were done on the available data solely.

GDELT indicators calculation. For the research work presented in this paper, the data from the Event table in GDELT Events database, available on Google BigQuery, was used. All event data starting from year 1979 was used for calculation of the statistical indicators. Since all other indicators were available at annual frequency, while the GDELT events are available at 15 minutes' frequency, the event data had to be summarized at annual level in order to be comparable with the rest of the indicators. A distinction was made between the events in which the geographic location of the first actor (actor 1) differs from the geographic location of the second actor (actor 2) and the events in which both actors are at the same geographic location (referred to as 'self' in the rest of the paper). For the first group of events, the mean and standard deviation were calculated on the two indices, AvgTone and Goldstein scale index, for each country appearing as geographic location of the first actor in each year. The same statistics was done for the countries appearing as geographic location of the second actor. For the second group of events, since the same country appears as geographic location of both the first and the second actor, the above described statistics was calculated only once and included in the feature set for that country and year. The countries that have not hosted any of the actors of any

of the events that happened in a particular year Y , value of zero was imputed for all GDELT indicators.

Besides the mean and standard deviation of the Goldstein index, an additional modified Goldstein index was calculated as suggested in [Trajanov et al., 2017]. Since the Goldstein index measures the impact one event has over the stability of the countries involved in it and is related to the event type, the authors suggest that not all events are equally important and equally influence the involved countries stability. The modified Goldstein index is calculated as given with

$$mGSi = \frac{\sum_e GSi_e \times NumArticles_e}{\sum_e NumArticles_e} . \quad \text{Equation 1, where the Goldstein}$$

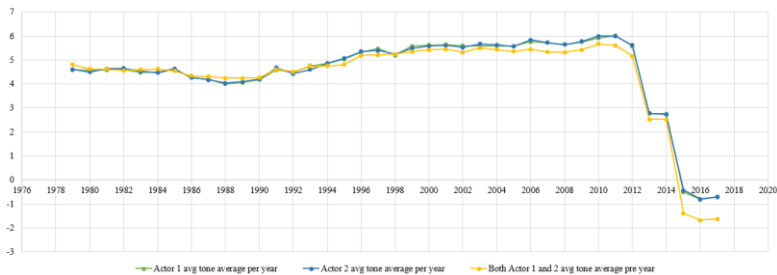
index for an event is multiplied by the number of articles that mention it and the sum of the products is then divided by the total number of articles that mention the events. Table 1 summarizes all GDELT derived indicators.

$$mGSi = \frac{\sum_e GSi_e \times NumArticles_e}{\sum_e NumArticles_e} . \quad \text{Equation 1}$$

Additional analyses were done to three indicators, i.e. AvgTone average per country and year for countries hosting actor 1, actor 2 or both actors. Figure 1 illustrates the change of the mean value of these indicators per year, calculated over all countries. The time series indicate a significant shift in the mean values around year 2015. For that reason, a shift of the values from the first two intervals was done, so that the mean value of all intervals is equal. With this heuristics, we made an effort to ensure having data that is comparable over the whole time period we are analyzing. The mean values per year after the change are shown in Figure 2.

Table 1. Indicators calculated from GDELT with annual frequency

| GDELT indicator | Description |
|---|---|
| avg (AvgTone _x), std (AvgTone _x), $x \in \{\text{Actor1}, \text{Actor2}, \text{Self}\}$ | Actor1/Actor2 avg tone mean and standard deviation for year Y, for events where $\text{geo}(\text{Actor1}) \neq \text{geo}(\text{Actor2})$. Self refers to events where $\text{geo}(\text{Actor1}) = \text{geo}(\text{Actor2})$. |
| avg (GSI _x), std (GSI _x), $x \in \{\text{Actor1}, \text{Actor2}, \text{Self}\}$ | Actor1/Actor2 Goldstein index mean and standard deviation for year Y, for events where $\text{geo}(\text{Actor1}) \neq \text{geo}(\text{Actor2})$. Self refers to events where $\text{geo}(\text{Actor1}) = \text{geo}(\text{Actor2})$. |
| avg (mGSI _x), $x \in \{\text{Actor1}, \text{Actor2}, \text{Self}\}$ | Actor1/Actor2 modified Goldstein index mean for year Y, for events where $\text{geo}(\text{Actor1}) \neq \text{geo}(\text{Actor2})$. Self refers to events where $\text{geo}(\text{Actor1}) = \text{geo}(\text{Actor2})$. |
| avg (Articles _x), $x \in \{\text{Actor1}, \text{Actor2}, \text{Self}\}$ | Actor1/Actor2 average number of articles that mentioned it in year Y, for events where $\text{geo}(\text{Actor1}) \neq \text{geo}(\text{Actor2})$. Self refers to events where $\text{geo}(\text{Actor1}) = \text{geo}(\text{Actor2})$. |

**Figure 1.** Change of the mean value of the three indicators per year, original values

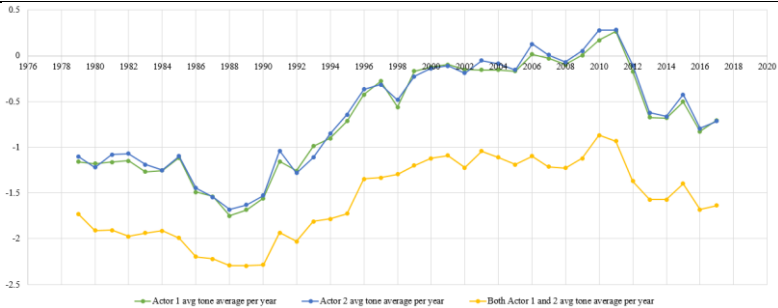


Figure 2. Change of the mean value of the three indicators per year, values after equalization of the intervals mean

FDI statistics and its challenges. When considering the use of FDI statistical data for research, the methodologies that have been used for its compilation in the past and their recent changes have to be taken into consideration. This subsection underlines some of the challenges related to the FDI statistics compilation and the official recommendations for overcoming them. There are few phenomena related to the foreign direct investments that have become more obvious with the rise of the MNEs financial structures complexity, leading to revision of the international standards for FDI statistics compilation. One of the phenomena is the use of special purpose entities (SPEs) by MNEs to manage their global investments and operations. The SPEs are defined as enterprises formally registered with a national authority, but ultimately controlled by non-resident enterprise, having little or no physical presence in the host country and almost all assets and liabilities in a form of investments from and to other countries. Including SPEs in FDI statistics may lead to double counting of FDI in cases when they are channeled through SPEs from multiple countries before the final host country (pass-through capital). Additionally, country's real inward on outward FDI statistics can be distorted. Another phenomenon is the financial

round-tripping in which the foreign investment has an actual host enterprise as ultimate investor, but the direct investor is from a foreign country [OECD, 2015 Feb] [OECD, 2015 Mar]. In order to adapt to the new economic circumstances, IMF and OECD revised their standards in new editions, i.e. the IMF Balance of Payments and International Investment Position Manual, 6th Edition, (BPM6) in 2009 [IMF, 2009] and the OECD Benchmark Definition of Foreign Direct Investment, 4th Edition, (BD4) in 2008 [OECD, 2008], making new recommendations for compilation of the national FDI statistics to make them more meaningful. Among the other, the new requirements include separate identification of the capital that goes through SPEs, presenting the statistics according to the country of the ultimate investor and the immediate investor, distinguishing FDIs by type of transaction and etc. [OECD, 2008]

In this subsection we elaborate the findings from several research papers that discuss the FDI statistics challenges. [Blanchard and Acalin, 2016] analyze the correlation between quarterly FDI inflows and outflows for 25 emerging-market countries, finding that some countries have significantly higher correlation than the average correlation. As suggested by the authors, it can be expected that this type of correlation is higher for FDI statistics calculated at longer intervals, but not at high frequency statistical data like the quarterly statistics. They conclude that even though the suggested corrections of the calculation methodology, e.g. separate treatment of SPEs, reduce the FDI bias, they do not eliminate it completely, i.e. “the measured FDI is not entirely true FDI”. [Dellis et al., 2017] further test these findings on the OECD database by comparing the regression results before and after removing the countries with high inward and outward FDI correlation from the dataset. Their conclusion is that although

changes in the magnitude of the coefficients can be noticed, the overall impact can be considered negligible.

In this research the annual FDI statistics from the World Bank's World Development Indicators dataset were used, which are compiled from the IMF Balance of Payments database, data from the United Nations Conference on Trade and Development and official national sources. In an attempt to avoid any bias and ensure relevance of our results, the dataset used in our research was tested for bias introduced by countries having very high correlation between the FDI net inflows and outflows. Since the statistical data used in this research is calculated on annual level, we eliminate only the countries that deviate significantly from the mean correlation calculated on the dataset. Pearson product-moment correlation was calculated on the FDI data time series for each country separately. We took into consideration only the countries whose correlation had p-value below 1% and rejected the rest of the correlations. On the dataset with 96 retained countries, only five had deviation from the mean correlation above the standard deviation and those are the countries that were filtered during the testing. Figure 3 illustrates the distribution of the correlation values over the dataset consisting of 96 countries with p-value below 1%.

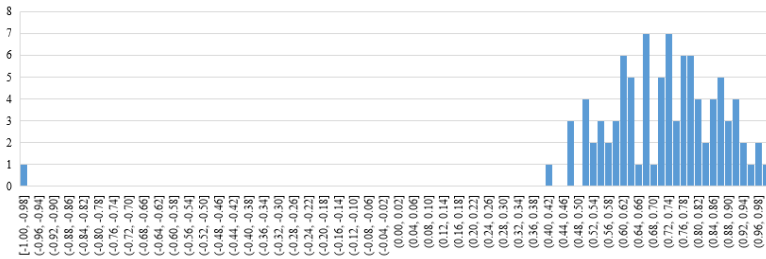


Figure 3. Distribution of the Pearson product-moment correlation between countries annual FDI net inflows and outflows on a selected dataset with 96 countries

In order to see if the five countries with high correlation between the annual FDI net inflows and outflows introduce bias in the correlation coefficients between the set of indicators and the FDI net inflows, a comparison between the correlation coefficients on the dataset before and after the filtering for the affected income categories was done. All five filtered countries belong to the high income category. For this analysis only the correlation coefficients whose p-value is below 1%, or 941 indicators, were included.

Figure 4 illustrates the relation between the correlation coefficients calculated on the original dataset and the filtered dataset, which is almost linear. The root-mean-square error (RMSE) of the correlation coefficients calculated on the filtered dataset and on the original dataset is equal to 0.028, whereas the mean absolute error (MAE) is equal to 0.018. As a conclusion, since the presented results indicate a small bias, the subsequent analyses are done on the original dataset, without filtering.

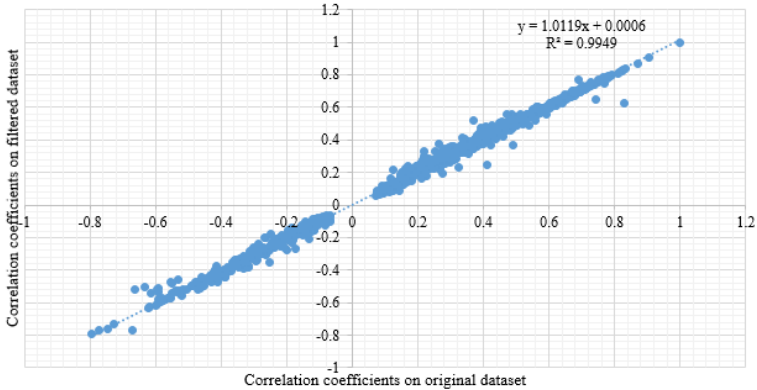


Figure 4. Relation between the correlation coefficients calculated on the original dataset and the filtered dataset. Only the indicators for which the correlation coefficient had p-value below 1% were included.

Country grouping. Since one of the goals of this research is to evaluate whether the same set of indicators are the most correlated with the annual FDI net inflows value for all countries, the countries had to be grouped by some relatedness criteria. At this stage of the research, it was considered that the level of countries' development is the most appropriate criteria, so the World Bank's classification of economies by income was selected. World Bank provides several classifications of the economies it monitors for easier aggregation, comparison and presentation of the statistical data⁶. It classifies the economies by geographic region, income group, operational lending categories and other criteria. The classification by income is done by use of

⁶ <https://datahelpdesk.worldbank.org/knowledgebase/articles/378834-how-does-the-world-bank-classify-countries>

the gross national income (GNI) per capita in U.S. dollars, converted from country's local currency with the Atlas method⁷. In the current classification, as retrieved in April 2018, the countries are divided into four income classes, low (GNI per capita of \$1,005 or less in 2016), lower-middle (GNI per capita between \$1,006 and \$3,955), upper-middle (GNI per capita between \$3,956 and \$12,235), and high (GNI per capita of \$12,236 or more). The assignment of countries to classes is done each July 1 and stays fixed for the entire fiscal year. 78 economies are classified as high income economies, 56 as upper-middle income, 53 as lower-middle income and 31 as low income economies⁸.

Indicators correlation calculation. In contrast to the Pearson product-moment correlation coefficient that measures the strength of a linear relation between two variables, the Spearman rank correlation coefficient measures the strength of the monotonic relation between the variables. Two variables that do not have a linear relation, might still have a monotonic relation, i.e. both may be increasing or decreasing following a different pattern than a linear one. Spearman rank correlation coefficient ranges between -1 and 1, receiving negative values when one of the variables is decreasing and the second one increasing, positive values when the two variables are increasing or decreasing and zero value when there is no correlation. Since determining whether there is any kind of relation between two indicators, not

⁷ <https://datahelpdesk.worldbank.org/knowledgebase/articles/378832-what-is-the-world-bank-atlas-method>

⁸ <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-andlending-groups>

just linear, is of interest in this research, the Spearman rank correlation coefficient was chosen. Once the indicators that have monotonic relation with the FDI net inflows are identified, techniques that are able to capture non-linear relations between the independent variables and the target variable can be used to come to a final predictive model. Because the number of missing values in the used datasets cannot be neglected, for each calculated correlation score additional information, i.e. the significance level (p-value) and the samples size on which the correlation was calculated, are presented. The calculations were done with Python's SciPy library⁹.

Results and Discussion

Correlation coefficients differences between income categories. The results indicate that differences between indicators correlation to the FDI net inflows exist between the different income categories. However, it has to be emphasized that the number of available data in the different income categories differs as well, so the correlation between a same pair of indicators in the different categories may have been calculated on significantly different sample sizes, i.e. may have different significance level (p-value). The purpose of this subsection is to analyze the differences between the correlation coefficients of the same indicators in the different income categories.

To have a set of reliable correlation coefficients on which the income categories will be compared, only the indicators whose correlation to FDI net inflows has p-value below 5% in all four

⁹ <https://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.spearmanr.htm>

income categories were retained. This resulted in a dataset of 643 indicators.

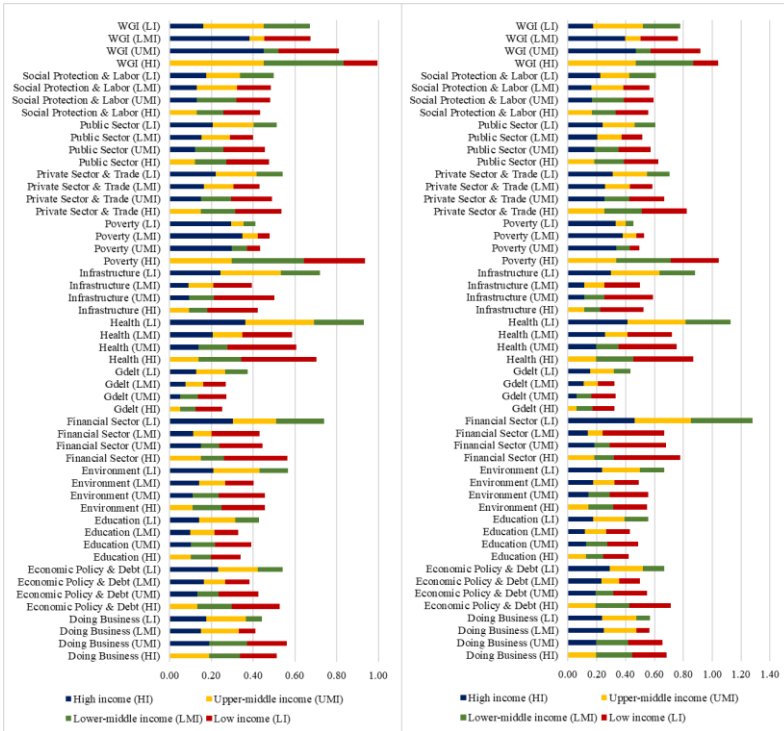


Figure 5. (a) Mean absolute error (MAE) and (b) root-mean-squared error (RMSE) between the correlations of selected indicators and FDI net inflows for the four income categories, summarized by indicator topic gives the MAE and RMSE between all income categories, indicating that the deviations are lowest between the high and upper-middle income categories, as well as the upper-middle and lower-middle income categories. Deviations are highest between the low income category and the high / upper-middle income categories. Figure 5 further divides the analysis by types of indicators in a stacked bar chart. Figure 5 (a) presents the MAEs between the income categories by

indicator type, whereas Figure 5 (b) presents the RMSEs. The errors, i.e. differences between the correlation coefficients for the different income categories on GDELTA indicators are among the smallest, whereas on the Worldwide Governance Indicators (WGI) and WDI health indicators are among the largest.

Table 2. Mean absolute error (MAE) and root-mean-squared error (RMSE) between the correlations of selected indicators and FDI net inflows for the four income

| | | High | Upper-middle | Lower-middle | Low |
|---------------------|-------------|------|--------------|--------------|------|
| High | MAE | 0.00 | 0.13 | 0.15 | 0.24 |
| | RMSE | 0.00 | 0.19 | 0.21 | 0.30 |
| Upper-middle | MAE | 0.13 | 0.00 | 0.12 | 0.22 |
| | RMSE | 0.19 | 0.00 | 0.14 | 0.28 |
| Lower-middle | MAE | 0.15 | 0.12 | 0.00 | 0.14 |
| | RMSE | 0.21 | 0.14 | 0.00 | 0.20 |
| Low | MAE | 0.24 | 0.22 | 0.14 | 0.00 |
| | RMSE | 0.30 | 0.28 | 0.20 | 0.00 |

categories

Statistics by indicator type per income category. This subsection presents summary statistics of the different indicator types for each income category separately, i.e. the mean correlation value per indicator type and its standard deviation. Each income category was analyzed separately by retaining only the correlation coefficients with p-value below 5%. The statistics is summarized in Table 3.

As the results indicate, the mean correlation coefficients are the highest for the World Development Indicators in topics

Infrastructure and Economic Policy & Debt, with little variability of this mean among the different income categories. GDELT indicators show a little lower mean correlation than the previous two indicator topics, but again the variability between the income groups of the mean correlation is low. As expected, the indicators in the topic Poverty from WDI have negative mean correlation to the FDI net inflows in all income categories.

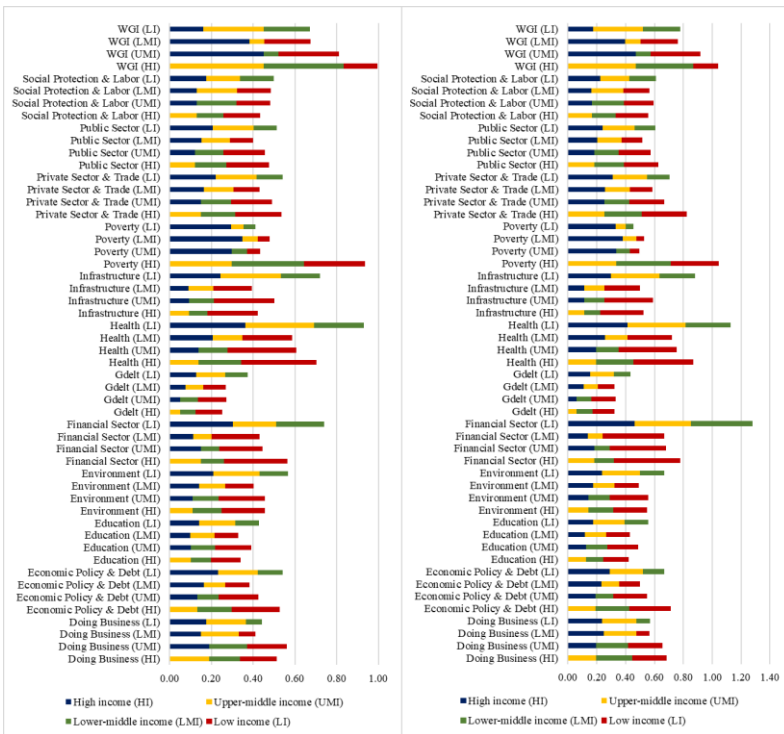


Figure 5. (a) Mean absolute error (MAE) and (b) root-mean-squared error (RMSE) between the correlations of selected indicators and FDI net inflows for the four income categories, summarized by indicator topic

Table 3. Mean correlation by indicator topic per income group. The standard deviation and average sample size on which the correlations were calculated are given in brackets.

| Indicator Topic / Income Group | High | Upper-middle | Lower-middle | Low |
|------------------------------------|-----------------------|-----------------------|-----------------------|----------------------|
| Doing Business | 0.02 (s=0.29, n=411) | 0.14 (s=0.33, n=438) | 0.14 (s=0.2, n=459) | 0.08 (s=0.27, n=255) |
| Economic Policy & Debt | 0.37 (s=0.35, n=1448) | 0.34 (s=0.35, n=1511) | 0.3 (s=0.27, n=1604) | 0.25 (s=0.25, n=969) |
| Education | 0.19 (s=0.27, n=971) | 0.22 (s=0.3, n=702) | 0.15 (s=0.26, n=811) | 0.25 (s=0.28, n=446) |
| Environment | 0.23 (s=0.27, n=1529) | 0.27 (s=0.29, n=1249) | 0.22 (s=0.22, n=1284) | 0.14 (s=0.21, n=821) |
| Financial Sector | 0.17 (s=0.34, n=1228) | 0.24 (s=0.27, n=1244) | 0.24 (s=0.23, n=1000) | 0.05 (s=0.34, n=691) |
| GDELT | 0.27 (s=0.32, n=1893) | 0.26 (s=0.32, n=1645) | 0.19 (s=0.31, n=1562) | 0.2 (s=0.25, n=1064) |
| Health | 0.04 (s=0.45, n=1286) | 0.09 (s=0.41, n=1168) | 0.08 (s=0.31, n=1242) | 0.02 (s=0.27, n=802) |
| Infrastructure | 0.48 (s=0.16, n=1165) | 0.54 (s=0.23, n=950) | 0.43 (s=0.18, n=846) | 0.33 (s=0.2, n=467) |
| Poverty | -0.47 (s=0.29, n=152) | -0.27 (s=0.07, n=325) | -0.16 (s=0.17, n=302) | -0.08 (s=0.28, n=86) |
| Private Sector & Trade | 0.07 (s=0.41, n=1267) | 0.19 (s=0.4, n=1003) | 0.2 (s=0.33, n=925) | 0.12 (s=0.32, n=660) |
| Public Sector | 0.21 (s=0.38, n=1307) | 0.23 (s=0.37, n=690) | 0.21 (s=0.25, n=636) | 0.22 (s=0.24, n=298) |
| Social Protect. & Labor | 0.06 (s=0.29, n=1079) | 0.06 (s=0.28, n=663) | 0.08 (s=0.23, n=614) | 0.09 (s=0.22, n=489) |
| WGI | 0.39 (s=0.08, n=1002) | -0.17 (s=0.26, n=952) | -0.05 (s=0.22, n=929) | 0.23 (s=0.04, n=542) |

Visualization of selected results. In this subsection the actual correlation coefficients of a selected set of indicators to the FDI net inflows are presented. A visual presentation that emphasizes the change of the correlation coefficients of one indicator in each

the four different income categories was selected as most suitable for the intended purpose, i.e. illustrating the correlation differences between the income categories. A separate parallel coordinates chart was created for the different indicator topics, due to the large number of indicators considered in the research. The main vertical axis contains the names of the indicators and each of the additional vertical axis refers to the correlation coefficients of one income category on a scale of -1 to 1.

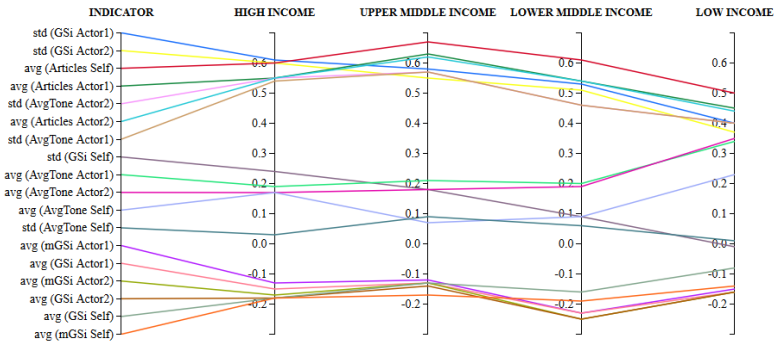


Figure 6. Spearman correlation coefficient between GDELT indicators and FDI net inflows per income group

Since most of the indicator topic contain larger number of indicators than we are able to present in this paper, for each chart a set of indicators was compiled by selecting the several indicators with highest correlation (positive or negative, as applicable) in each income category. When ranking and selecting the indicators in one income category, to ensure relevance of the results, we took into consideration only those having correlation coefficients with significance level below 5%, calculated on sample size larger than 500. Once the indicators for one income category were selected, their correlation coefficients for all other income categories were retrieved, but here no restrictions on the

significance level and sample size were posed, to ensure completeness of the chart data. For that reason, we encourage the readers to look at the appendix section, where all correlation coefficients from the charts, together with their p-values and sample size are available. Figure 6 presents the GDELT indicators and their correlation to FDI net inflows in the four income categories. Figure 7 to Figure 11 present selected WDI indicators from Economic Policy & Debt, Financial Sector, Private Sector & Trade, Infrastructure and Poverty respectively. Figure 12 presents selected set of Doing Business indicators. The charts confirm the previously elaborated findings on the variability (to different extents) of one indicator correlation coefficients between the income categories.

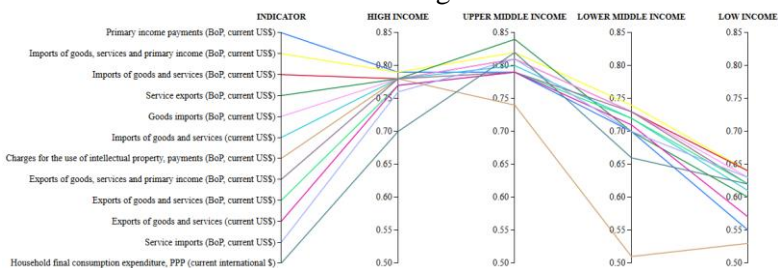


Figure 7. Spearman correlation coefficients between selected WDI (Economic Policy & Debt) and FDI net inflows per income category

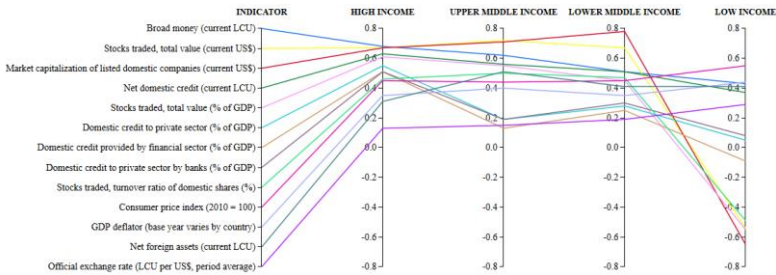


Figure 8. Spearman correlation coefficients between selected WDI (Financial Sector) and FDI net inflows per income category

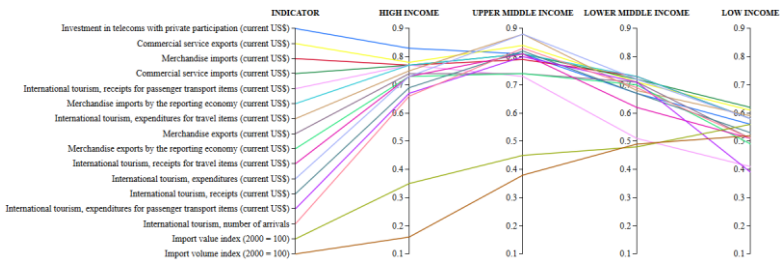


Figure 9. Spearman correlation coefficients between selected WDI (Private Sector & Trade) and FDI net inflows per income category

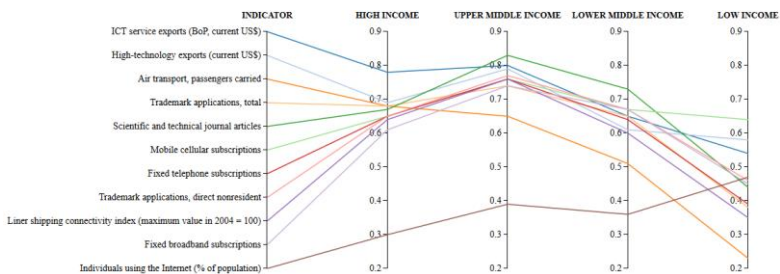


Figure 10. Spearman correlation coefficients between selected WDI (Infrastructure) and FDI net inflows per income category

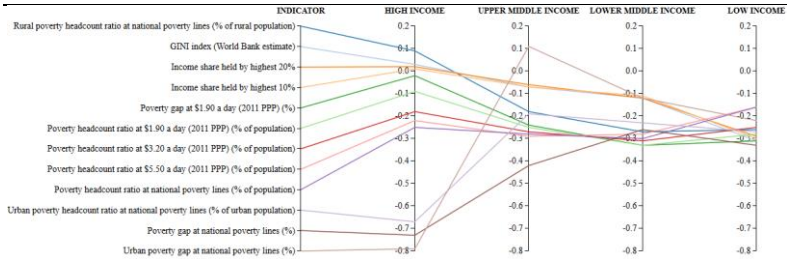


Figure 11. Spearman correlation coefficients between selected WDI (Poverty) and FDI net inflows per income category

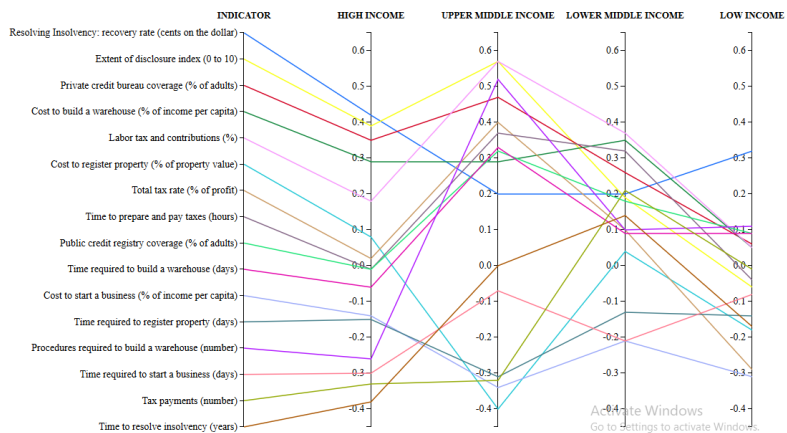


Figure 12. Spearman correlation coefficients between selected Doing Business indicators and FDI net inflows per income category

Conclusion

Identifying the country indicators that are related to profitable foreign direct investments is of great importance to both the investors selecting a country for future investments and for countries trying to increase their FDI inflows. This paper presents detailed analysis of the correlation between large set of economic, governance and geopolitical indicators, compiled from World

Bank's open datasets and GDELT, to countries annual FDI net inflows, indicator often used in research as target indicator when modeling FDI inflows. The results identify the most important groups of indicators for the countries with common income per capita (i.e. GNI per capita) and the differences, as well as trends, of the correlation coefficients that exist between these groups of countries. Through different types of statistical analysis and visualizations, directions on the indicators that are useful for development of FDI net inflows predictive models are given.

Bibliography

OECD, OECD Benchmark Definition of Foreign Direct Investment: 4th Edition (BMD4), Organisation for Economic Cooperation and Development, 2008.

IMF, Balance of Payments and International Investment Position Manual: Sixth Edition (BPM6), International Monetary Fund, 2009.

Dunning JH, Lundan SM. Multinational enterprises and the global economy. Edward Elgar Publishing; 2008.

Botrić V, Škuflić L. Main determinants of foreign direct investment in the southeast European countries. *Transition Studies Review*. 2006 Jul 1;13(2):359-77.

Beugelsdijk S, Hennart JF, Slangen A, Smeets R. Why and how FDI stocks are a biased measure of MNE affiliate activity. *Journal of International Business Studies*. 2010 Dec 1;41(9):1444-59.

Leino T, Ali-Yrkkö J. How well does foreign direct investment measure real investment by foreign-owned companies?: Firm-level analysis, 2014.

Kaufmann D, Kraay A, Mastruzzi M. The worldwide governance indicators: methodology and analytical issues. *Hague Journal on the Rule of Law*. 2011 Sep;3(2):220-46.

Kaufmann D, Kraay A. Governance indicators: Where are we, where should we be going?. *The World Bank Research Observer*. 2008 Jan 31;23(1):1-30.

Leetaru K, Schrodt PA. Gdelt: Global data on events, location, and tone, 1979–2012. *InISA annual convention 2013 Apr 3* (Vol. 2, No. 4, pp. 1-49).

Trajanov D, Vodenska I, Cvetanov G, Chitkushev L. Data Driven Analysis of Trade, FDI and International Relations on Global Scale. *The 13th Annual International Conference on Computer Science and Education in Computer Science, 2017, Jun 29 - Jul 03*

Blanchard O, Acalin J. What Does Measured FDI Actually Measure?. 2016 Oct.

Dellis K, Sondermann D, Vansteenkiste I. Determinants of FDI inflows in advanced economies: Does the quality of economic structures matter?. 2017

OECD, How Multinational Enterprises Channel Investments Through Multiple Countries. *Organisation for Economic Cooperation and Development, 2015 Feb*

OECD, FDI Statistics by the Ultimate Investing Country. *Organisation for Economic Cooperation and Development, 2015 Mar*

Authors' Information



Ana Gjorgjevikj, PhD student, ss “Cyril and Methodius” University, Faculty of Computer Science and Engineering, Skopje, Macedonia, Ruger Boskovik 16 Skopje – Macedonia, ana.gorgevic@gmail.com.

Major Fields of Scientific Research: Data Science, Machine Learning, Deep Learning, Big Data



Kostadin Mishev, PhD student, ss “Cyril and Methodius” University, Faculty of Computer Science and Engineering, Skopje, Macedonia, Ruger Boskovik 16 Skopje – Macedonia, kostadin.mishev@finki.ukim.mk.

Major Fields of Scientific Research: Data Science, Big Data. Machine Learning



Irena Vodenska, PhD, Associate Professor, Boston University, Metropolitan College, Boston, USA, vodenska@bu.edu.

Major Fields of Scientific Research: Quantitative Finance, Complex Networks, Big Data Analytics



Lubomir Chitkushev, PhD, Associate Professor, Boston University, Metropolitan College, Boston, USA, LTC@bu.edu.

Major Fields of Scientific Research: Computer Networks, Health Informatics, Information Security, Complex Systems and Data Analytics



Dimitar Trajanov, PhD, Professor, ss “Cyril and Methodius” University, Faculty of Computer Science and Engineering, Skopje, Macedonia, Ruger Boskovik 16 Skopje – Macedonia, dimitar.trajanov@finki.ukim.mk.

Major Fields of Scientific Research: Data Science, Semantic Web, Big Data, Computer Networks, Parallel processing

Appendix

Detailed results. The subsequent tables present the correlation coefficients, together with the p-values and sample size on which the correlations were calculated for each of the charts in the Results section. One table corresponds to one chart from the Results section and the presentation order is the same.

Table 4. Spearman correlation coefficients between GDELT indicators and FDI net inflows per income category (sample size and p-values above 5% given in brackets)

| Indicator / Income category | High | Upper middle | Lower middle | Low |
|-----------------------------|-------------------|-------------------|-------------------|-------------------|
| std (GSi Actor1) | 0.61 (n=1893) | 0.577 (n=1645) | 0.528 (n=1562) | 0.396 (n=1064) |
| avg (Articles Self) | 0.597 (n=1893) | 0.671 (n=1645) | 0.607 (n=1562) | 0.501 (n=1064) |
| std (GSi Actor2) | 0.597 (n=1893) | 0.545 (n=1645) | 0.513 (n=1562) | 0.372 (n=1064) |
| avg (Articles Actor1) | 0.551 (n=1893) | 0.634 (n=1645) | 0.541 (n=1562) | 0.451 (n=1064) |
| std (AvgTone Actor2) | 0.551 (n=1893) | 0.574 (n=1645) | 0.462 (n=1562) | 0.397 (n=1064) |
| avg (Articles Actor2) | 0.547 (n=1893) | 0.623 (n=1645) | 0.537 (n=1562) | 0.44 (n=1064) |
| std (AvgTone Actor1) | 0.539 (n=1893) | 0.574 (n=1645) | 0.457 (n=1562) | 0.402 (n=1064) |
| std (GSi Self) | 0.241 | 0.177 | 0.09 | -0.008 |

| | | | | |
|----------------------|------------------------------|--------------------|--------------------|------------------------------|
| | (n=1893) | (n=1645) | (n=1562) | (n=1064, p=0.787) |
| avg (AvgTone Actor1) | 0.191 (n=1893) | 0.208 (n=1645) | 0.2 (n=1562) | 0.337 (n=1064) |
| avg (AvgTone Actor2) | 0.174 (n=1893) | 0.182 (n=1645) | 0.187 (n=1562) | 0.348 (n=1064) |
| avg (AvgTone Self) | 0.167 (n=1893) | 0.068 (n=1645) | 0.095 (n=1562) | 0.232 (n=1064) |
| std (AvgTone Self) | 0.032 (n=1893, p=0.17) | 0.086 (n=1645) | 0.064 (n=1562) | 0.015 (n=1064, p=0.63) |
| avg (mGSi Actor1) | -0.13 (n=1893) | -0.12 (n=1645) | -0.233 (n=1562) | -0.149 (n=1064) |
| avg (GSi Actor1) | -0.149 (n=1893) | -0.13 (n=1645) | -0.234 (n=1562) | -0.155 (n=1064) |
| avg (mGSi Actor2) | -0.17 (n=1893) | -0.134 (n=1645) | -0.254 (n=1562) | -0.156 (n=1064) |
| avg (GSi Actor2) | -0.177 (n=1893) | -0.142 (n=1645) | -0.248 (n=1562) | -0.157 (n=1064) |
| avg (GSi Self) | -0.181 (n=1893) | -0.134 (n=1645) | -0.162 (n=1562) | -0.083 (n=1064) |
| avg (mGSi Self) | -0.182 (n=1893) | -0.17 (n=1645) | -0.19 (n=1562) | -0.137 (n=1064) |

Table 5. Spearman correlation coefficients between selected WDI (Economic Policy & Debt) and FDI net inflows per income category (sample size given in brackets, all p-values below 5%)

| Indicator / Income category | High | Upper middle | Lower middle | Low |
|--|-------------------|---------------------|---------------------|------------------|
| Primary income payments (BoP, current US\$) | 0.793 (n=1893) | 0.789 (n=1625) | 0.695 (n=1636) | 0.546 (n=799) |
| Imports of goods, services and primary income (BoP, current US\$) | 0.788 (n=1893) | 0.817 (n=1625) | 0.737 (n=1636) | 0.645 (n=799) |
| Imports of goods and services (BoP, current US\$) | 0.781 (n=1893) | 0.815 (n=1625) | 0.735 (n=1636) | 0.643 (n=799) |
| Service exports (BoP, current US\$) | 0.781 (n=1893) | 0.842 (n=1625) | 0.701 (n=1636) | 0.6 (n=799) |
| Goods imports (BoP, current US\$) | 0.78 (n=1893) | 0.809 (n=1625) | 0.734 (n=1636) | 0.63 (n=799) |
| Imports of goods and services (current US\$) | 0.776 (n=2074) | 0.802 (n=1782) | 0.718 (n=1712) | 0.606 (n=998) |
| Charges for the use of intellectual property, payments (BoP, current | 0.776 (n=1663) | 0.744 (n=1475) | 0.507 (n=1337) | 0.527 (n=662) |

| | | | | |
|---|-------------------|-------------------|-------------------|------------------|
| US\$) | | | | |
| Exports of goods, services and primary income (BoP, current US\$) | 0.775 (n=1893) | 0.795 (n=1625) | 0.729 (n=1631) | 0.622 (n=799) |
| Exports of goods and services (BoP, current US\$) | 0.773 (n=1893) | 0.794 (n=1625) | 0.723 (n=1636) | 0.62 (n=799) |
| Exports of goods and services (current US\$) | 0.768 (n=2074) | 0.786 (n=1782) | 0.711 (n=1712) | 0.571 (n=998) |
| Service imports (BoP, current US\$) | 0.76 (n=1893) | 0.81 (n=1625) | 0.705 (n=1636) | 0.633 (n=799) |
| Household final consumption expenditure, PPP (current international \$) | 0.702 (n=1299) | 0.819 (n=1019) | 0.655 (n=1065) | 0.619 (n=582) |

Table 6. Spearman correlation coefficients between selected WDI (Financial Sector) and FDI net inflows per income category (sample size and p-values above 5% given in brackets)

| Indicator / Income category | High | Upper middle | Lower middle | Low |
|---|-------------------|-------------------|-------------------|-------------------------------|
| Broad money (current LCU) | 0.683 (n=1354) | 0.619 (n=1778) | 0.511 (n=1760) | 0.434 (n=1074) |
| Stocks traded, total value (current US\$) | 0.667 (n=1247) | 0.725 (n=638) | 0.671 (n=366) | -0.539 (n=10, p=0.108) |
| Market capitalization of listed domestic companies (current US\$) | 0.666 (n=1102) | 0.709 (n=530) | 0.778 (n=294) | -0.648 (n=10) |
| Net domestic credit (current LCU) | 0.629 (n=1624) | 0.563 (n=1768) | 0.51 (n=1731) | 0.375 (n=1070) |
| Stocks traded, total value (% of GDP) | 0.612 (n=1235) | 0.554 (n=634) | 0.451 (n=366) | -0.552 (n=10, p=0.098) |
| Domestic credit to private sector (% of GDP) | 0.548 (n=1575) | 0.185 (n=1742) | 0.282 (n=1744) | 0.053 (n=1025, p=0.088) |
| Domestic credit provided by financial sector (% of GDP) | 0.51 (n=1571) | 0.126 (n=1730) | 0.248 (n=1728) | -0.094 (n=1025) |
| Net foreign assets (current LCU) | 0.313 (n=1624) | 0.512 (n=1781) | 0.412 (n=1760) | 0.412 (n=1074) |
| Stocks traded, turnover ratio of domestic shares (%) | 0.458 (n=1075) | 0.499 (n=510) | 0.468 (n=281) | -0.491 (n=10, p=0.15) |
| Consumer price index (2010 = 100) | 0.448 (n=2017) | 0.439 (n=1660) | 0.454 (n=1688) | 0.546 (n=872) |

| | | | | |
|---|-------------------|-------------------|-------------------|-------------------|
| GDP deflator (base year varies by country) | 0.351 (n=2122) | 0.4 (n=1850) | 0.355 (n=1856) | 0.44 (n=1089) |
| Domestic credit to private sector by banks (% of GDP) | 0.508 (n=1576) | 0.188 (n=1743) | 0.304 (n=1748) | 0.085 (n=1027) |
| Official exchange rate (LCU per US\$, period average) | 0.134 (n=2086) | 0.152 (n=1864) | 0.195 (n=1865) | 0.293 (n=1169) |

Table 7. Spearman correlation coefficients between selected WDI (Private Sector & Trade) and FDI net inflows per income category (sample size given in brackets, all p-values below 5%)

| Indicator / Income category | High | Upper middle | Lower middle | Low |
|--|----------------|---------------------|---------------------|----------------|
| Investment in telecoms with private participation (current US\$) | 0.831 (n=86) | 0.815 (n=759) | 0.665 (n=717) | 0.561 (n=314) |
| Commercial service exports (current US\$) | 0.782 (n=1893) | 0.841 (n=1625) | 0.706 (n=1636) | 0.61 (n=799) |
| Merchandise imports (current US\$) | 0.772 (n=2277) | 0.793 (n=1901) | 0.733 (n=1886) | 0.578 (n=1186) |
| Commercial service imports (current US\$) | 0.769 (n=1893) | 0.813 (n=1625) | 0.717 (n=1636) | 0.624 (n=799) |
| International tourism, receipts for passenger transport items (current US\$) | 0.766 (n=907) | 0.73 (n=789) | 0.512 (n=763) | 0.41 (n=359) |
| Merchandise imports by the reporting economy (current US\$) | 0.766 (n=2252) | 0.811 (n=1822) | 0.727 (n=1810) | 0.576 (n=1202) |
| International tourism, expenditures for travel items (current US\$) | 0.75 (n=1192) | 0.881 (n=1030) | 0.681 (n=1009) | 0.588 (n=470) |
| Merchandise exports (current US\$) | 0.738 (n=2286) | 0.741 (n=1901) | 0.713 (n=1886) | 0.513 (n=1186) |
| Merchandise exports by the reporting economy (current US\$) | 0.729 (n=2237) | 0.743 (n=1842) | 0.703 (n=1821) | 0.495 (n=1206) |
| International tourism, receipts for travel items (current US\$) | 0.727 (n=1203) | 0.814 (n=1038) | 0.623 (n=1025) | 0.511 (n=466) |
| International tourism, expenditures (current US\$) | 0.726 (n=1220) | 0.877 (n=1048) | 0.724 (n=1036) | 0.581 (n=493) |
| International tourism, receipts (current US\$) | 0.687 (n=1271) | 0.822 (n=1062) | 0.668 (n=1070) | 0.525 (n=522) |
| International tourism, expenditures for passenger transport items (current US\$) | 0.673 (n=896) | 0.804 (n=852) | 0.708 (n=897) | 0.386 (n=413) |

| | | | | |
|---|-------------------|-------------------|-------------------|-------------------|
| International tourism, number of arrivals | 0.657 (n=1298) | 0.83 (n=1057) | 0.69 (n=1039) | 0.507 (n=520) |
| Import value index (2000 = 100) | 0.353 (n=1616) | 0.445 (n=1645) | 0.48 (n=1645) | 0.557 (n=1011) |
| Import volume index (2000 = 100) | 0.162 (n=1175) | 0.382 (n=1321) | 0.494 (n=1439) | 0.515 (n=932) |

Table 8. Spearman correlation coefficients between selected WDI (Infrastructure) and FDI net inflows per income category (sample size given in brackets, all p-values below 5%)

| Indicator / Income category | High | Upper middle | Lower middle | Low |
|---|-------------------|-------------------|-------------------|-------------------|
| ICT service exports (BoP, current US\$) | 0.775 (n=1367) | 0.795 (n=1081) | 0.651 (n=955) | 0.541 (n=392) |
| High-technology exports (current US\$) | 0.692 (n=1370) | 0.793 (n=988) | 0.613 (n=803) | 0.582 (n=426) |
| Air transport, passengers carried | 0.681 (n=1967) | 0.653 (n=1588) | 0.51 (n=1661) | 0.232 (n=902) |
| Trademark applications, total | 0.681 (n=1701) | 0.741 (n=1247) | 0.65 (n=1165) | 0.385 (n=463) |
| Scientific and technical journal articles | 0.667 (n=728) | 0.832 (n=747) | 0.728 (n=719) | 0.444 (n=424) |
| Mobile cellular subscriptions | 0.655 (n=2162) | 0.758 (n=1838) | 0.667 (n=1836) | 0.639 (n=1156) |
| Fixed telephone subscriptions | 0.653 (n=2227) | 0.764 (n=1813) | 0.645 (n=1808) | 0.388 (n=1113) |
| Trademark applications, direct nonresident | 0.65 (n=1645) | 0.774 (n=1199) | 0.667 (n=1130) | 0.455 (n=418) |
| Liner shipping connectivity index (maximum value in 2004 = 100) | 0.641 (n=693) | 0.759 (n=558) | 0.602 (n=486) | 0.35 (n=208) |
| Fixed broadband subscriptions | 0.61 (n=965) | 0.739 (n=743) | 0.669 (n=664) | 0.447 (n=299) |
| Individuals using the Internet (% of population) | 0.298 (n=1422) | 0.385 (n=1251) | 0.362 (n=1137) | 0.467 (n=677) |

Table 9. Spearman correlation coefficients between selected WDI (Poverty) and FDI net inflows per income category (sample size and p-values above 5% given in brackets)

| Indicator / Income category | High | Upper middle | Lower middle | Low |
|---|-----------------------|-------------------|-------------------|---------------------------|
| Rural poverty headcount ratio at national poverty lines (% of rural population) | 0.094 (n=22, p=0.676) | -0.184 (n=158) | -0.275 (n=159) | -0.258 (n=58, p=0.051) |

| | | | | |
|---|-------------------------------|-------------------------------|-------------------------------|------------------------------|
| GINI index (World Bank estimate) | 0.027 (n=378, p=0.595) | -0.073 (n=459, p=0.12) | -0.125 (n=358) | -0.305 (n=91) |
| Income share held by highest 20% | 0.019 (n=378, p=0.71) | -0.063 (n=459, p=0.175) | -0.117 (n=358) | -0.29 (n=91) |
| Income share held by highest 10% | 0.009 (n=378, p=0.862) | -0.068 (n=459, p=0.145) | -0.109 (n=358) | -0.296 (n=91) |
| Poverty gap at \$1.90 a day (2011 PPP) (%) | -0.019 (n=378, p=0.712) | -0.238 (n=469) | -0.333 (n=389) | -0.311 (n=93) |
| Poverty headcount ratio at \$1.90 a day (2011 PPP) (% of population) | -0.088 (n=378, p=0.088) | -0.251 (n=469) | -0.329 (n=388) | -0.283 (n=93) |
| Poverty headcount ratio at \$3.20 a day (2011 PPP) (% of population) | -0.176 (n=378) | -0.274 (n=469) | -0.309 (n=389) | -0.251 (n=93) |
| Poverty headcount ratio at \$5.50 a day (2011 PPP) (% of population) | -0.217 (n=378) | -0.289 (n=469) | -0.282 (n=389) | -0.156 (n=93, p=0.135) |
| Poverty headcount ratio at national poverty lines (% of population) | -0.251 (n=106) | -0.281 (n=303) | -0.302 (n=243) | -0.163 (n=67, p=0.187) |
| Urban poverty headcount ratio at national poverty lines (% of urban population) | -0.67 (n=26) | -0.193 (n=167) | -0.226 (n=160) | -0.268 (n=58) |
| Poverty gap at national poverty lines (%) | -0.729 (n=12) | -0.415 (n=94) | -0.26 (n=143) | -0.327 (n=57) |
| Urban poverty gap at national poverty lines (%) | -0.793 (n=14) | 0.106 (n=58, p=0.427) | -0.118 (n=115, p=0.211) | -0.216 (n=52, p=0.124) |

Table 10. Spearman correlation coefficients between selected Doing Business indicators and FDI net inflows per income category (sample size and p-values above 5% given in brackets)

| Indicator / Income category | High | Upper middle | Lower middle | Low |
|---|------------------|------------------|------------------|-------------------------------|
| Resolving Insolvency: recovery rate (cents on the dollar) | 0.416 (n=690) | 0.199 (n=675) | 0.197 (n=697) | 0.316 (n=387) |
| Extent of disclosure index (0 to 10) | 0.394 (n=613) | 0.567 (n=604) | 0.194 (n=617) | -0.055 (n=342, p=0.308) |
| Private credit bureau coverage (% of adults) | 0.354 (n=638) | 0.471 (n=613) | 0.261 (n=624) | 0.06 (n=366, |

| | | | | |
|--|-------------------------------|-------------------------------|------------------------------|-------------------------------|
| | | | | p=0.253) |
| Cost to build a warehouse (% of income per capita) | 0.295 (n=613) | 0.294 (n=595) | 0.347 (n=616) | 0.051 (n=328, p=0.358) |
| Labor tax and contributions (%) | 0.178 (n=597) | 0.572 (n=572) | 0.368 (n=577) | 0.048 (n=340, p=0.376) |
| Cost to register property (% of property value) | 0.078 (n=652) | -0.405 (n=617) | 0.039 (n=634, p=0.325) | -0.184 (n=366) |
| Total tax rate (% of profit) | 0.019 (n=613, p=0.646) | 0.399 (n=604) | 0.098 (n=617) | -0.285 (n=340) |
| Time to prepare and pay taxes (hours) | -0.006 (n=613, p=0.887) | 0.372 (n=604) | 0.32 (n=617) | -0.042 (n=340, p=0.437) |
| Public credit registry coverage (% of adults) | -0.009 (n=638, p=0.825) | 0.319 (n=613) | 0.18 (n=624) | 0.088 (n=366, p=0.094) |
| Time required to build a warehouse (days) | -0.057 (n=613, p=0.161) | 0.333 (n=595) | 0.092 (n=616) | 0.086 (n=328, p=0.12) |
| Cost to start a business (% of income per capita) | -0.137 (n=690) | -0.337 (n=675) | -0.205 (n=697) | -0.314 (n=387) |
| Time required to register property (days) | -0.153 (n=652) | -0.309 (n=617) | -0.125 (n=634) | -0.142 (n=366) |
| Procedures required to build a warehouse (number) | -0.265 (n=613) | 0.521 (n=595) | 0.102 (n=616) | 0.107 (n=328, p=0.054) |
| Time required to start a business (days) | -0.298 (n=690) | -0.068 (n=675, p=0.078) | -0.213 (n=697) | -0.084 (n=387, p=0.1) |
| Tax payments (number) | -0.328 (n=613) | -0.322 (n=604) | 0.206 (n=617) | -0.014 (n=340, p=0.797) |
| Time to resolve insolvency (years) | -0.381 (n=648) | 0.002 (n=589, p=0.955) | 0.137 (n=536) | -0.17 (n=311) |

CSECS 2018, pp. 061 - 075

Computer Science and Education in Computer Science
14th Annual International Conference
ISSN 2603-4794

June 29 – 30, 2018, Boston, USA

SOFTWARE SYSTEM READABILITY METRICS


Latchezar Tomov, PhD,

New Bulgarian University, Department of Informatics

***Abstract:** In previous works we presented conceptual model of software understandability, based on the notion that all software are systems and the understanding of code is related to their structure and not only the sum of their code. We further this work by extending the system to include its observer and propose software system readability metrics based on a model of the observer and the software system using graph theory and information theory.*

***Keywords:** software quality, readability, understandability, system design, graph theory, probability theory*

***ACM Classification Keywords:** D.2.8 - Metrics (This is just an example, please use the correct category and subject descriptors for your submission. The ACM Computing Classification Scheme: <http://www.acm.org/class/1998/>)*



Introduction

In previous work [Tomov and Ivanova, 2015] we presented conceptual model of software understandability in which we presented several layers of readability, related to the hierarchical nature of systems. System readability is different from the readability of its elements and that is related to its structure and level of disorder independently from the same properties of its subsystems and elementary blocks. We want to formalize this model here in order to define metrics for system readability. The first step to that is to extend the system to include the observer and the way he/she process information – a simplified model of mental mappings that are being loaded into the working memory of the reader in dependence with the changes in programming languages that need different *modes* of thinking.

In these and previous research we distinguish the readability and the complexity of the software – they are correlated but are not the same thing. Readability is as much property of the observer as it is property of the system, while complexity can be defined in terms of probability theory and be linked directly to reliability. The involvement of the observer in the system does not necessary translates into subjective approach to readability, since the way the human brain operates is not as unique per person to require different model for every individual. An excellent overview of the domain is given in [Kahneman, 2011]. What the inclusion of the observer of our system adds is a cost function from switching into one mental mapping into another when a different programming language and or programming paradigms and level of programming are introduced during “reading” the system. We propose metric which is as abstract and independent from the choice of classical code readability metrics [Buse and Weimer,

2008], [Buse and Weimer, 2010], [Bjorstler et.al, 2015], [Namani and Kumar, 2012], [Sedano, 2016] and all metrics in the reviews of [Pahal and Chillar, 2017] and [Tashtoush et.al, 2013] as possible. The metrics are either based on entropy of tokens [Cholewa, 2017], or on expert judgement [Sedano, 2016], or some natural language readability measure [Pahal and Chillar, 2017]], or statistically derived [Pahal and Chillar, 2017], [Tashtoush et.al, 2013]. All of them and any of them can be used as “volume metrics” when we calculate software system readability based on our simple, yet abstract, mental model.

Software system structure

The structure of a **software system** (here we understand a **system of programs** under that term) is a network of software components, called “modules”, each of which is written in one programming language, which can be different for each module. An example is a web service in .NET framework that depend on different libraries, some in Visual Basic, some in C, that also depend on different libraries, etc. A limitation on such software system is that most often it is *directed acyclic graph* – because no circular dependencies *should* exist. We call the lack of cycles in the graph a necessary condition for the software system to be **proper system**.

A software system can depend on other software systems, which can be viewed as encapsulated as modules with their relevant properties estimated separately.

An example of graph of software system, depending on systems is given in Fig.1:

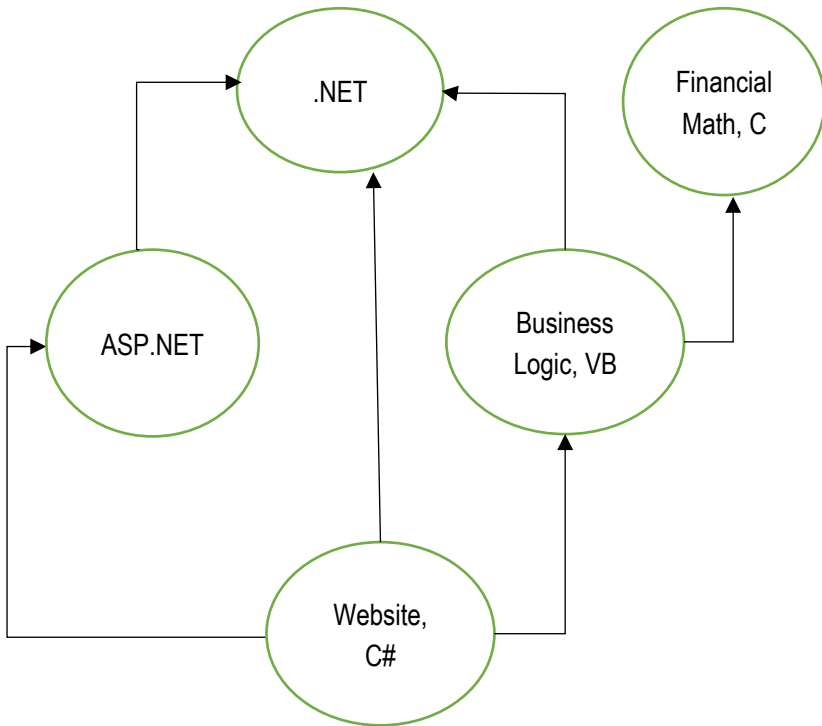


Figure 1 Software system of financial web site in encapsulated form – every module is a subsystem

Software system and the observer

One of the main problems in software engineering is long-term support of existing software systems [Rugaber, 2000], [Collar and Valerdi, 2014], which comprise of different modules. They are written at different times with different programming languages and using different programming paradigms and different levels of abstractions and coding conventions, a.k.a. “styles”. Most developers have specialized in one language, using particular

style and set of programming paradigms, for which they have mapped symbols and their meaning in conceptual networks which are most developed because of the specialization and the constant practice - these particular networks are closes to the mind. Trying to read code that requires different conceptual networks means that they have to switch the mode of work and have to “load” other conceptual networks, therefore reading of modules in other languages, style and using different sets of paradigms requires disproportional to the volume of the code effort – an effort is necessary to adapt to the new way of thinking, for example changing from object-oriented to functional programming. New dictionaries with data need to be used, if for example one switches from C# to Java where paradigms are mostly shared, but syntax is different. We consider readability of program system with regards to a model of cognition with **finite capacity** of working memory and **modes** of thinking that change proportionally to the changes in programming concepts and languages in different parts of the code. The effort to read code is therefore not only cognitive effort in the presence of static mapping of words and rules in formal languages which is proportional to the volume of code, but also the **effort** to adapt the mapping to the changes in the code which is proportional to the number of changes from one part to another.

This mapping can be viewed as a function $G(L)$, that has its domain the set of all words over an alphabet that define some formal language L . This function takes into account separately all programming paradigms and other higher level constructs but which also are expressed over that same set. The co-domain of the function is the set of all words in the natural language (English can be used as this language for simplicity)

Software system heterogeneity measures

Def. Programming concepts – an extension of programming paradigms that includes the usage of pointer arithmetic and explicit work with memory as a **separate** concept.

We will define software modules as elementary building blocks of software systems

Def. Software module $M_j = (L_j, C_{L_j}, W_j, F_{C_{L_j}})$ is an ordered tuple, consisting of formal language L_j , set of programming concepts belonging to that language $C_{L_j} = \{C_1, C_2 \dots C_{m_j}\}$, a volume measure W_j and a set of frequencies of usage of programming concepts $F_{C_{L_j}} = \{F_{C_1}, F_{C_2} \dots F_{C_{m_j}}\}$ where the frequency of usage of programming concepts is the relative volume of the module written with it:

$$F_{C_i}^j = \frac{W_{C_i}}{W_j} \quad (1)$$

An important constraint when estimating frequencies is that they should behave as probabilities:

$$\sum_{i=1}^{m_j} F_{C_i}^j = 1 \quad (2)$$

Def. Software module heterogeneity is the Shannon entropy of programming concepts usage inside of a module:

$$H(M_j) = - \sum_{i=1}^{m_j} F_{C_i}^j \log(F_{C_i}^j) \quad (3)$$

This has useful application in software project management as a tool to understand if the software modules are being built with a coherent style of programming or are randomly evolving as a consequence from certain deficits of qualification and control.

Software modules comprise program systems

Def. Program system – an ordered pair $S = (V, E)$, comprised of set of vertices V and edges E with adjacency matrix $A = \{a_{ij}\}$, $i, j = 1 \dots n$. The set of vertices V is a set of software modules $V \equiv M = \{M_1, M_2 \dots M_n\}$.

Def. The number n is called the size of the system

The set of formal languages of the system is the union of the sets of the software modules $L_g = \bigcup_{j=1}^n L_j$

The set of programming concepts is the union of the sets of programming concepts $C_g = \bigcup_{j=1}^n C_{L_j}$ with cardinality bounded by the sum of cardinalities of the sets when all programming concepts are different:

$$|C_g| \leq m = \sum_{j=1}^n m_j \quad (4)$$

Def. Program system heterogeneity is the Shannon entropy of programming concepts usage inside of the system:

$$H(S) = -\sum_{j=1}^n \left(\sum_{i=1}^{m_j} F_{C_i}^j \log \left(F_{C_i}^j \right) \right) \quad (5)$$

$$F_{C_i}^j = F_{C_i}^j \cdot \frac{W_j}{W} = \frac{W_{C_i}}{W_j} \cdot \frac{W_j}{W} = \frac{W_{C_i}}{W} \leq F_{C_i}^j \quad (6)$$

$$\sum_{j=1}^n \sum_{i=1}^{m_j} F_{C_i}^j = 1 \quad (7)$$

Statement 1: The program system heterogeneity is at most the sum of software modules heterogeneities:

$$H(S) \leq \sum_{j=1}^n H(M_j) \quad (8)$$

$$\sum_{j=1}^n H(M_j) = -\sum_{j=1}^n \left(\sum_{i=1}^{m_j} F_{C_i}^j \log \left(F_{C_i}^j \right) \right) \quad (9)$$

Since $F_{C_i}^j \leq F_{C_i}^j$ we have

$$-F_{C_i}^j \log \left(F_{C_i}^j \right) \leq -F_{C_i}^j \log \left(F_{C_i}^j \right) \quad (10)$$

From there (8) holds.

The equality is reached when the entropy is maximal

$$H_{max}(S) = 2^{n.m} \quad (11)$$

Then all modules have the same volume, all have different programming concepts (27) - and every concept has equal share of $\frac{W}{m.n}$.

$$|C_g|_{H(S)=H_{max}(S)} = n.m \quad (12)$$

Otherwise, if there are modules with much larger volume measure than others, written in small number of programming concepts, these concepts will be much more frequent in the system than they are in the modules – for example, if one module has two programming concepts with 50% share, but the module volume measure is 90% of the system volume measure, those two programming concepts will have at least 45% share each.

Def. Program system **language heterogeneity** is the entropy of formal languages in the system:

$$H_{L_g}(S) = -\sum_{j=1}^n \frac{W_j}{W} \log\left(\frac{W_j}{W}\right) \quad , \quad W = \sum_{j=1}^n \frac{W_j}{W} \quad (28)$$

Statement 2: The entropy is maximal, when the cardinality of the set of formal languages of the system is equal to its size (20) and is minimal, when the cardinality of the set of formal languages is equal to one (21)

$$|L_g|_{H_{L_g}(S)=H_{L_g}(S)_{max}} = n \quad (29)$$

$$|L_g|_{H_{L_g}(S)=H_{L_g}(S)_{min}} = 1 \quad (30)$$

This is so, since each software module by definition has one formal language and the entropy is minimal when all modules share the same language and its volume measure is equal to the

volume measure of the system. The entropy is maximal if each module has different language and then the size of the system is the same as the cardinality of its set of formal languages L_g .

Software system readability measures

Software or program system reading we view as a process of traversal of a graph with weights, proportional to individual volume measures (or the traditional readability ratings) and to the effort of the reader to switch between different modes of thinking by loading data into his working memory, related to different words and rules in different formal languages and to different programming concepts. Switching from C# to Java needs effort independent of the effort to read these programs, switching from high level programming to pointer arithmetic and direct usage of addresses is similar, as well for example switching from imperative to functional paradigms and from class based OOP to prototypal, since all these can be intermixed.

In order to measure Total Software System Readability we ascribe weights $\omega = \{\omega_{ij}\}$ to the adjacency matrix $A = \{a_{ij}\}$ to form an adjacency matrix $A^* = \{a_{ij}^*\}$ of a weighted version of the system's graph $G = (V, E)$

$$a_{ij}^* = \omega_{ij} \cdot a_{ij} \quad (13)$$

The weights ω_{ij} are functions of the volume measure for the $j - th$ module W_j and the cardinalities of set differences between the programming concept sets $|C_{L_{i-j}}|$ and the formal language sets $|L_{i-j}|$

$$C_{L_{i-j}} = C_{L_i} \setminus C_{L_j} \quad (14)$$

$$L_{i-j} = L_i \setminus L_j \quad (15)$$

$$\omega_{ij} = F \left(W_j, |C_{L_{i-j}}|, |L_{i-j}| \right) \quad (16)$$

The idea behind these weights is that the reader of the code has limited working memory and each programming concept requires different mode of thinking in order to translate the code into natural language – a different **map** is required. The difference of languages is independent from the programming concepts, because it is difference between two sets of words in an alphabet behind which there are different meanings of the same words and different syntax. Two modules using the same sets of programming concepts but written in different languages require remapping. Here we propose two measures - a linear functionality (17) and a Euclidean norm (18):

$$\omega_{ij_l} = W_j \left(1 + |C_{L_{i-j}}| + f(|L_{i-j}|) \right) \quad (17)$$

$$\omega_{ij_e} = W_j \left(\sqrt{1 + |C_{L_{i-j}}|^2 + f(|L_{i-j}|)^2} \right) \quad (18)$$

The function $f(L_{i-j})$ and its form are questions of future research in cognitive science. Here we propose this version:

$$f(L_{i-j}) = \begin{cases} 0, & |L_{i-j}| = 0 \\ \alpha |L_{i-j}| \geq |L_{i-j}|, & |L_{i-j}| > 0 \end{cases} \quad (19)$$

This is general definition that will apply even for multilingual software modules.

Def. Total Program System Readability is the reciprocal of the cost of the system with adjacency matrix $A^* = \{a_{ij}^*\}$:

$$R_s = \frac{1}{\sum_{i,j} a_{ij}^*} = \frac{1}{\sum_{i,j} a_{ij} \omega_{ij}} \quad (20)$$

This is the reciprocal of sum of all node strengths

$$s_i = \sum_{j=1}^n a_{ij} \omega_{ij} \quad (21)$$

$$R_s = \frac{1}{\sum_{i,j} a_{ij} \omega_{ij}} = \frac{1}{\sum_i s_i} \quad (22)$$

The measurement of the cost of the weighted graph reflects our understanding that during reading of software system when one must understand its behavior – the “how” readability, he or she has to traverse all dependent modules more than once, because each software module has to be read in the context of the previous module that depend on it for a specific purpose and because the reader doesn’t have infinite memory. Switching from different languages and programming concepts multiplies the effort to read even for developers who are fluent in all of them. Thus, program system readability becomes structural property, not only a volume property. The weight of reading all code is not only determined by its volume, no matter how that is measured, but by its **distribution** and **connectivity**. Breaking software into modules has multiple advantages for reliability, and extensibility, but does not decrease the effort for studying the system – and in some cases may substantially increase it.

The three stages of comparison of different program systems are first by their total readability, second by their language entropy and third, by their programming concepts entropy. Systems with the same total readability may differ – one can be with simpler structure and smaller sets of formal languages and concepts but may have larger entropies that make its reading similar challenge.

A main challenge from this measure is to study the optimal topology of program systems, which maximizes (20) for fixed sets L_g and C_g and variable α in (20). Another, long term

challenge is to find the best fit for (20) based on cognitive research that would represent correctly the burden of switching between two formal languages. The correlation between readability and the two different entropies of the system is also a problem for future research.

Conclusion

The difference between a system and a set is the interaction between elements for achieving a common goal. Every software system has behavior largely determined by its structure which is incorporated in reliability measures [Jatain and Mehta, 2014]. We try to do the same with readability by proposing a specific metric for readability of multilingual software systems and by taking into account the observer, the reader and his/hers specifics (finite memory, different methods). Our future work is related both to implementation of this method and to searching for optimal topology for maximal readability with constraints from reliability. Implementation issues are related mostly with identification of programming concepts usage and frequency estimations, since a single block of code can express multiple programming concepts and programming concepts implementation vary from language to language. There are possibilities for expert identification, with or without machine learning as an assistant similar to what is done in [Goeues et.al, 2013] and [Long and Rinard, 2016], which means probabilistic approach, mixed with expert review. The estimation of the model effectiveness will again be verified with statistical research, again with expert reviews and the acceptance from the community.

Bibliography

- [Buse and Weimer, 2010] R. P. L. Buse and W. R. Weimer, “A metric for software readability,” in International Symposium on Software Testing and Analysis
- [Buse and Weimer, 2008] R.P.L. Buse and W.R.Weimer, A metric for Software Readability, ISSTA '08 Proceedings of the 2008 international symposium on Software testing and analysis pp 121-130, Seattle, WA, USA — July 20 - 24, 2008, DOI: 10.1145/1390630.1390647
- [Bjorstler et.al, 2015] J. Bo`rstler et. al, Beauty and the Beast: on the readability of object-oriented example programs, Software Quality Control 24(2), February 2015, DOI 10.1007/s11219-015-9267-5
- [Cholewa, 2017] M.Cholewa. Shannon information entropy as complexity metric of source code. 2017, 468-471. 10.23919/MIXDES.2017.8005255.
- [Collar and Valerdi, 2014] E. Collar, Jr and R. Valerdi, Role of Software Readability on Software Development Cost, 2014
- [Goeues et.al, 2013] C. Le Goeues et.al, Current Challenges in Automatic Software Repair, Software Qual J (2013) 21: 421. <https://doi.org/10.1007/s11219-013-9208-0>
- [Jatrain and Mehta, 2014] A. Jatrain and Y. Mehta, Metrics and Models for Software Reliability: A Systematic Review. IEEE conference, ICICT At: Ghaziabad, February 2014, 10.1109/ICICT.2014.6781281.

-
- [Kahneman, 2013] D. Kahneman, Thinking, Fast and Slow, Farrar, Straus and Giroux; 1st edition (April 2, 2013) ISBN-13: 978-0374533557
- [Long and Rinard, 2016] F. Long and M. Rinard, Automatic Patch Generation by Learning Correct Code, POPL '16 Proceedings of the 43rd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, pp 298-312, St. Petersburg, FL, USA — January 20 - 22, 2016
ACM New York, NY, USA ©2016
doi>10.1145/2837614.2837617
- [Namani and Kumar, 2012] R. Namani and J. Kumar, A New Metric for Code Readability, IOSR Journal of Computer Engineering ISSN: 2278-0661, ISBN: 2278-8727 Volume 6, Issue 6 Nov. - Dec. 2012, pp 44-48
- [Pahal and Chillar, 2017] A. Pahal. R.S.Chillar, Code Readability: A Review of Metrics for Software Quality, International Journal of Computer Trends and Technology (IJCTT) – Volume 46 Number 1- April 2017
- [Posnett et.al., 2011] D. Posnett, A.Hindle and P. Devanbu, A Simpler Model of Software Readability, MSR '11 Proceedings of the 8th Working Conference on Mining Software Repositories pp 73-82, Waikiki, Honolulu, HI, USA — May 21 - 22, 2011, DOI: 10.1145/1985441.1985454
- [Rugaber, 2000] S. Rugaber, The use of domain knowledge in program understanding. Annals of Software Engineering9: 143-192, 2000
- [Sedano, 2016] T. Sedano, Code Readability Testing, an Empirical Study, DOI10.1109/CSEET.2016.36

- [Selvarani et.al, 2009] R. Selvarani et.al, Software Metrics Evaluation Based on Entropy, Handbook of Research on Software Engineering and Productivity Technologies: Implications of Globalization, IGI Global, pages 139-148, 2009
DOI: 10.4018/978-1-60566-731-7.ch011
- [Tashtoush et.al, 2013] Y.Tashtoush et.al, Impact of Programming Features on Code Readability, International Journal of Software Engineering and Its Applications Vol.7, No.6 pp.441-458, 2013
<http://dx.doi.org/10.14257/ijseia.2013.7.6.38>
- [Tomov and Ivanova, 2015]. Latchezar P. Tomov, Valentina Ivanova, Software understandability model, CSECS 2015, June 4-7 2015, Boston, MA USA

Authors' Information



Latchezar TOMOV, PhD, assistant professor at NBU-Sofia, Bulgaria

Major Fields of Scientific Research: Control theory, software quality, project management, applied mathematics

CSECS 2018, pp. 077 - 091

Computer Science and Education in Computer Science
14th Annual International Conference
ISSN 2603-4794

June 29 – 30, 2018, Boston, USA

AESTHETIC METRICS AND THE EVOLUTION OF C LANGUAGE FAMILY

Latchezar Tomov

New Bulgarian University, Department of Informatics

Abstract: In previous works he defined *the notions of intellectual aesthetic and proposed an ordinal measure of the aesthetic of function. We further this work in order to clarify the relationships between aesthetics and readability on both code and system level and to define aesthetics of code (previously we defined the aesthetics of functionality). We analyze the main drives behind the evolution of C languages, one of which is the aesthetic motive, at the expense of readability*

Keywords: *software quality, readability, software aesthetics, system design, graph theory.*

ACM Classification Keywords: *D.2.8 - Metrics (This is just an example, please use the correct category and subject descriptors for your submission. The ACM Computing Classification Scheme: <http://www.acm.org/class/1998/>)*

Introduction

Aesthetics and readability of code matter long before the first digital computer was created. Proper notations such as the invention of Gauss – the sign for congruence “ \equiv ” for modular arithmetics in *Disquisitiones Arithmeticae* [Gauss, 1801] and the whole concept of modular arithmetics simplify and shorten existing proofs – for example the divisibility of 9:

Explanation without modular arithmetics

If the sum of digits of a number is divisible by 9, the number itself is divisible by 9. We can obtain the result by the following transformation of the digital representation:

$$\sum_{i=0}^n a_i 10^i = \sum_{i=0}^n a_i (10^i - 1 + 1) = \sum_{i=0}^n a_i (10^i - 1) + \sum_{i=0}^n a_i \quad (1)$$

Here we have to explain further that $10^i - 1$ is represented by

$$10^i - 1 = \sum_{j=0}^{i-1} 9 \cdot 10^j = 9 \sum_{j=0}^{i-1} 10^j \quad (2)$$

which is divisible by 9. Thus (1) becomes:

$$\sum_{i=0}^n a_i (10^i - 1) + \sum_{i=0}^n a_i = \sum_{i=0}^n a_i (9 \sum_{j=0}^{i-1} 10^j) + \sum_{i=0}^n a_i \quad (3a)$$

Since the sum of digits is divisible by 9

$$\sum_{i=0}^n a_i = 9 \cdot m \quad (3b)$$

and in the first sum we have $n + 1$ numbers that are all divisible by 9, the whole number is divisible by 9

$$\sum_{i=0}^n a_i (10^i - 1) + \sum_{i=0}^n a_i = 9(\sum_{i=0}^n a_i (\sum_{j=0}^{i-1} 10^j) + m) \quad (4)$$

Explanation with modular arithmetic:

The number 10 is congruent with 1 by modulo of 9:

$$10 \equiv 1 \pmod{9} \quad (5)$$

By the rules of congruence $10 \cdot 10 \equiv 1 \cdot 1 \pmod{9}$ and we can see that:

$$10^i \equiv 1 \pmod{9} \quad (6)$$

Thus the sum of all such number by these rules is congruent with one by modulo of 9 and

$$\sum_{i=0}^n a_i 10^i \equiv \sum_{i=0}^n a_i \cdot 1 \pmod{9} \equiv 0 \pmod{9} \quad (7)$$

because $\sum_{i=0}^n a_i \equiv 0 \pmod{9}$ was given.

We can see how **much shorter** and more meaningful that explanation is, how much **more intuitive** is for the familiar with modular arithmetic, how much more abstract it is – and how **much harder** it is for people that are not introduced to that theory and notation.

The first proof is more readable and the second is more aesthetic.

Another example is a problem, often given in job interview – find the sum of the first n integers. A typical solution given by candidates is in the form of loop (in C#):

```
public static int SumOfNDigits(int n)
{
    int sum = 0;
    for (int i = 1; i <= n; i++)
        sum += i;
    return sum;
}
```

This is readable code – we can see that at each iteration the next integer is being added to the sum. The names of the variables carry enough meaning. The level of programming is relatively high – no pointer arithmetic to distract us from the purpose of the method and its mechanism of work, no complex programming paradigms that need translation into the reader’s mind.

There are two types of readability – the “**What**”-readability and the “**How**”-readability. This code answers both question equivocally. This

program is readable, but not very aesthetic and will not be accepted as correct answer to the interview question. The reason is that there is simple, yet elegant formula for calculation of the first n numbers:

$$\sum_{i=1}^n i = \frac{n(n+1)}{2} \quad (8)$$

The code is vastly more elegant, more efficient and shorter:

```
public static int SumOfNDigits(int n)
{
    return n * (n + 1) / 2;
}
```

Aesthetics is tightly connected with efficiency of code, but not with readability.

Statement 1. Maximizing Readability minimizes the effort to understand,

Statement 2. Maximizing aesthetics minimizes the effort to express our *already formed* understanding.

Aesthetics implies knowledge because it is appreciation of simplicity, harmony, elegance and order. Readability implies ease of use and proximity to the ideas in the natural language, while aesthetics requires codification of these ideas in an orderly and meaningful way.

In the next chapter we will define more precisely measures for aesthetics that distinguish it from readability of code, building on top of previous research [Tomov, 2016b], [Pahal and Chillar, 2017], [Buse and Weimer, 2008] and [Posnett et.al., 2011]. The third chapter is devoted to software system readability and aesthetics as long as software system code heterogeneity – a closely related concept. We will propose metric, based on graph model of the software system for readability and entropy based measures of software heterogeneity. Software system aesthetic will be

discussed at the level of code and functionality and we will show the lack of meaning of software system aesthetic on the “code” level.

Aesthetics and readability of code

In previous work [Tomov 2016] we defined intellectual aesthetics as:

Def.3 Intellectual aesthetics:

- *the sense of beauty and awe before certain natural forms and relationships, as expressions of order and truth following from rational explanation; and yet a source of great pleasure and seems to point to values and truths that originates from the human mind;*
- *the whole area of sensory experience that brings us the feelings of beauty and awe; the area of human production that we call the “fine arts and sciences” or the production of aesthetic objects;*
- *A number of themes and issues associated specifically with the human intellectual activity, such as deduction and logical reasoning; various theory proving techniques; and principles of vastness, simplicity, efficiency, harmony, and composition in mathematics, logic and crafts.”*

We also defined aesthetics of functions as **relation between vastness and simplicity** with metrics for vastness - the sum of cardinality of the domain and the co-domain of the function. We didn't introduce metrics for simplicity. The reason for this is that simplicity is the inverse of complexity and the question of defining complexity is self-referential:

Statement 3. *Complexity cannot explain itself.* The measure of complexity is itself complex and subject to measuring, thus creating an infinite regression.

We will use as approximate measure for complexity of code its volume in the simplest possible form – **number of meaningful symbols**

Def. Meaningful symbols – symbols that contain information that is used in code execution

Def. Aesthetics of code is the aesthetics of functionality, embodied in particular programming language.

Thus the metric (9) from [Tomov, 2016a] takes the form (10)

$$M = \frac{V}{c} \quad (9)$$

$$M = \frac{\sum_{i=1}^m |I_i| + \sum_{j=1}^n |O_j|}{N_{ms}} \quad (10)$$

Here $|I_i|$ is the cardinality of the set of values of the i -th input argument and $|O_j|$ is the cardinality of the set of values for the j -th output argument of the function. In the case of the programs for the sum of the first n numbers we have one input argument and one output argument with the same cardinality, which is the number of different values that the System.Int32 type can represent

$$|I_1| = |O_1| = 2^{32} - 1 \quad (11)$$

The number of characters without white spaces for the first version of the program is $N_{ms} = 80$ and for the second version is $N_{ms} = 50$. Clearly the second program has bigger measure of aesthetics because they have the same numerator and in the second case the denominator is smaller

$$M_1 = \frac{2(2^{32}-1)}{80} < M_2 = \frac{2(2^{32}-1)}{50} \quad (12)$$

The maximum vastness of a method is when all inputs and outputs are generic:

$$v = \sum_{i=1}^m |I_i| + \sum_{j=1}^n |O_j| = m \cdot T + n \cdot T = (n + m) \cdot T \quad (14)$$

Here T is the cardinal number either of the natural numbers, or of the reals

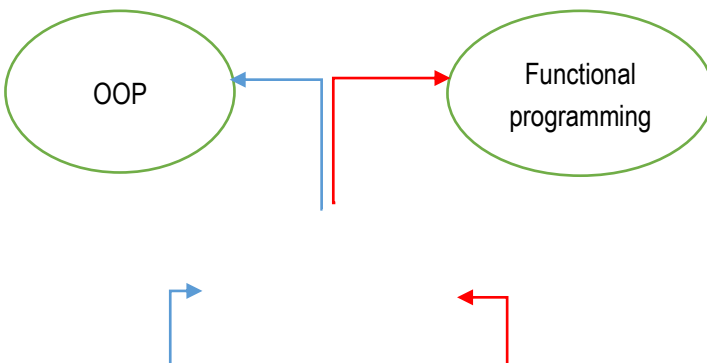
$$T = |N| \cdot \sum_{k=1}^{\infty} |t_k| \geq |N| \quad (15)$$

Here $|t_k|$ is the set of all possible values for the type t_k which in OOP languages could be any user-defined type. It is not clear whether the set of all types resembles the power set of the integers and is therefore uncountable or not.

The minimum number of meaningful symbols N_{ms} is for abstract methods or methods in interfaces for OOP – and it is **constrained** by the requirements for readability of methods which must have meaningful names. By the same approach, **in general** interfaces are more elegant than abstract classes which are more elegant than standard classes – C# as example. This is true for aesthetics of code, but not for **aesthetics of function** since they have no functional implementation – only classes which inherit them will have.

Early evolution of C

The appearance of high-level programming languages is itself an aesthetic revolution in programming. If we draw an imaginary axis with the one side touching the level of binary or hexademicals, a.k.a microcode or bytecode, and on the other side the natural human language and the abstract language of mathematics as branches, the appearance of C is a step before the division of readability and aesthetics (Fig.1)



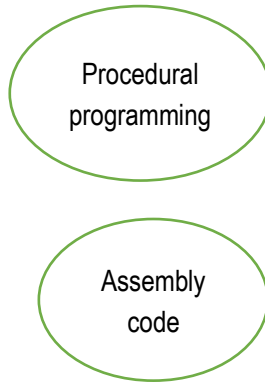


Figure 1. Evolution of programming paradigms in two directions – the **blue** arrows point increase in readability and the **red** – increase in aesthetics. Functional and OOP programming paradigms are different branches from the evolutionary tree

The figure shows the dominant directions of evolution, in every language there are multiple paradigms and every paradigm in certain cases can achieve both increase of readability and aesthetics, since readability is also inversely related to the size of the code. An example of this is the Fibonacci recursion routine:

```
int fibonacci(int i) {  
    if(i == 0) {  
        return 0;  
    }  
    if(i == 1) {  
        return 1;  
    }  
    return fibonacci(i-1) + fibonacci(i-2);  
}
```

Recursion is a primary example of the benefits of high level programming over low level for the human experience. It achieves generality – the domain and co-domain of this function are the sets of *at least* **int16** and not specific numbers. The size of the body of the function does not depend on the size of the number, whereas the actual calculation body grows exponentially with the size of the input in this case because of the two recursive calls to the same function per call or linearly in better written recursion. If this functionality was not captured by recursion it would again need some high level construction to achieve independence of the program length from the size of the input:

```
int fibonacci(int n) {  
  
    int first = 0, second = 1, next, i;  
  
    for (c = 0; c < n; c++)  
    {  
        if (c <= 1)  
            next = c;  
        else  
        {  
            next = first + second;  
            first = second;  
            second = next;  
        }  
    }  
    return next;  
  
}
```

Setting some algorithmic differences aside, (the recursion was not optimally written), the second program is more readable and less aesthetic. The sequence of steps here is explicit and more intuitive for the

novel programmer or the one without extensive mathematical background. The cardinality is the same, the length in meaningful symbols is again independent from the size of the input, but it is larger than in the previous example. Both types of programs are more readable and aesthetics than the low level versions or implementations with ad hoc calculations that are proportional in size to the size of the input – something which loops and recursions avoid.

C++/C# and OOP

Object-oriented programming and functional programming are two different, but related paradigms that focus on eliminating unnecessary code repetition, encapsulation of logic and improving the relationship of the programmer with the code. Both are ideas, taken from pure mathematics. OOP inherits from Bertrand Russell's type theory [Russell, 1908], which is already used in procedural programming, but adds the ability of the developer to create its own types, called **classes** and to create **instances** of these classes, called **objects**. Each class descends from one or more other classes or **class signatures** called **interfaces**, forming a hierarchy and a network like structure of the program code. This allows part of the code to be shifted up along the hierarchy to avoid repetition and thus to decrease the size of the code. That increases both readability and aesthetics, but it is optimizes for the first since the code guidelines and the aim of OOP in C++ and C# is for more literate and verbose code as continuation of Knuth's work [Knuth, 1984]. While the base of OOP is aesthetic, its main purpose is to replace mathematical functions with human readable names of methods. While the abstraction and encapsulation decreases the size of the program, the notation diverges from the most aesthetic notation, which is the shortest one – the mathematical notation. Furthermore, states and data in objects increase

the complexity of the entity and introduces **time** indirectly in the timeless notion of types, which is further away from the optimal aesthetics of mathematical notation. The “**what**” readability increases more than the “**how**” readability since code is just abstracted away, not decreased except with the elimination of code repetition. One of the features that not increase readability and somewhat increase aesthetics is runtime polymorphism as shown in that listing from [Wagner et.al, 2015]:

```
var shapes = new List<Shape>
{
    new Rectangle(),
    new Triangle(),
    new Circle()
};

// Polymorphism at work #2: the virtual method
// Draw is invoked on each of the derived classes, // not the base class.
foreach (var shape in shapes)
{
    shape.Draw();
}
```

Here we have two factors of increased readability – first, we use **foreach** which has syntax closer to natural language than the standard **for** loop, and second, we invoke the same method on different objects in a consistent manner.

The language C# introduced many concepts that increased readability and some that decrease it at the expense of increased aesthetics. The **strong typing** rules fixed the sizes of the basic type thus increasing readability of the code – no need to read special documentation to understand how exactly the language standard is implemented. Here an integer is exactly 32 bits, not minimum of 16. The strict casting rules forced developers to make casting explicit where information may be lost and to call special methods for conversion where it is illegal. For example conversion between the type Boolean and the type Int – mixing the two types in C introduced not only errors, but decreased the logical coherence of the code thus making it less readable. The boolean **true** is converted to 1, but every non-zero integer is converted to **true**. This makes very hard reading code in conditional operators where logical and arithmetic operations are intermixed.

A strong point for readability of C# is the appearance of LINQ that introduced SQL like syntax in the language. An example of that is the answer of the following question in Microsoft developer network – “Can I convert a foreach and if Statement into LINQ?” [Taylor, 2017]:

```
foreach (var code in Globals.myCodes)
{
    if (code.Code == bodyTypeCode)
    {
        bodyType = code.Description;
    }
}
```

```
//Extension method
var bodyType = Globals.myCodes.FirstOrDefault(c => c.Code ==
bodyTypeCode)?.Description;
```

```
//LINQ syntax
var bodyType = (from c in Globals.myCodes
```

```
where c.Code == bodyTypeCode
select c.Description).FirstOrDefault();
```

The pure LINQ syntax is example for increased readability – no need to **assemble the iterations of the code into a logical chain in order to extract the general idea** – it is directly written as a query expression. The extension method is an example for increased aesthetics and decreased readability, thanks to the Lambda Calculus [Cardone and Hindley, 2006]. The invented by Alonso Church system relies on functions, not on types and mixing the two different approaches in programming at will decreases the logical coherence of the code and increases the effort to read it when they are together, which is more often than not the case with LINQ queries. Thus, the late developments of C# for which we will dedicate special series of articles leads the language astray from the path of ever-increasing readability.

Conclusion

The research on software aesthetics will continue with implementation of metrics and in-depth analysis of programming languages and programming paradigms aesthetics, especially the evolution of C, C++ and C#.

Bibliography

-
- [Buse and Weimer, 2008] R.P.L. Buse and W.R.Weimer, A metric for Software Readability, ISSTA '08 Proceedings of the 2008 international symposium on Software testing and analysis pp 121-130, Seattle, WA, USA — July 20 - 24, 2008, DOI: 10.1145/1390630.1390647
- [Cardone and Hindley, 2006] F. Cardone and J.R. Hindley, History of Lambda-calculus and Combinatory Logic, University Mathematics Department Research Report No. MRRS-05-06
- [Gauss, 1801] C.F.Gauss, W.C. Waterhouse, Arthur A. Clarke, J. Brinkhuis, C. Greiter, *Disquisitiones Arithmeticae*, Springer 1986
- [Kernighan and Plauger, 1982] B. W. Kernighan and P. J. Plauger. 1982. *The Elements of Programming Style* (2nd ed.). McGraw-Hill, Inc., New York, NY, USA.
- [Knuth, 1984] D. Knuth, *Literate Programming*, *The Computer Journal*, Volume 27 Issue 2, May 1984, pp 97 - 111
- [Pahal and Chillar, 2017] A. Pahal. R.S.Chillar, *Code Readability: A Review of Metrics for Software Quality*, *International Journal of Computer Trends and Technology (IJCTT)* – Volume 46 Number 1- April 2017
- [Posnett et.al., 2011] D. Posnett, A.Hindle and P. Devanbu, *A Simpler Model of Software Readability*, MSR '11 Proceedings of the 8th Working Conference on Mining Software Repositories pp 73-82, Waikiki, Honolulu, HI, USA — May 21 - 22, 2011, DOI: 10.1145/1985441.1985454
- [Russell, 1908] B. Russell, *Mathematical Logic as Based on the Theory of Types*, *American Journal of Mathematics*, Vol. 30, No. 3 (Jul., 1908), 222-262.

- [Tomov, 2016a] Latchezar P. Tomov, The Role Of Aesthetics In Software Design, Development And Education Part I: Review And Definitions, CSECS 2016, July 04-07 2016, Fulda, Germany
- [Tomov, 2016b] Latchezar P. Tomov, The Role Of Aesthetics In Software Design, Development And Education Part II: Applications of Aesthetics, CSECS 2016, July 04-07 2016, Fulda, Germany
- [Wagner et.al, 2015] B Wagner et.al., Polymorphism (C# Programming Guide), 20 July 2015, <https://docs.microsoft.com/en-us/dotnet/csharp/programming-guide/classes-and-structs/polymorphism> Retrieved at 7/05.2018
- [Taylor, 2017] M. Taylor, Can I convert a foreach and if Statement into LINQ, 31 July, 2017 MSDN <https://social.msdn.microsoft.com/Forums/en-US/6f120b18-fe9b-4c68-92b2-9e334c4d62b6/can-i-convert-a-foreach-and-if-statement-into-linq?forum=csharpgeneral> Retrieved at 7/05.2018

Authors' Information



Latchezar TOMOV, PhD, assistant professor at NBU-Sofia, Bulgaria

Major Fields of Scientific Research: Control theory, software quality, project management, applied mathematics

CSECS 2018, pp. 093 - 108

Computer Science and Education in Computer Science
14th Annual International Conference
ISSN 2603-4794

June 29 – 30, 2018, Boston, USA

REAL-TIME COMPARISON OF MOVEMENT IN BULGARIAN FOLK DANCES

Zlatka UZUNOVA

New Bulgarian University, Department of Informatics

***Abstract:** The most challenging part of creating the Virtual Reality System for Motion Capture Analysis and Visualization for Folk Dance Training is the process of correct comparison of the movements and subsequent assessment. The purposed method uses combined analysis of several body tracking features for movement comparison. The reference dance movements are recorded with two Kinetic sensors, so that they are clean and precise. But the user of the training system, will most likely use one Kinetic sensor, so complex movements like fine spring movements, leg-crossings, body spins, and quick, sharp movements, are difficult to compare correctly because they could not be tracked successfully. The article proposes methods including approximation of movements and an assumption of their proper execution so that the learning process can be carried out.*

Keywords: *Virtual reality system, Motion capture, Machine learning, Bulgarian folk dance, Kinect, Leap Motion, Unity - Game Engine.*

ACM Classification Keywords: *Human computer interaction (HCI), Virtual reality, Motion capture*

Introduction

Dancing is an instrument for non-verbal communication. Every pose, each jump and body movement, is an expression of a feeling. It is hard to catch and compare the feeling, with which the dance is imbuing us.

What we can do about it, is to learn the players to do the dance movements correctly, and the feeling will come as a result, when sound and body movement become one.

For the purposes of this research is used a combination between machine learning and algorithm for prediction of the body movements. To achieve the correct timing, the dance is separated in poses and short gestures, which define each beat. This way, we can learn the system to make correct comparison of the movements and subsequent assessment.

In the last years, there have been many studies made in the field of Gesture recognition. Besides the pure theoretical articles, there are many practical approaches to human-computer interaction. There are also applications, which help restore body mobility, others that guide the movement in the field of robotics, and so on.

For systems, that purpose to detect and record body movements in real time, the Kinect-2 sensor has proven itself as good and stable instrument [Călin, 2016], and because of that, it is the tool of choice for the development of this system.

The combination between Kinect as a Visual Gesture Builder and the game engine Unity, makes the task achievable.

Basic Approach

For the purpose of this article have been examined 2 main approaches to gestures recognition with Kinect.

Heuristic approach

The first one is the conventional method. With this method, there are many programming tasks to be done, before we achieve it. The data obtained from the Kinect sensor have complex relations and should be analyzed individually. One of the biggest challenges is the correct choice of substantial dependencies for each movement. It is necessary to code an algorithm that works for any bone size. It is good to select a specific set of bones to track, and to combine several parameters to properly track the movements. Using logical functions, the bone-to-bone relationships for each individual posture are described. It is important to note that a waiting period should be provided for analysis of the previous frames [Microsoft, 2013].

Heuristic approach is appropriate in cases where match-detection can be performed algorithmically. Its great advantage is that it does not depend on pre-training of the system. It is used to detect discrete static gestures – for example – is the body of a man upright, or lying on the ground [Amini, et. et al., 2016].

Machine learning

This approach has many advantages. The maximum capacity of modern technologies is used. The system learns itself to recognize specific movements through a large amount of previously inputted data. The aim is to recognize movements that have never been part of the practical training of the system. Nowadays, the systems being trained, serve many different purposes. The complexity of the tasks is constantly growing and the importance

of machine learning algorithms is undeniable [Amini, et. et al., 2016].

Used Algorithms

Several types of algorithms are used in the machine learning practice, depending on the type of input learning material and the purpose of the system.

One of the main algorithms used for gesture recognition is "Dynamic Time Warping" (DTW). It is used in any data that can be transformed to a linear chart. It is an algorithm that compares sequences that can be transposed into variations of sequences of velocity and time [Chuan-Jun Su, et. al., 2014]. In static systems with a fixed gesture classification standard, DTW is very accurate, but for dynamic systems that are trained and used in real time, it is preferable to use Hidden Markov Models (HMM) [Călin, 2016]. Hybrid systems combining different approaches, such as Hidden Markov Models and Neural Networks, are used as well [Wu, et. al., 2017].

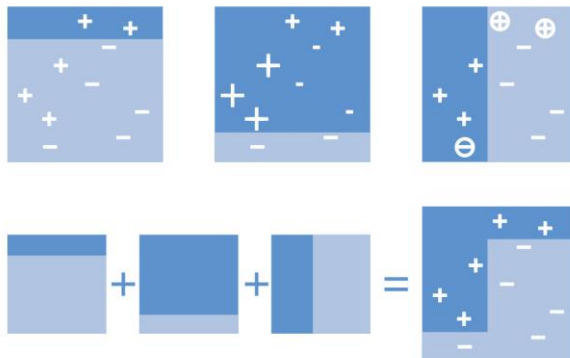


Figure 1. Linear weighted combination of three different weak classifiers compiled the strong nonlinear classifier [Ripoll, 2016]

Kinect v2 has an official tool that conveniently integrates gestures into applications. This is the Visual Gesture Builder (VGB). It contains numerous machine learning techniques that support the training and recognition of gestures, and their subsequent implementation in a suitable environment [Rahman, 2017]

For the purposes of this study are being used algorithms, integrated in Visual Gesture Builder (VGB) - Adaptive Boosting (AdaBoost) and Random Forest Regression (RFR).

VGB uses AdaBoost machine learning algorithm through a discrete indicator AdaBoostTrigger. A powerful tool that heavily depends on the number of samples, and uses binary values to distinguish the different types of predefined discrete gestures [Jin, et. al., 2015].

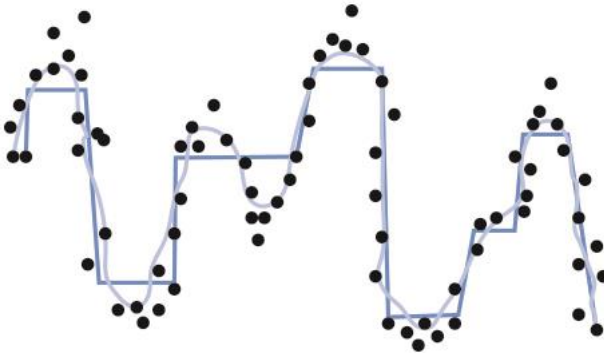


Figure 2. Training samples and the Boosted Decision Tree Regression

VGB uses RFR machine learning algorithm through a continuous indicator Random Forest Regression Progress (RFRProgress). Via detection technology it is estimated the stage of gesture execution. For the correct estimation of the progress in a

continuous gesture, it is good to note in advance its building discrete gestures.

Proper and accurate tagging of the input data plays an important role in reducing the latency and increasing the accuracy of each of the algorithms used.

Method

The method consists of several crucial stages.

Dividing the dance into beats and tacts. Each beat is defined as a discrete gesture and each tact as progress gesture defined by the beat that contains it.

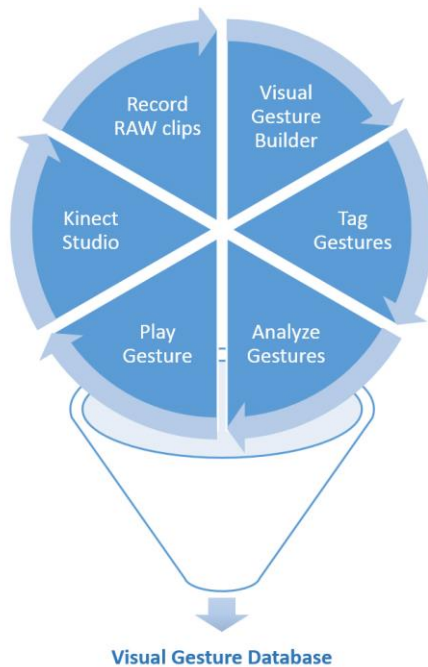


Figure 3. Cycle of creating Visual Gesture Database

A system is trained to recognize the corresponding gestures and predefined poses. The gestures are performed many times by different people with different proportions. Each gesture is played at a different angle to the sensor's camera. The gestures are recorded at a different speed than the normal. It is preferable for the gestures to be performed very slowly, the speed after that won't be affected, but the capture is better.

The demonstrated movement during training is done with captured movement by 2 Kinect sensors, followed by a feet stabilization post-processing. Unwanted vibrations and movements are cleared out. The reproduced final movement is accurate and precise.

On the other hand, the captured movement of the user during the training is by done only by one sensor. That's why the trained system is designed to recognize movement from one sensor. Complex gestures are recognized with high precision via VGB [Microsoft, 2013]. Nevertheless, at some point, the system might not to recognize correctly performed movements due to the specifics of Bulgarian folk dances. [Hu MC, et. al., 2015] use a combined approach to assess the posture match during the learning process by extracting and combining data from information, obtained from the bone system and from the user's silhouette. To reduce the chance that a gesture will not be correctly recognized, this method uses a combination between two approaches.

The assessment of the match is done not only by the trained system through visual gestures, but at times it is combined with gesture data from the Kinect sensor. This is particularly necessary for gestures that can overlap key bones. A condition is set to check if there is a chance that a specific gesture has been done, by

estimating the position or rotation of a particular body joint. If the predefined movement is captured correctly and the final position is reached, and the offset is with the correct direction and angle, it can be assumed that the movement in the middle was correctly performed whether the trained system recognizes the gesture, or not. For this purpose, a few simple calculations on basic bones are made. All bones of the body are tracked simultaneously by the Kinect SDK. In two arrays, all bone positions and their tracking states are stored. It should be noted that all joint positions are stored in world coordinates and the measurement unit is meters. We do not need the whole model of the bone system, and we don't need to calculate the coordinates for each joint, but only the key elements for the tact. For the different tacts, the positions of the individual joints are tracked; the change of the angle between them at different times is tracked as well. In the vast majority of cases, we use that each movement can be defined as a directional vector. It is necessary to determine where the bone is at the beginning and what should be its minimum displacement in a certain direction.

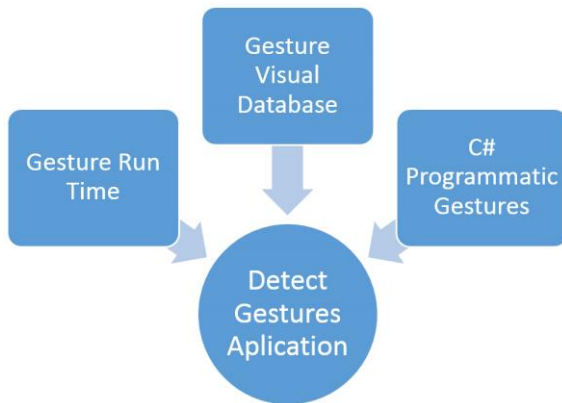


Figure 4. Cycle of creating Visual Gesture Database

The correct calibration of the Kinect, used by the user, is essential. There are two basic settings that must be set before the training starts: the distance from the ground to the sensor in meters, and the rotation angle between the Kinect and the ground, in degrees. Those can be set manually, or automatically. The next step for proper training is the ability to detect whether the user is captured, or is out of the range of the cameras. And the third essential step is to determine the times in which the player has turned his back on the sensor. Various approaches can be used for this. The most intuitive solution is to add a face-recognition system. This way, it will always be known whether the user is facing the sensor, or has turned his back to it, when initializing the application [Filkov, 2015].

With this methodology, movements are compared in several stages. During each stage, is done an appropriate assessment method that provides Immediate Feedback, Score Report, and Slow Motion Replay to the player [Uzunova et. al., 2016]. Initially, we compare only the vertical displacement of the Hip bone. This determines whether the character moves rhythmically with the music. At this stage, only one parameter is compared, so the comparison is done programmatically. The next stage is recognizing the specific beats. The goal is for the player to grasp the individual key elements of each tact – discrete gestures are used for this purpose, paying attention only to the position of major bones. Next is recognition tact by tact – for the purpose are used progress gestures, which are composed of pre-learned discrete gestures. Lastly, one whole dance unit is tracked, using a combined assessment between visual gestures and programmatic gestures.

For the quality reproduction of the captured motion, because the motion is captured by only one sensor, it is good have a variant,

in which the user will control only certain bones of his avatar. The remaining bones that are non-essential and don't participate in the gesture recognition, could be pre-animated and played back at the appropriate time. This way, by combining real user gestures and pre-recorded dance elements, the learner's avatar will be sufficiently plausible to visualize, while the learning simulation will continue more smoothly. For this purpose, it is best to mix Kinect-captured movement with Mecanim animation, by an avatar mask. The mask is good to exclude the bones important to the sensor capture, which are tied to the gesture identification database, but without root-joint.

Experimental work

During the experimental work, attempts were made to train the system with different parts of Bulgarian folk dances with various complexity. The obtained data show, that the concept of the method is doable.

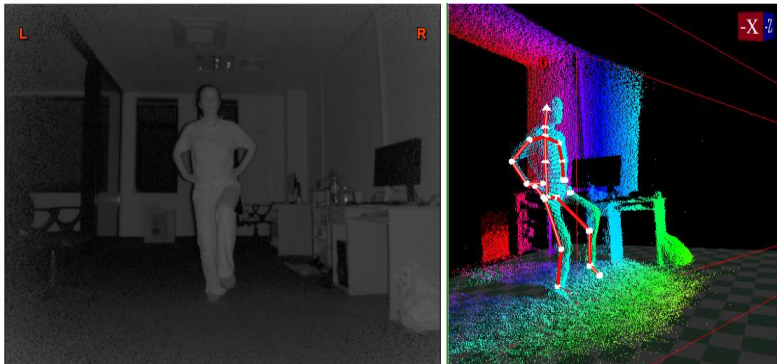


Figure 5. Kinect Studio Clips in Visual Gesture Builder
2D View (left) 3D View (right)

The focus of this study is on one of the difficult Bulgarian folk dances - Graovsko Horo.

| | Name | Value | Type |
|---|-------------------------|-------|-------|
| * | GraovskoHoroT01Progress | 1 | FLOAT |
| | GraovskoHoroT01B2 | True | BOOL |
| | GraovskoHoroT01B0 | | BOOL |
| | GraovskoHoroT01B1 | | BOOL |

Figure 6. GraovskoHoroT01Progress is a continuous indicator includes three discrete gestures “GraovskoHoroT01B0”, “GraovskoHoroT01B1” and “GraovskoHoroT01B2”. Value “1” indicates that the gesture is 100% done. Value “True” indicates that the person is in the process of performing the discrete gesture “GraovskoHoroT01B2”.

It is divided into 10 tacts, each of which is 3 beats. For every beat, there is one discrete gesture. Every 3 beats make up one progress gesture. They were captured using the Kinect v2 sensor, and the database is made using the VGB tools.

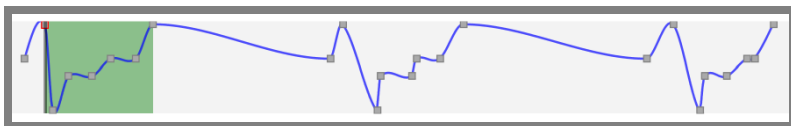


Figure 7. Process of training

Dynamic gestures weights vary between 0 and 1. For each beat they are as follows: B0 (0 - 0.4) B1 (0.4-0.6) B2 (0.6 - 1).

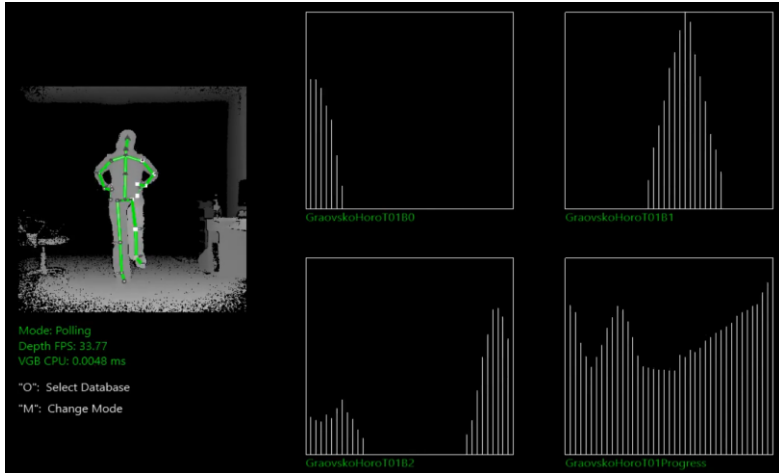


Figure 8. The Visual Gesture Builder Viewer, preview of how database classifies gestures

The system is trained and the resulting database is used in Unity application. Scripts to track the movement of certain bones for each beat were added.

```

Top 10 contributing weak classifiers:
VelocityY( FootRight ) using inferred joints, fValue < -2.000003, alpha = 2.568879
Angles( SpineMid, Head, FootLeft ) using inferred joints, fValue < 4.000000, alpha = 2.361723
DiffMuscleForceZ( HipRight, SpineMid ) using inferred joints, fValue < -0.600000, alpha = 1.870100
Angles( AnkleRight, KneeRight, HipRight ) using inferred joints, fValue < 128.000000, alpha = 1.814810
Angles( AnkleLeft, KneeLeft, HipLeft ) using inferred joints, fValue < 158.000000, alpha = 1.811998
DiffPositionZ( HipLeft, Neck ) using inferred joints, fValue < 0.000000, alpha = 1.714548
DiffMuscleForceX( KneeLeft, SpineShoulder ) using inferred joints, fValue >= 0.200000, alpha = 1.682334
MuscleForceY( KneeRight ) using inferred joints, fValue < -0.000003, alpha = 1.639042
Acceleration( FootRight ) rejecting inferred joints, fValue >= 1.100000, alpha = 1.620712
MusclePower( AnkleRight ) using inferred joints, fValue < -3.000000, alpha = 1.615147

```

Figure 9. Part of a log file from the project, that shows for a specific gesture the top 10 weak classifiers, which can be used to implement gesture detection.

The main problems arise in the last stage of evaluating the matches, namely when a whole pattern is played and the body rotates to move in the direction of flow of the horo. A solution for the moment is the player to perform the dance without following

the horo's flow during this movement, but only to learn the necessary steps.

Conclusion

This article proposes a method for correctly comparing the movements in building a Bulgarian dances training system.

Two benchmarking approaches are combined - a pre-trained system with its database and algorithmic scoring of matches. The experiments are in the initial phase but prove that the method can be successfully implemented.

In the future it will be created a stable training system in which it will be available to track and analyze the movement of several people at the same time.

Bibliography

- [Ripoll, 2016] Marina Ballester Ripoll, Gesture recognition using a depth sensor and machine learning techniques, <http://hdl.handle.net/10251/77950>, DÜSSELDORF, 2016
- [Rahman, 2017] Mansib Rahman, Understanding How the Kinect Works, Apress, Berkeley, CA, 13 August 2017
- [Jin, et. al., 2015] X. Jin, Y. Yao, Q. Jiang, X. Huang, J. Zhang, X. Zhang, K. Zhang, "Virtual personal trainer via the kinect sensor", Communication Technology (ICCT) 2015 IEEE 16th International Conference, pp. 460-463, 2015.

- [Microsoft, 2013] Microsoft, Visual Gesture Builder: A Data-Driven Solution to Gesture Detection, Xbox One XDK, September 18, 2013
- [Filkov, 2015] Rumen Filkov, Kinect v2 Tips, Tricks and Examples, January 25, 2015
- [Hu MC, et. al., 2015] Hu MC, Chen CW, Cheng WH, Chang CH, Lai JH, Wu JL., Real-Time Human Movement Retrieval and Assessment With Kinect Sensor, IEEE Trans Cybern, 2015
- [Chuan-Jun Su, et. al., 2014] Chuan-Jun Su, Chang-Yu Chiang, Jing-Yan Huang, Kinect-enabled home-based rehabilitation system using Dynamic Time Warping and fuzzy logic, Applied Soft Computing, Volume 22, Pages 652-666, September 2014
- [Călin, 2016] A. D. Călin, "Gesture recognition on kinect time series data using dynamic time warping and hidden markov models", 2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), pp. 264-271, Sept 2016.
- [Wu, et. al., 2017] Huiyue Wu, Jianmin Wang, Xiaolong Zhang, Combining hidden Markov model and fuzzy neural network for continuous recognition of complex dynamic gestures, The Visual Computer, October 2017
- [Uzunova et. al., 2016] Uzunova Z., Chotrov D., Maleshkov S., Virtual reality system for motion capture analysis and visualization for folk dance training, Computer Science Education & Computer Science Research Journal: ISSN 1313-8624, CSECS 2016

Authors' Information



Zlatka UZUNOVA, has received BSc. (2008) degree in computer systems and technologies and MSc. (2015) degree in multimedia and computer animation from the New Bulgarian University of Sofia. Currently she is PhD student at New Bulgarian University, Department of Informatics (zl.uzunova@gmail.com).

Major Fields of Scientific Research: Computer graphics, Virtual reality

SIMULATION AND EVALUATION OF SCENARIOS IN A GAS STATION USING SIMUL8 SOFTWARE

Denis Ramos de Oliveira, João Pedro Fonseca de Barcelos, Thiago Henrique Nogueira

Universidade Federal de Viçosa, Institute of Exact and Technological Sciences

Abstract: *The simulation consists of reproducing the functioning of a system by creating a model, considering its statistical data and its premises, to understand the behavior of the variables of interest. Within the simulation, several software are used. One of the most popular in organizational environments and in universities is the Simul8. This software uses two-dimensional animations, through representation in entities, resources, activities, entry points, output points, and queues. Its purpose is to simulate the operation of complex discrete models, and by manipulating the variables in the model, it is possible to simulate different scenarios and find effective solutions. As a complementary tool to the simulation of scenarios, it is relevant to cite the financial analysis, which through indicators such as EAC, MAR, NPV, PB, can indicate the economic viability of an investment. In this context, the objective of this paper was to develop a scenario analysis, through simulation in the software Simul8, in a gas station, which has an area of 150 m² and offers the services of filling up gasoline, washing the car, and calibrating the tires, to its customers. In another situation, the purchase of an adjacent land of 80 m² was analyzed. Therefore, it was possible to determine the number of pumps and employees that generates the highest profit, considering the available space of the two situations, and to develop a financial analysis to demonstrate the viability of the investment. It was demonstrated that the number of pumps and employees that generated the highest monthly profit was 4 and 6, respectively, for the first situation, with 150 m², and 6 and 5 for the second, with 230 m². It was verified that investment in the acquisition of the land is feasible because it generated both an NPV and EAC positives. However, this investment is not suggested, since the indicators obtained for the initial situation, with 150 m², are higher, since no investments will be made in this situation and the profit values for the two scenarios are similar. Values of investment in pumps and other costs, as well as expenses, are not part of the scope of this paper, which may explain the high monetary values found based on the assumptions considered.*

Keywords: *Simulation; Simul8; Gas station; Scenarios.*

MAIN HEADING

1. Introduction

Simulation is a technique that seeks to imitate, or simulate, the functioning of a system. In order to understand a system, some assumptions about the various activities and resources that compose it are

necessary, which are usually described through mathematical or logical distributions, thus forming a model that is used to understand the way the corresponding system works and the behavior of its variables. In the simulation, a computer is used in order to evaluate a model numerically, and data are collected to estimate the desired real characteristics of the model (LAW; KELTON; KELTON, 1991).

The main reason for using the simulation is the impossibility of using conventional analytical methods, such as algebra, calculus, probability theory, etc., in the analysis of complex problems. Unlike Operational Research, the simulation does not provide an optimal value for the variables of interest, but it can simulate different scenarios in order to find a solution that returns a desired value for these variables in highly complex problems. Many problems in the areas of production, logistics and services can be analyzed through simulation, given their dynamic and probabilistic nature. For this, all system activities, their flow, duration and resources must be defined a priori (BANKS, 1998; CONCANNON et al., 2007; O'KANE et al., 2000).

Some software are used as simulation tools. Simul8 is one of the most popular in the most diverse organizations and universities in the world. It is used to simulate discrete events, allowing the user to create a visual model of the analyzed system through its representation through entities, queues, activities, resources, points of entry and exit in the system. Once this model is created in the software, the simulation can be started, using random numbers generated by the model, considering the time distributions of the model (SHALLIKER; RICKETTS, 2002).

Thus, Simul8 can be used to simulate the operation of the most varied types of system, from complex models of production systems to service models, in a clear and succinct way, through 2-dimensional animations that demonstrate the flow of entities through various activities in one model. Its results can be harnessed in order to find a solution that improves the performance of the system and the variables of interest (CONCANNON et al., 2007; FICOVÁ; KUNCOVÁ, 2013).

Ferreira (2001) argues that in the current scenario of high market competition, in addition to understanding the functioning of production processes, it is necessary for an organization to carry out a financial analysis that demonstrates the economic viability of a project. Thus, this type of analysis can work as a complement to the scenario analysis performed in a simulation.

To make a financial analysis, can be used parameters such as the Minimum Attraction Rate (MAR), which is the minimum rate that an investor expects to receive for making an investment. In addition, can be used the Equivalent Annual Cost (EAC), that consists of defining an equivalent series equivalent to the annual cash flow. The Net Present Value (NPV), which is the sum of the values that the investment will profit or pay, in an initial time, and the Payback (PB), that is the time necessary to recover the resources invested in a project can be used too (FERREIRA, CAVALCANTE, 2001; KOPITTKE; CASAROTTO FILHO, 2000; BRAGA, 1998).

Considering this context of importance of the simulation and the use of software for the simulation of scenarios, we developed in this work a scenario analysis, through Simul8, in a gas station, which receives customers through three different traffic routes, with different arrival rates, and provides the services of filling up, washing and calibration.

The objective, therefore, of this paper was to evaluate different scenarios for the company under study, through simulation, considering as variables of the problem the number of employees, gas pumps, calibration stations and available space in the position. Initially, a space of 150 m² was taken, which is the area that the station currently has. Then, an alternative scenario was analyzed, with the possible acquisition of an adjacent land of 80 m². In the latter case, in addition to the simulation, a financial analysis was carried out to determine the feasibility of this investment. Therefore, we tried to determine, through simulation in Simul8 software, the configuration of variables that provides the greatest profit for the company.

2. Theoretical Framework

2.1 The application of Simulation and Simul8 Software in the evaluation of scenarios

Simulation has become an essential tool in many technical and scientific areas today. There are records of its presence for centuries. The military area has used it constantly in order to simulate the development of weapons and strategies. The modern aviation industry has developed software capable of simulating high-precision flights. Space programs have made use of this technique for training employees and evaluating scenarios. Nuclear programs, which have already demonstrated just how devastating the consequences of a failure in their systems can be, has also made extensive use of simulation (BRADLEY, 2006).

In addition, the simulation has been used frequently by several organizations in order to achieve better results through the simulation of scenarios. Montevecchi et al. (2007) demonstrates the importance of simulation to represent different scenarios and strategies of a company. Other simulation applications cited by Law; Kelton and Kelton (1991) are:

- Development and analysis of manufacturing systems;
- Evaluation of hardware and software for computer systems;
- Evaluation of new military tactics and tactics;
- Determination of stock resupply policies;
- Development of communication systems;
- Development of transport systems, such as ports, highways, airports, subway;
- Development of evaluation systems for hospitals, restaurants, and services in general.
- Financial or economic analysis of systems.

Several software have been used for simulation. Among those currently used, one of the greatest significance is the Simul8. This was developed by the US company SIMUL8 Corporation and began to be used between mid-1994 and 1995 in the United States, and every subsequent year, has been re-released with innovations and improvements in its functionalities. It is considered as the preferred simulation tool of notable companies, such as McDonald's, Ford, Hewlett Packard, as well as a variety of small companies and universities from various countries (CONCANNON, HUNTER, TREMBLE, 2003; GREASLEY, 2003).

Its popularity can be justified by its ease of use as a discrete event simulation tool that provides the resolution of a multitude of complex problems. A discrete system is one whose state of the variables changes instantly at discrete times, and a continuous system is one that changes the state of variables continuously over time. Because it is an object-oriented modeling tool, Simul8 incorporates a programming and modeling language that allows the user to develop accurate, flexible and robust simulations faster (CONCANNON; HUNTER; TREMBLE, 2003). The Simul8, according to Bangsow (2010); Concannon et al. (2007), by allowing the user to create a visual model of the system analyzed through two-dimensional animations, representing the objects directly on the computer screen, it is used 6 basic representations, namely:

- Work Entry Point: These are points that represent the entry of similar entities in the model. Entities that appear in the same work entry point can differentiate themselves through specific attributes.
- Work Item: are the entities that circulate in the model, whose behavior is desired to analyze.
- Storage bin: are the queues of the model, where entities expect to be met.

- Work Center: are the points where the activities on the entities occur.
- Work Exit Point: are the exit points of the entities.
- Resource: are the resources needed to perform the activities on the entities in the Work Centers. They can be shared in various activities, dependent on schedules, have certain levels of efficiencies, etc.

These representations that are schematized in simul8 can be visualized by means of figure 1, as follows:

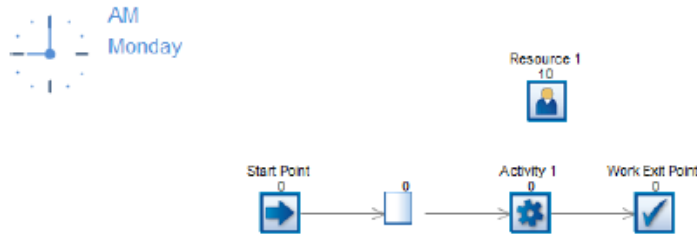


Figure 1. Representations in Simul8

Source: Elaborated by the authors (2017).

Note that all objects, except the resources, are connected by arrows that demonstrate the sequence of movements that entities will follow in the model, from its appearance in the Work Entry Point to its exit in the Work Exit Point. After modeling the system in simul8, the simulation can be performed. The animation in the software demonstrates the flow of entities, and after completion of the simulation, the variables of interest, such as profit, queue time, resource utilization, etc., can be easily accessed and analyzed. Several simulation rounds can be performed under different circumstances, and system and variable performance can be analyzed statistically (SHALLIKER; RICKETTS, 2002).

3. Methodology

The study of this work was developed in a gas station that has 150 m² of area to be used to supply its three services: filling up, washing and calibration. Thus, it is considered that in relation to the technical procedures used, this work has used the technique of a study case. According to the propositions of Brasileiro (2013), a study case is the analysis of a specific case, and is generally used in descriptive research, with the purpose of obtaining a deep and detailed understanding of a given situation, which in this paper will be the operation of the gas station.

This technique was used in this work mainly using a quantitative approach, which according to Almeida (2014) is one that basically manipulates measurable and numerical information, such as probabilities, statistics, deviations, etc. The following data were provided for the simulation of the gas station operation:

- Cars can pass in front of the station from 3 different flows A, B and D: A, with arrivals every 10 seconds, B, which has arrivals every 30, and C, with arrivals every 12 seconds, being 15% of the cars coming from this last flow return in D.
- 10% of the cars that pass in front of the station enter to do some service.
- 70% of the cars entering the station will only fill up, 15% fill up and wash the car, 5% fill up, wash and calibrate the tires, 5% only wash, 5% only calibrate.
- The time required to fill up a car is described by an Exponential of 3 minutes, in which each liter is supplied in 5 seconds; Wash: Normal (13, 3) min; Calibrate: Triangular (2,3,4) min.

- Each car occupies an average space of 6 m² and each fuel pump a space of 8 m².
- The gross profit (sales price - cost of the product) generated by each customer in the filling up process is equal to 30 cents per liter of fuel supplied.
- The gross profit from the car wash, which is made by an employee, is R\$ 8 per wash, and the calibration process is free.

Despite the quantitative approach of this work, a qualitative analysis of the data for the simulation was made from the following information provided:

- Once inside the station, the cars can: fill up, calibrate the tires, or do the washing service.
- The filling up process can be performed at any of the fuel pumps and the access queue to these pumps is a single row of cars for each pump.
- When accessing one of the pumps the car must wait for an employee to carry out the programming of the machine.
- After scheduling, the employee is released and the customer (car) must wait for the automatic fill up, activity whose duration depends on the volume supplied.
- The pump for the gas is automatic and after the filling up, the customer must wait for the employee to finish the operation (remove the hose from the car and collect the charge).
- The process of calibrating the tires is carried out by the customer in a reserved place, not requiring employees.
- The washing process requires an employee to carry out the entire process.
- The monthly cost of an employee is R\$ 1700.00.

Therefore, from this information provided, Simul8 software was used for the development of the simulation to determine the number of pumps and employees that returns the highest profit.

After the simulation was performed considering an available space of 150 m², another scenario was simulated through the purchase of an adjacent land of 80 m², which has a value of R\$ 130,000. Considering this last scenario, a financial analysis was carried out, with a given MRA of 15% per year to determine the viability of the investment.

4. Results and Discussions

Based on the statistical data provided and the complementary information, the simulation process considered a period of execution of 1 month, considering that the cost of the employees is monthly. The warm up time, in which the software does not consider the runs in order to reduce the variability of the initial results, was 11,009 seconds, which was the simulated period in which 23 cars left the station after being served. This number was considered the maximum number of cars the station can receive at the same time, considering only one pump in the gas station.

To determine the number of simulation rounds to be performed, a 5% change in the number of entities that exited the system in one month was considered. Therefore, it was calculated that the number of replications, that is, of simulations using different random numbers, would be equal to four. It was decided to run the 12-hour daily (from 08:00 a.m. to 8:00 p.m.), 7 days a week, without shifts, since no variation was considered in the number of employees. In addition, the following assumptions were considered:

- The points of entry direct the cars according to the probabilities already explained, according to the percentage discipline of the routing out of each of these points.

- The first car, when entering the station, makes the space available to be subtracted by the space corresponding to the maximum number of pumps installed. In this same activity (set space), which has fixed time equal to 0, the cars are directed to have their service assigned according to the probabilities defined, according to the percentage discipline of their routing out.
- In the following activities (set filling up, set filling up and washing, set filling up, washing and calibration, set washing, calibration), that has fixed time equal to 0, the entities receive the attributes of the services that will perform in the gas station (to fill up, to wash and to calibrate).
- In the next activity, which is the sorting, with fixed time 0, there is the analysis of the attributes of the entity and from these, their targeting according to the service that the same should receive. Its direction is by means of a label.
- The entity also reduces by 6 m² the value of the variable that represents the space available at the post when entering space, as this is the average space occupied by an entity at the station. The variable that corresponds to the space available increases by 6 m² when the car leaves the gas station.
- After sorting, the entity can follow 4 paths, depending on the service attributes previously received, choosing the lowest queue:
- It goes to the service of filling up, passing first by the programming of the pump, activity of fixed time equal to 0 and that needs an employee resource, normally allocated, and needs a pump, that gets stuck in the activity. Then it goes into the filling up activity, which has a time described by an exponential distribution of 3 minutes, and finally there is the removing of the pump and charging, with fixed time 0, which needs an employee and releases the resource pump.
- Go to the washing service, which is performed by an employee and described by a normal distribution of 13 and 3 minutes. It was not considered a specific space at the station for this service. The customer, when needing the washing of his car, will be attended by an employee who is available, who will wash the car. Thus, the number of simultaneous activities that can be performed in this service will be limited by the number of employees available, since this is the only associated resource, considering that the materials needed for car washing are unlimited.
- Go to calibration, which requires no resource and is described by a triangular distribution of 2, 3 and 4 minutes.
- Go to the street, because the costumer may not find available space at the station.
- After each of these three services, there is a further sorting to determine if the entity needs another service. If it does not require it, it will be directed to the "profit calculation" activity, which has fixed time equal to 0, and where there is the calculation of the profit variable based on the variables that describe the time of filling up, the number of cars washed, and the cost of employees. After this activity, the entity (the car) goes to Exit Point.
- As the filling up activities can be carried out by a maximum of 18 pumps of 8 m², considering the space of 150 m², the number of activities that can be performed at the same time in these activities (replicate) will be greater than 18.
- Considering also that the number of employees may be greater than that of pumps, will be assigned a number of replications greater than 18 for the washing activity, since the only resource necessary for the car wash will be the employee.
- Because the calibration service does not generate profit, and it is desired to find the number of pumps and employees that maximize profit, it was considered only a calibration activity (replicate = 1), that is, only one calibrator at the gas station. Calibration was also considered as a service that is not essential to the customer. Thus, it is expected that if there is a queue larger than one entity in the space reserved for calibration, the car (entity) will move to another more important service, leaving the

calibration as the last service. If he still finds a queue in this service, he will go to the street, which also happens if this is the only service needed and there is a queue at this activity.

The model of the station, considering these premises, can be visualized by means of the figure 2 below, elaborated using the software Simul8.

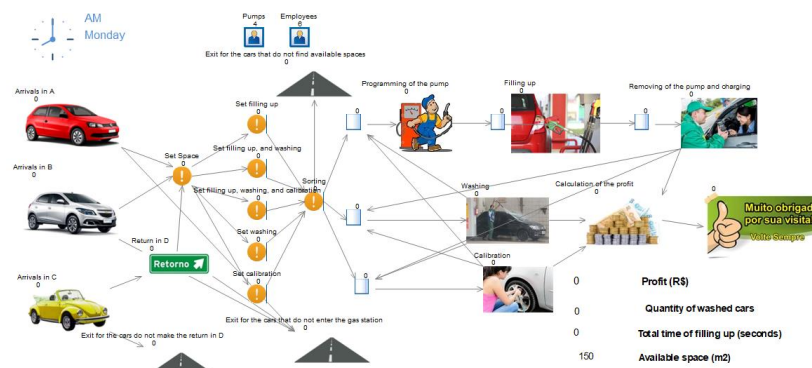


Figure 2. Model of the gas station in Simul8

Source: Elaborated by the authors (2017).

Considering these different assumptions, the simulation was performed for different scenarios, varying the number of pumps and employees at the station, in order to determine the configuration of these variables that generates the highest profit. A diagram, drawn up using Microsoft Excel software, of the monthly profit generated, considering the different scenarios, varying the number of pumps and employees, can be visualized by means of figure 3 below:

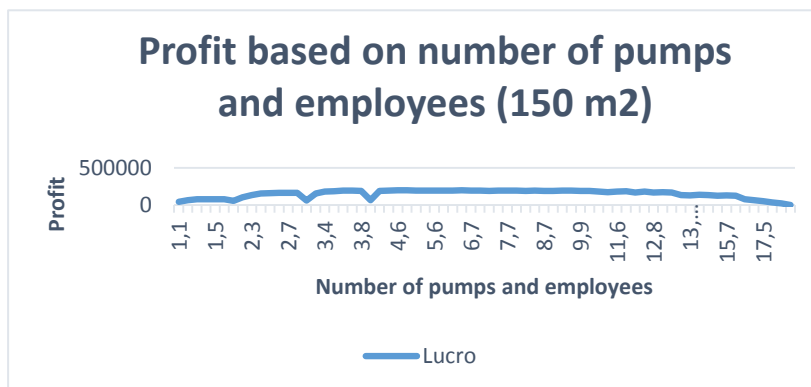


Figure 3. Profit in function of the number of pumps and employees, considering 150 m².

Source: Elaborated by the authors (2017).

Based on this chart, the profit figures generated were classified in descending order, in order to facilitate the visualization. The highest profits, according to the configurations of the variables, can be visualized by means of the following figure 4, elaborated using the software Microsoft Excel:

| Number of pumps/Number of Employees | Profit (R\$) |
|-------------------------------------|--------------|
| 4,6 | 198,034.53 |
| 6,6 | 197,427.39 |
| 4,7 | 197,316.90 |
| 4,5 | 197,001.55 |
| 7,6 | 196,484.69 |
| 5,7 | 196,193.46 |
| 6,7 | 196,064.26 |
| 3,6 | 195,603.76 |
| 5,6 | 195,563.70 |
| 4,8 | 195,491.07 |
| 7,7 | 194,631.16 |
| 6,8 | 194,209.62 |
| 5,5 | 193,935.11 |
| 9,8 | 193,455.95 |
| 6,5 | 193,411.12 |

Figure 4.- Decreasing profit based on the number of pumps and employees, considering 150 m².

Source: Elaborated by the authors (2017).

Then, by simulating the various scenarios, considering a space of 150 m², it can be noticed that the configuration that generates the highest monthly profit is 4 pumps and 6 employees, returning a monthly profit of around R\$ 198,034.53, considering the previously mentioned premises, such as the operation for 7 days in the week with 12 hours daily. The representation in Simul8 of this scenario can be visualized by means of figure 5 below:

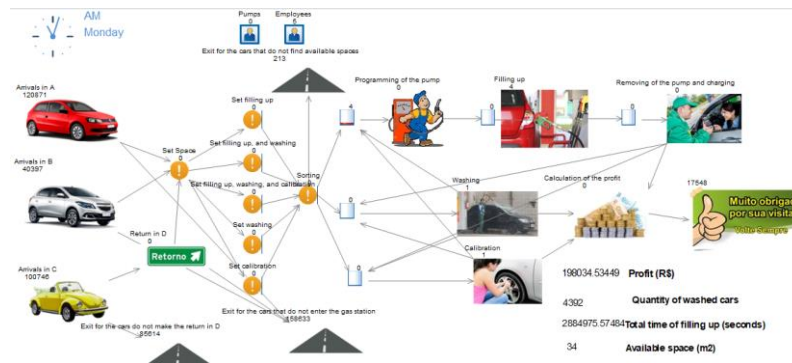


Figure 5. Simulation with 4 pumps and 6 employees

Source: Elaborated by the authors (2017).

In addition, the simulation was done considering the acquisition of a land of 80 m². The results obtained, considering the same assumptions, initially changing only the variable that corresponds to the space available at the station, can be visualized by means of the following figure, elaborated using Microsoft Excel software:

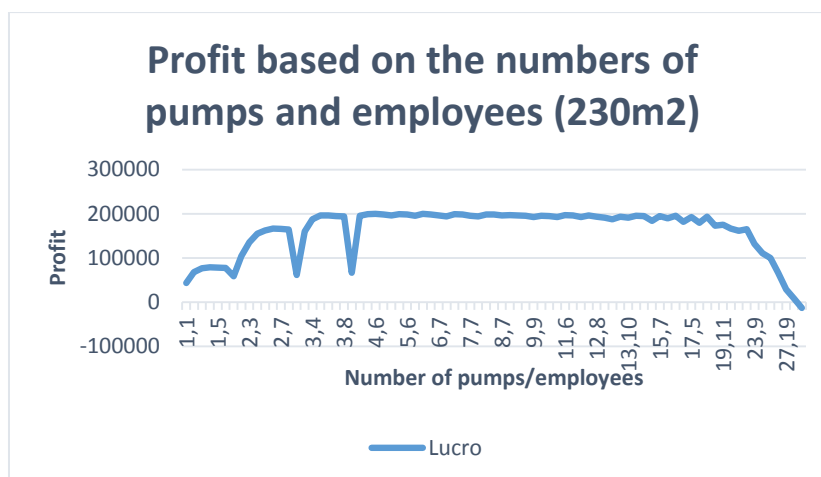


Figure 6. Profit due to the number of pumps and employees, considering 230 m².

Source: Elaborated by the authors (2017).

Similarly, these results were classified in a decreasing way, to facilitate the visualization, which can be observed by means of the figure 7, elaborated using the software Microsoft Excel:

| Number of pumps/Number of employees | Profit (R\$) |
|-------------------------------------|--------------|
| 6,5 | 199,975.17 |
| 4,6 | 199,709.94 |
| 7,5 | 199,530.95 |
| 5,5 | 199,413.31 |
| 4,5 | 199,161.21 |
| 4,7 | 198,738.67 |
| 8,6 | 198,690.08 |
| 6,6 | 198,639.89 |
| 8,5 | 198,388.76 |
| 7,6 | 198,222.11 |
| 5,6 | 198,211.36 |
| 9,6 | 197,444.22 |
| 11,6 | 196,791.43 |
| 9,7 | 196,702.90 |
| 4,8 | 196,699.26 |

Figure 7. Decreasing profit based on the number of pumps and employees, considering 230 m².

Source: Elaborated by the authors (2017).

Note that, considering a space of 230 m², the configuration that generates the highest monthly profit is 6 pumps and 5 employees. The representation in Simul8 of this scenario can be visualized through figure 8 below:

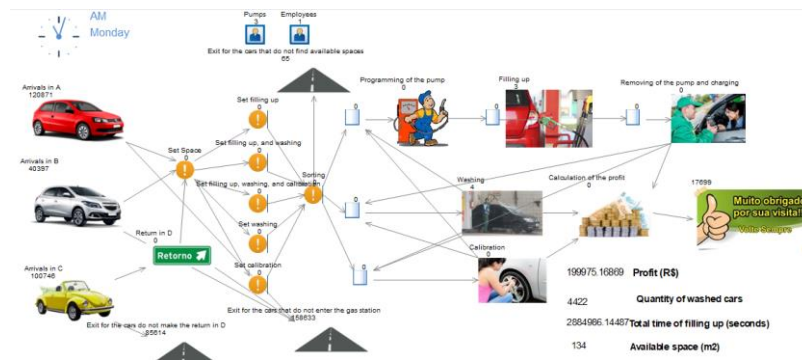


Figure 8. Simulation with 6 pumps and 5 employees

Source: Prepared by the authors (2017).

From the scenario that returns the highest profit for the situation with a space available of 230 m², a financial analysis was made, considering as parameters the value of the Minimum Attractiveness Rate of 15% per year, in order to determine the feasibility of this investment in a time horizon of 1 year. This period will be the time taken to pay the investment to be made in the acquisition of adjacent land under that rate.

For the development of the financial analysis, the Net Present Value and the Equivalent Annual Cost were calculated by the following equation, where P is the principal amount invested in the initial time and A the value of the installments to be received:

$$P / A: \frac{A((1+i)^n - 1)}{i((1+i)^n)}$$

The steps followed were:

- Convert the given MAR of 15% a year to a monthly rate, considering compound interest and using the following mathematical expression that gives the equivalence of the two rates:

- $(1 + ia) = (1 + ip)^n$,

- where ia is the annual rate equivalent (15%), ip is the rate of the given period, and n is the number of periods (12).

Thus, the MAR generated was 1.17% per month.

- Transform the monthly profit value, considered as uniform installments (A) of R\$ 199,975.17 in a Principal (P) value at initial time 0, estimating that this profit value will repeat over the time horizon of 1 year, since there is no seasonality at the station and once the data are statistically described.

- Take the value of this Principal (P) found from the monthly profit obtained with 6 pumps and 5 employees, and subtract the Principal (P) value to be invested for the acquisition of the land (R\$ 130,000.00).

From these calculations, the following values were obtained for the Net Present Value and for the Equivalent Annual Cost for the 80 m² land acquisition scenario, considering a time horizon of 1 year.

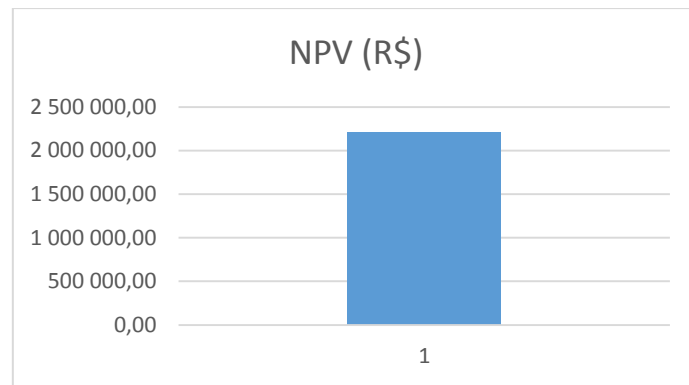


Figure 9. Net Present Value

Source: Elaborated by the authors (2017).

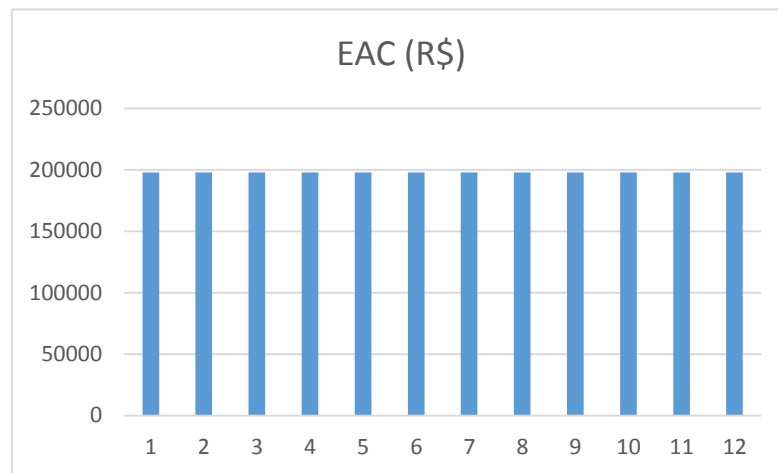


Figure 10. Equivalent Annual Cost

Source: Elaborated by the authors (2017).

It was verified that the value found for the NPV was R\$ 2,096,747.70. In addition, it is noted that the investment will generate a monthly profit, or EAC, of R \$ 188,300.40. Therefore, the positive value for these indicators demonstrates the viability of the investment.

In this context, the Payback calculation was performed to determine how long this investment will be paid, considering the value of the EAC (R\$ 188,300.40) as the monthly return (M) during the 1 year period. The calculations were made using the following equation:

$$\text{Payback} = \frac{M}{(1+i)^n}$$

Thus, it is noticed that the Payback was of 1 month, as it can be visualized by means of the following figure, elaborated through the software Microsoft Excel:

| Period | Payback |
|--------|--------------|
| 1 | 186,122.76 |
| 2 | 370,093.07 |
| 3 | 551,935.82 |
| 4 | 731,675.61 |
| 5 | 909,336.76 |
| 6 | 1,084,943.32 |
| 7 | 1,258,519.05 |
| 8 | 1,430,087.42 |
| 9 | 1,599,671.66 |
| 10 | 1,767,294.71 |
| 11 | 1,932,979.25 |
| 12 | 2,096,747.70 |

Figure 11. Payback

Source: Elaborated by the authors (2017).

In addition, since the amounts of taxes to be paid by the gas station were not provided, the scenario was considered with the payment of the Real Profit rate, which is characterized by a rate of 15% of the Gross Profit of a company. Thus, considering the Gross Profit, the NPV and EAC values found were as follows, based on a profit of R \$ 169,978.89, which corresponds to a decrease of 15% of the value of the profit previously found (R\$ 199,975.17):

| Scenario without taxes | NPV (R\$) | EAC (R\$) |
|------------------------|--------------|-------------|
| | 2,096,747.70 | 188,300.40 |
| Scenario with taxes | NPV - taxes | EAC - taxes |
| | 1,762,735.54 | 158,304.12 |

Table 4. Indicators considering the Real Profit

Source: Elaborated by the authors (2017).

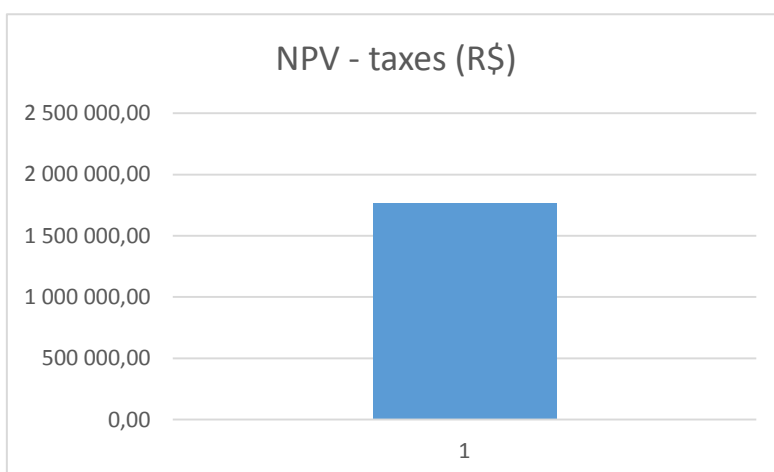


Figure 12. Discounted NPV

Source: Elaborated by the authors (2017).

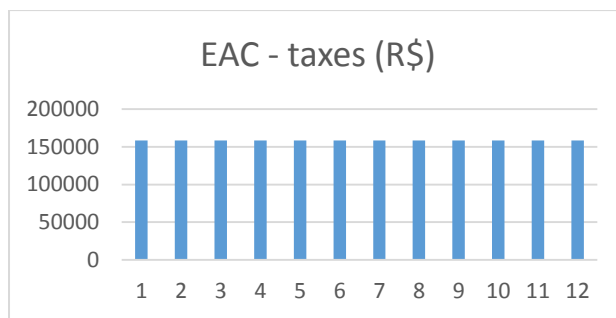


Figure 13. Discounted EAC

Source: Elaborated by the authors (2017).

It should be noted that despite the decrease in NPV and EAC values, these values remain positive, which demonstrates the feasibility of the investment in the acquisition of the land.

Finally, a comparison was made between the two situations, with 150 m² and 230 m², through the financial analysis, using the same indicators, in order to verify which one returned the best values.

In an analogous way to the calculations made for the 230 m² situation, the values obtained for the NPV and EAC considering the situation with available space of 150 m² were the following:

| NPV (R\$) | EAC (R\$) |
|-------------------|------------|
| 2,205,138.50 | 198,034.53 |
| MAR (% per month) | |
| 1.17% | |

Figure 14. TMA, VPL and EAC considering 150 m²

Source: Elaborated by the authors (2017).

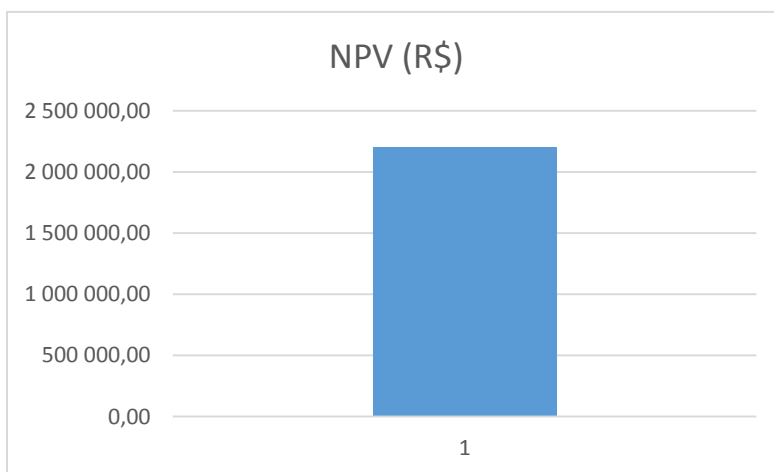


Figure 15. NPV (monetary unit)

Source: Elaborated by the authors (2017).

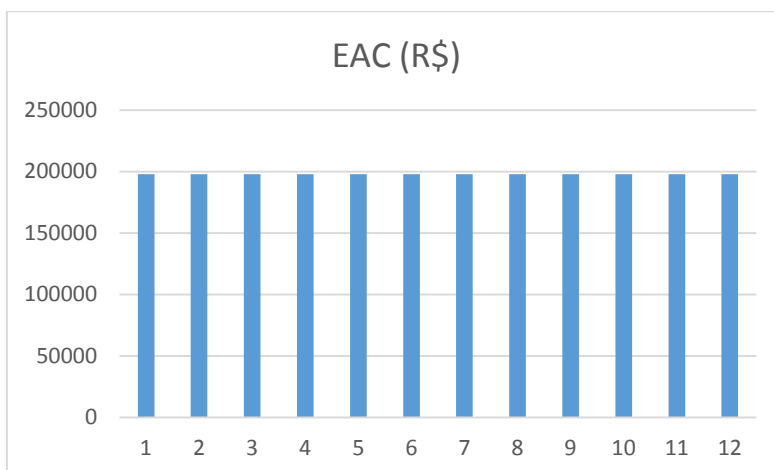


Figure 16. EAC

Source: Elaborated by the authors (2017).

Considering the Payback, if there is any investment, the values would be as follows:

| Period | Payback |
|--------|--------------|
| 1 | 195,744.33 |
| 2 | 389,224.93 |
| 3 | 580,467.99 |
| 4 | 769,499.38 |
| 5 | 956,344.68 |
| 6 | 1,141,029.17 |
| 7 | 1,323,577.85 |
| 8 | 1,504,015.40 |
| 9 | 1,682,366.25 |
| 10 | 1,858,654.53 |
| 11 | 2,032,904.08 |
| 12 | 2,205,138.50 |

Figure 17. Payback

Source: Elaborated by the authors (2017).

Considering the scenario with the payment of Real Profit, we have the following indicators:

| Taxes | % | Value of taxes |
|-------------|-----|----------------|
| Real Profit | 15% | 29,705.18 |

Figure 18. Tax

Source: Elaborated by the authors (2017).

| NPV - taxes (R\$) | EAC - taxes (R\$) |
|-------------------|-------------------|
| 1,874,367.72 | 168,329.35 |

Figure 19. NPV and EAC with taxes

Source: Elaborated by the authors (2017).

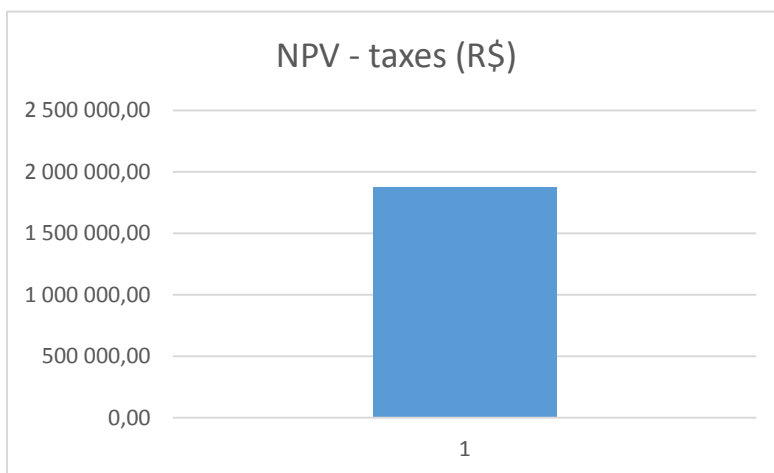


Figure 20. NPV considering taxes
Source: Elaborated by the authors (2017).

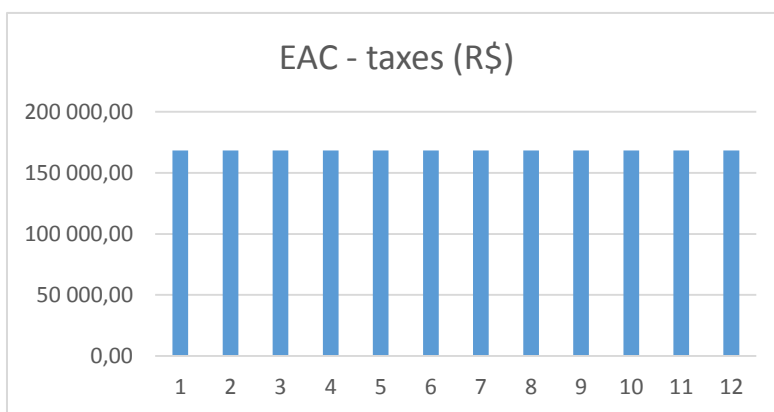


Figure 21. Discounted EAC
Source: Elaborated by the authors (2017).

The comparison between the two scenarios, with the acquisition of the land and without the acquisition, considering the scenarios with the presence and absence of the Real Income taxes, can be visualized through the following figure:

| | Monthly Profit | |
|-----------------------|---------------------|-----------------|
| | Not buying the land | Buying the land |
| Not considering taxes | 198,034.53 | 199,975.17 |
| Considering Taxes | 168,329.35 | 169,978.89 |
| | Investment | |
| | Not buying the land | Buying the land |
| Not considering taxes | 0.00 | 130,000.00 |
| Considering Taxes | 0.00 | 130,000.00 |
| | NPV | |
| | Not buying the land | Buying the land |
| Not considering taxes | 2,205,138.50 | 2,096,747.70 |
| Considering Taxes | 1,874,367.72 | 1,762,735.54 |
| | EAC | |
| | Not buying the land | Buying the land |
| Not considering taxes | 198,034.53 | 188,300.40 |
| Considering Taxes | 168,329.35 | 158,304.12 |

Figure 22. Comparison of scenarios
Source: Elaborated by the authors (2017).

The comparison between these same scenarios can be visualized graphically from the following figures:

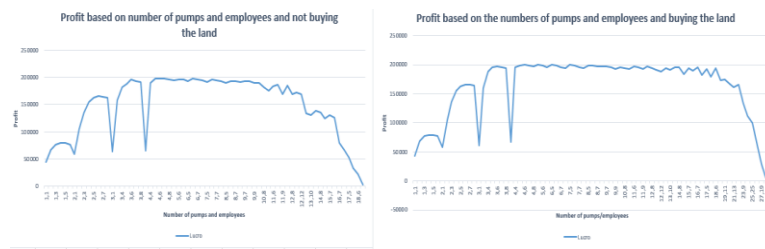


Figure 23. Graphs of profits based on the number of pumps and employees

Source: Prepared by the authors (2017).

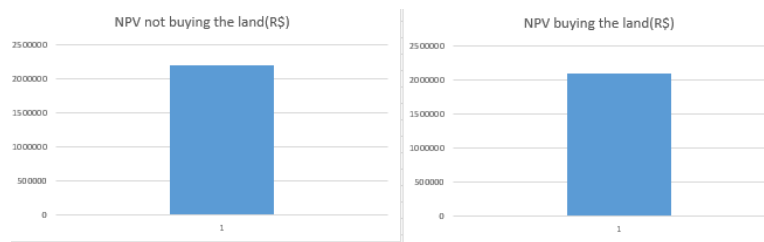


Figure 24. Graphs of NPVs without considering taxes

Source: Elaborated by the authors (2017).

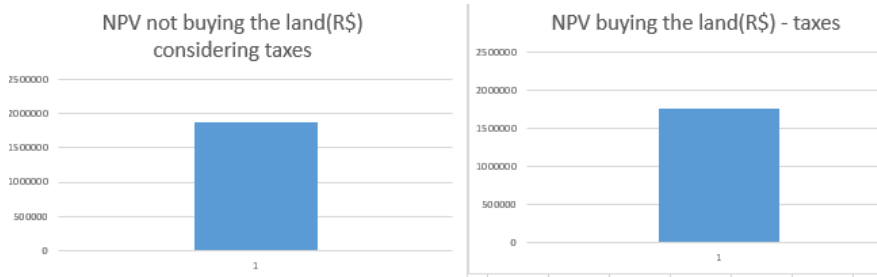


Figure 25. Graphs of NPVs considering taxes

Source: Elaborated by the authors (2017).

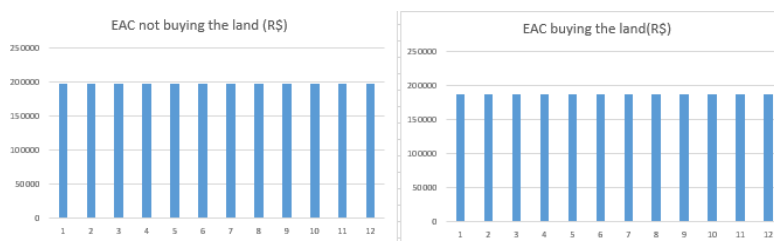


Figure 26. EAC's charts disregarding taxes

Source: Elaborated by the authors (2017).

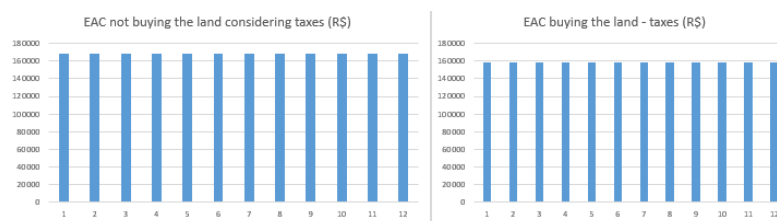


Figure 27. EAC's graphs considering the taxes

Source: Elaborated by the authors (2017).

Therefore, it is possible to perceive that the indicators of the scenario without the purchase of the land of 80 m² have a higher NPV and EAC and consequently better returns, since the values of the monthly profits are close to those obtained with the acquisition of the land. In addition, it is not necessary to do any investment that will decrease these indicators, such as land purchase, pumps, employees.

Thus, considering that in the scenario with 230 m² still have to be made investments in the purchase of pumps, payments of other taxes and expenses that did not fall within the scope of this work, the NPV and EAC calculated for this scenario will suffer a decrease. Thus, it is suggested that the acquisition of the adjacent land is not carried out, because if it is done, the calculated financial indicators will suffer a decrease because of the reasons cited and will become even smaller than the situation found for 4 pumps and 6 employees.

Conclusions

Based on the analysis performed at the gas station through the simulation, using Simul8 software, and considering the statistical data and assumptions of the model, it was possible to simulate several scenarios, manipulating the model variables, to determine the highest value of the model variable of interest, which was profit. Thus, through the simulations carried out, it was possible to find that the number of pumps and employees that generated the most profitable scenario was 4 and 6, respectively, considering an available space of 150 m².

Note the installation of 4 pumps that there is an available space of about 118 m² before entering the station the first car. When more pumps are installed, there are more cars that do not enter the station due to the reduction of available space. With the increase in the number of employees beyond 6, the cost increases in a ratio higher than the profit that these employees can generate by doing more services, which can be more idle, which justifies the fact that the profit has not increased when increases the number of pumps beyond 4 and employees beyond 6.

For the second situation, with a space of 230 m², it is noted that the scenario that generated the highest profit was with the number of pumps equal to 6 and with 5 employees. This increase in the number of pumps and employees, when compared to the first situation can be justified by the increase of the available area in the station. Thus, while more pumps can be installed, more cars can enter the station. However, although the EAC and NPV indicators indicate positive values for this scenario, they are lower than those found for the first situation, due to the investment to be made, which demonstrates the infeasibility of the investment in the acquisition of the land.

Specifically, with the methodology presented in this work, we used statistical data regarding the time intervals of arrivals and activities at the station. The purchase price of the initial 150 m² of the station was not taken into account, nor the purchase price of the pumps, so the profits are high. Costs with taxes, expenses and other variables were also not provided. Thus, these data, which are absent, provided cases, would make the profit become smaller and more concrete.

Therefore, by analyzing the results presented by this software, it was possible to conclude that Simul8 was a useful tool to develop a scenario analysis through simulation. Through it was possible to demonstrate the performance of the organization, as well as allowed the manipulation of variables of interest and obtaining desired values to determine the viability of an investment.

It is assumed that this methodology can be used in studies for other organizations with the purpose of simulating its operation and determining a configuration that returns a desired value for its variables of interest, such as profits, queue times, etc. If there is the possibility of making an investment to simulate other scenarios, a financial analysis, such as that performed in this work, using parameters

such as EAC, MAR, NPV and Payback can be useful in order to demonstrate the viability of an investment.

References

ALMEIDA, Mário. *Elaboração de projeto, TCC, dissertação e tese*. 2. ed. São Paulo: Atlas, 2014.

BRAGA, Roberto. *Fundamentos e técnicas de administração financeira*. São Paulo: Atlas, 1998

BRASILEIRO, Magaly Matias. **Manual de produção de textos acadêmicos e científicos**. São Paulo: Atlas, 2013.

Banks, J., 1998. *Handbook of Simulation*. USA, John Willey & Sons

Bangsow S. 2010. *Manufacturing Simulation with Plant Simulation and SimTalk*. Springer-Verlag Berlin Heidelberg

CALVACANTE, Francisco. Valor Presente Líquido (VPL). Disponível em: <http://www.cavalcanteassociados.com.br/article.php?id=61>. Acesso em: 20 de Novembro de 2017.
CARDOSO, Suzana;

Concannon, K. H., Hunter, K. I., & Tremble, J. M. (2003, December). Dynamic scheduling II: SIMUL8-planner simulation-based planning and scheduling. In *Proceedings of the 35th conference on Winter simulation: driving innovation*(pp. 1488-1493). Winter Simulation Conference.

Concannon, K. et al. 2007. *Simulation Modeling with SIMUL8*. Visual Thinking International, Canada.

Dlouhý, M., Fábry, J., Kuncová, M. and T. Hladík. 2011. *Business Process Simulation* (in Czech). Computer Press.

FERREIRA, Nelson. (2011) *Análise de Viabilidade econômica e suas implicações na obtenção do crédito bancário*. Disponível em: <http://www.lume.ufrgs.br/bitstream/handle/10183/77628/000894682.pdf?sequence=1>. Acesso em: 20 de Novembro de 2017.

Ficová, P. and M. Kuncová. 2013. „Looking for the equilibrium of the shields production system via simulation model“. In *Modeling and Applied Simulation 2013*. (Athens, Sept. 25-27). Genova : DIME Università di Genova, 50–56.

Greasley, A. 2003. *Simulation modelling for business*. Innovative Business Textbooks, Ashgate, London.

KOPITTKÉ, H. Bruno e CASAROTTO FILHO, Nelson. *Análise de Investimentos*. São Paulo: Atlas, 2000.

Law, A. M., Kelton, W. D., & Kelton, W. D. (1991). *Simulation modeling and analysis* (Vol. 2). New York: McGraw-Hill.

Montevecchi, J. A. B., et al. 2007. Application of design of experiments on the simulation of a process in an automotive industry. In WSC'07 Proceedings of the 39th Conference on Winter Simulation, IEEE Press Piscataway, NJ, USA, 1601-1609.

O'Kane, J.F. et al., 2000. Simulation as an essential tool for advanced manufacturing technology problems. Journal of Materials Processing Technology 107/2000, 412-424

Shalliker, J. and C. Ricketts. 2002. An Introduction to SIMUL8, Release nine. School of Mathematics and Statistics, University of Plymouth.

Author's Information



***Author One, Denis Ramos de Oliveira,
Production Engineer
Universidade Federal de Viçosa, Institute of
Exact and Technological Sciences
Denis.ramos@ufv.br***



***Author Two, João Pedro Fonseca de
Barcelos, Production Engineer
Universidade Federal de Viçosa, Institute of
Exact and Technological Sciences
jpfbarcelos@gmail.com***



***Author Three, Thiago Henrique Nogueira,
Dr.
Universidade Federal de Viçosa, Institute of
Exact and Technological Sciences
Research Area: Operational Research
thnogueira.ufv@gmail.com***

GEOGRAPHICALLY WEIGHTED REGRESSION IN ANALYSIS OF INFORMATION AND COMMUNICATION TECHNOLOGY DEVELOPMENT IN INDONESIA

Dwi Puspita Sari, Jamilatuzzahro, Vijay Kanabar

Metropolitan College, Boston University
Data Science Indonesia

***Abstract:** The main purpose of this paper was to analyze the development of Information and Communication Technology (ICT) in Indonesia with Geographically Weighted Regression (GWR) to enable the identification of the variability of regression coefficients in the geographical. The study has been conducted using the statistical data for 33 provinces in Indonesia. There were 3 independent variables defined during this study as factors that influence the development of ICT in Indonesia: location, economic prosperity, and education, and internet users in Indonesia has been defined as dependent variable. The result of this research shown the highly correlation between those three independent variables toward defined dependent variable shown by the value of R square of 0.7310782 which means the dependent variable can be well explained by independent variables.*

***Keywords:** ICT, GWR, spatial data analysis, Indonesia*

1. Introduction

Information and communication technology (ICT) covers two aspects: information technology and communication technology. Information technology includes the process and use of tools in information management, while communication technology includes the process of transferring data from one device to another device. This research focused on the development of computing device as information technology devices and internet access as information communication technology tools used by Indonesian population in 33 provinces in Indonesia.

Indonesia is ranked sixth among internet users in the world in 2018 with a population of about 123 million people accessing the internet ^[1]. The large number of the population who are connecting and accessing the internet with various purposes clearly indicates that internet users have access to the use of information technology tools such as computers, laptops, smartphones, and tablets.

There are numerous studies about the development of ICT in Indonesia as well as the factors influence the development and deployment of ICT infrastructure in Indonesia. The three main factors that highly influence ICT development in Indonesia, namely location, economic prosperity, and education. The main goal of this paper is to identify ICT development in Indonesia using GWR model approach. This study estimates the distance in influencing ICT development between each province. The GWR model approach enables researchers to identify the correlation between the independent variables defined in the

literature review section and the development and deployment of ICT process in Indonesia. The identification of percentages for the ICT development across the country is also defined in analysis and discussion section of this paper.

This paper consists of five sections. In section two, the introduction of ICT in Indonesia, the factors affect the development of ICT in Indonesia, and spatial data are elaborated in details. In section three, methods used during the research are described. The methods are GWR, Inferencing Analysis of Geographically Weighted Regression, Goodness of Fit Test, Parameter Coefficient Test, and Coefficient of Determination (R Square) Local. In section four, the result of analysis for ICT development in Indonesian is defined and well discussed. The analysis has been conducted using data from some organizations related to ICT in Indonesia and dataset sourced from Indonesia Government official website. The data collection includes the technology usages and education distribution in Indonesia. Section five is a final section that provides the summary and research conclusion. In this section also, the estimation parameter of internet users is defined based on findings formula. In the conclusion section, the most factor influences the development of ICT in Indonesia is defined and ranked based on the data analysis. This section estimates the highest ICT development in the future based on findings formulation. In addition, the future work is briefly discussed.

2. Literature Review

Indonesia is categorized as a very diverse country with more than 16,000 islands across the territory with more than 300 different vernaculars, various education levels of population, and also diversity of ethnicity and cultures in each state with uneven economic level, it can be expected that the determinants of ICT development are diverse in the geographical space based in each province. This section consists of three sections: Information and Communication Technology (ICT) in Indonesia, Factors of ICT Development, and Spatial Data.

2.1 Information and Communication Technology (ICT) in Indonesia

Technology changes the way people communicate, it establishes the interaction with distance matter becomes easier and efficient. Internet not only connects between one person another but also one organization to another organization or partner. Association of Internet Service Providers in Indonesia (APJII) survey shown that there was a significant increasing amount of internet technology usage in each year in Indonesia. In past years, there were 132.7 million of 256.2 million Indonesian population who accessed the internet, 50.7% accessed the internet using mobile devices and computers followed by mobile devices and computers only with 47.6% and 1.7% ^[2]. It shows that about one-third the Indonesian population has an access to computer technology and internet connection.

2.2 Factors of ICT Development

The era of globalization and moderation cannot be avoided by countries in the world from various aspects of life with any positive or negative impacts. Positive impacts can be ease in transferring various science, technology, and any other beneficial knowledge from well-developed country to Indonesia, however the negative impacts easily arise and becomes harmful to Indonesia as an example of the ease of the Indonesian population to access the foreign countries cultures without having a certain filter from the Indonesian nation that is easily entered and influence the existence cultures ^[3]. Internet-connected technologies have changed the way people communicate such as decreasing number of face-to-face meetings due to the widespread use of social media and applications that support virtual encounters. It is undeniable that technological advances and the use of technology wisely greatly help today's human's daily activities to simplify their lives and improve the quality of life for the better living standard.

During this study, there were 3 main factors identified that influenced the development of ICT in Indonesia: location, economic prosperity, and education.

2.2.1 Location

Indonesia is a unitary state, which is a republic country that is located with longitude and latitude at 0.7893o S and 113.9213o E. The bandwidth of the radius size in geographically weighted regression method can be analogous to the radius of a circle, so that an area within the circle radius is still considered to have an influence. In this case, the territory is defined by the province in Indonesia. If the weights used are kernel functions, then the selection of bandwidth becomes very important because the bandwidth is used as a balance control between the suitability of the curve to the data and the data graduation. A considerable bandwidth value will cause an even greater bias due to the over-smoothing model caused by the number of observations used.

2.2.2 Economic Prosperity

In 2017, the population of Indonesian below the poverty line is 26.58 million people.

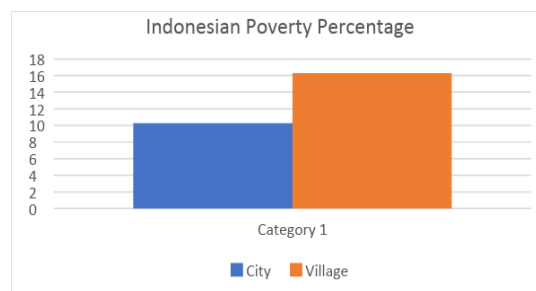


Figure 2.1 Indonesian Poverty Percentage

Source: Badan Pusat Statistics - Statistics Indonesia (2014)

Gross Regional Domestic Product (GDRP) also known as a gross domestic product of region (GDPR) is defined as the amount of value-added goods and services generated from all economic activities in the certain area. In this study, researchers used GDRP to define the correlation between the development of ICT in Indonesia in each province. The increasing amount of GDPR in each year supports the development of ICT in Indonesia that shown from the increasing number of devices purchasing in each year.

2.2.3 Education

In 2017, the total percentage of illiteracy in Indonesia reached 16.52% which is divided into 3 different categories based on population age: 4.5% for age younger than 15 years old, 0.94% for age in between 15 to 44 years old, and 11.8% for age older than 45 years old ^[4].

In 2014, the total formal education institutions ranging from elementary school to college level reach 132,407 institutions ^[5].

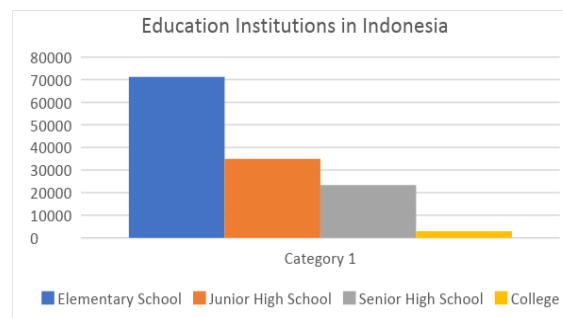


Figure 2.2 Education Institution in Indonesia

Source: Badan Pusat Statistics - Statistics Indonesia (2014)

In 2017, the total Indonesian population who have completed education from elementary school to tertiary degree reach 305.68 million people ^[6].

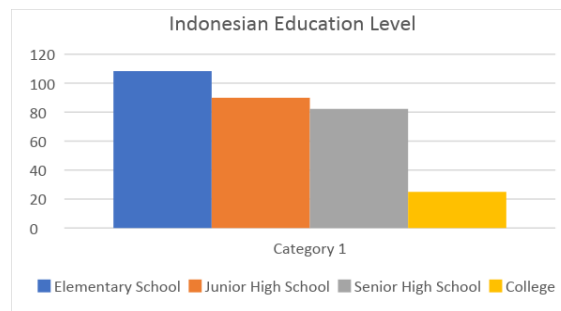


Figure 2.3 Indonesian Education Level

Source: Badan Pusat Statistics - Statistics Indonesia (2017)

In this study, the education factor in each province defines as the percentage of illiterate Indonesians. The researchers found that there was a negative correlation between education and internet users in Indonesia that indicates the highest illiterate Indonesian in each province influence the number of internet users.

2.3 Spatial Data

Spatial data are also known as a geospatial data is a data that has a georeferenced reference in which various attribute data lies in various spatial units. Spatial data contains information about attributes and location information. As an illustration of the number of people infected with diseases in various regions, it shows that spatial data is dependent data because it comes from a different location indicating a dependency between the measurement value and the location.

Tobler's first law of geography in Schanbenberger and Gotway states that "everything is related to everything else, but near things are more related than distant things" which means: everything is interconnected one with the other, however, something closer has more influence than something far away^[7].

The condition of the location of the observation will be different from the other observation location. However, the condition of an observation site will have a close relationship with the adjacent observation location. These relationships are called spatial effects. Spatial effects arise due to differences in environmental and geographical characteristics between the observation sites so that each observation may have a different variance or there is a difference in the influence of predictor variable to the variable response for each location of observation. This spatial effect is then referred to as spatial diversity or spatial heterogeneity.

A statistical method is required that anticipates spatial heterogeneity. The statistical method is a geographically weighted regression or geographically weighted regression (GWR)^[8].

Spatial coordinate variables of longitude and latitude are the variables used in weighing in the formation of GWR models. Longitude is a longitudinal line connecting between the north and south sides of the earth (poles) used to measure the east-west side of the coordinates of a point in the hemisphere. While the latitude is the transverse line between the north pole and south pole that connects between the eastern and western sides of the earth which are used as a measure in measuring the north-south side of the coordinate of a point in the hemisphere.

3. Methods

3.1 Geographically Weighted Regression

Geographically Weighted Regression (GWR) was first introduced by Fortheringham in 1967. The GWR model is the development of the classical linear or ordinary linear regression (OLR) model. The GWR model is a regression model that aims to model the continuous response variable by taking into account the spatial or location aspect, the spatial heterogeneity. Spatial heterogeneity occurs when a single independent

variable gives unequal responses at different sites within a study area. The approach taken by GWR is the point approach. Each parameter value is estimated at each point of the observation location so that each point of the observation location has different parameter values.

The geographically weighted regression model can be written as follows:

$$y_i = \beta_0(u_i v_i) + \sum_{k=1}^p \beta_k(u_i v_i) x_{ik} + \varepsilon_i, \quad i = 1, 2, 3, \dots, n$$

- y_i : observation value of the i -th response variable
 x_{ik} : estimated value of the k -predictor vector of the i -location observation
 $(u_i v_i)$: defines the geographical location (longitude, latitude) at the i -th sight location
 $\beta_0(u_i v_i)$: constants on the i -th observation
 $\beta_k(u_i v_i)$: observed value of the k -predictor variable at the i -th observation location
 ε_i : The i -th observational error, assumed to be $N N \sim (0, \sigma^2)$

Inferential statistics are data analysis techniques used to determine the extent to which the similarity between the results obtained from a sample with the results to be obtained in the population as a whole. Statistics of inference need to be done to determine whether the parameters in the Geographically weighted regression model are significant or not. The following inferential statistical tests are applied in the GWR model.

3.2 Coefficient of Determination (R Square) Local

The value of determination coefficient (R Square) can be used to predict how big contribution of free variable influence (X) to the dependent variable (Y) provided that test results in regression analysis are significant.

R square values are formulated as follows:

$$R^2 = 1 - \left(\frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \underline{y}_i)^2} \right) \left(\frac{n-1}{n-p-1} \right)$$

with

- y_i : observation response to $-i$
 \underline{y} : average
 \hat{y}_i : prediction response to $-i$
 n : number of observations
 p : number of parameters in the model

4. Results and Discussion

Over more than 16,000 islands in Indonesia, there are 6 big islands with the highest population among others, namely Sumatera, Jawa, Bali and Nusa, Kalimantan, Sulawesi, and Papua and Maluku. Among those 6 big islands, Jawa has the highest internet users with about 86.3 million population and on the other hand Papua and Maluku island has about 3.3 million population.

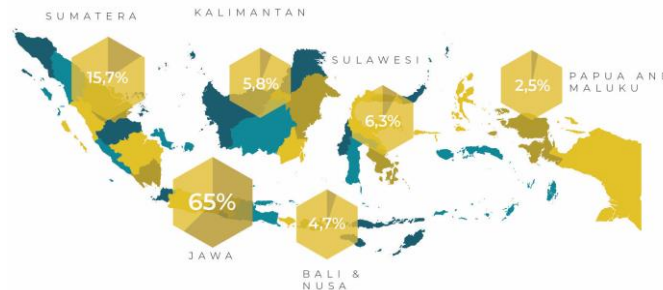


Figure 4.1. Internet Users in Indonesia
Source: APJII 2016

To model internet users' estimation in Indonesia that was influenced by three factors mentioned earlier, this study has used statistical modelling for each province. Linear regression method was used to measure whether there is a correlation between three defined independent variables and dependent variable. The results shown that there is less significant correlation between defined independent variables and dependent variable.

Based on the linear modeling above, the P-Value has been defined for each independent variable: Gini Ratio, income per capita, and illiteracy level of Indonesian population. The significance level for three of the independent variables shown 0.0345, 0.1889, and 0.9193 respectively for Gini Ratio, income per capita, and illiteracy level of Indonesian population, and the overall P-Value is 0.007267. As the P-Value shown for each individual independent variable and overall variables, it shows that the P-Value is below than 0.05 as a result the regression cannot describe this model and it assumed that this model is influenced by geographical location, so spatial regression modeling such as GWR is needed in this case.

GWR uses bandwidth determination for the model estimation with the bandwidth value of 145219838. In this case, the bandwidth with criteria of minimum cross validation (CV) has been chosen, so bandwidth was gathered from minimum CV of 0.09101971.

The feasibility test of GWR model has obtained P-value 0.04102 that is less than 0.05 so it is concluded that GWR model approach is the most suitable method to use.

The value of R square is 0.7310782. From the output above, the R square is 0.7310782, the dependent variable of 73.11% can be explained by the independent variable. The statistics description of parameter estimation obtained for each province in Indonesia is shown in the table 4.1 below.

Table 4.1. Estimation Parameter for 33 Provinces in Indonesia

| Province | Intercept | Gini Ratio | GDRP | Illiteracy |
|----------|-------------|------------|------------|-------------|
| Aceh | -3160720.33 | 12474804 | -13013.521 | -553169.577 |
| Sumut | -879355.91 | 5742782 | 36297.34 | -697277.795 |
| Sumbar | -89656.15 | 3407254 | -12794.981 | -897546.882 |
| Riau | -1223726.48 | 6268264 | -46321.117 | -527723.299 |
| Jambi | 1174427.01 | -1513551 | - | -534590.363 |
| Sumsel | 847860.09 | -1256467 | 275719.709 | -387130.284 |
| Bengkulu | 112558.2 | 1523260 | - | -671079.781 |

| | | | | |
|------------|--------------|----------|------------|-------------|
| lu | | | 227186.126 | |
| Lampung | -4163840.21 | 12889308 | 277764.023 | -437299.874 |
| Babel | -1047205.16 | 3956186 | 286930.737 | -95141.358 |
| Kep. Riau | -6385518.9 | 20175021 | 186625.387 | -361704.912 |
| Jakarta | -12752256.17 | 36848162 | 225567.617 | -450397.234 |
| Jabar | -9050225.76 | 28159568 | 168836.061 | -773919.896 |
| Yogya | -10964820.4 | 33735328 | -58017.87 | -685310.575 |
| Jatim | -12212724.43 | 36530558 | 3150.165 | -476685.733 |
| Jateng | -10640460.93 | 32663740 | -72981.758 | -659728.714 |
| Banten | -14187853.08 | 39756180 | 222711.711 | 28571.954 |
| Bali | -9396722.35 | 28295622 | -73530.2 | -510474.772 |
| NTB | -9165334.12 | 27376924 | -79287.388 | -481736.088 |
| NTT | -4946728.38 | 14514857 | -78110.003 | -69260.554 |
| Kalbar | -4364904.05 | 14210497 | 174759.876 | -311959.4 |
| Kalteng | -7285211.24 | 22332411 | 111321.862 | -435497.892 |
| Kalsel | -8501833.19 | 25724486 | -80314.561 | -451597.262 |
| Kalut | -4569033.37 | 13812070 | 115185.187 | -235403.103 |
| Sulut | -1893401.06 | 5802634 | -74881.782 | -31537.374 |
| Sulteng | -2530627.52 | 7792689 | 113271.912 | -132343.116 |
| Sulsel | -5833721.83 | 17237334 | -105283.46 | -297559.98 |
| Sulteng | -3185795.84 | 9491701 | -85787.298 | -133444.165 |
| Gorontalo | -1807132.56 | 5697482 | -85276.016 | -90872.138 |
| Sul. Barat | -3882982.75 | 11753460 | -85763.692 | -165078.04 |
| Maluku | -2153115.04 | 6418001 | -51569.274 | 10552.715 |
| Mal.Utara | -1188427.67 | 3784704 | -37122.973 | -1763.505 |
| Pap. Barat | -1596119.61 | 4793267 | -40344.304 | 16388.731 |
| Papua | -3512679.78 | 10243602 | -85268.917 | 19387.976 |

The estimation for internet users in each province in Indonesia calculation can be formulated based on this formula. The formula for internet users for each province is written as follow:

$$Y = \text{Intercept} + \text{Gini Ratio (A)} + \text{GDPR (B)} + \text{Illiteracy(C)}$$

APJII survey shows that 65% of internet users in Indonesia is in Jawa Island where the capital city of Indonesia is located. It has been identified that there is about 83% population in Jakarta has an access to the internet. There is 8.48 million Indonesian population who have access to the internet from about 10.2 million of the population ^{[9] [10]}. Jakarta capital territory is divided into six sub-territories: North Jakarta, South Jakarta, West Jakarta, East Jakarta, Central Jakarta, and Kepulauan Seribu Island.

This study focused more on Jawa Island and more specifically to DKI Jakarta province as a capital city of Indonesia. Based on the formulation above, the estimation formula for internet users in DKI Jakarta is written as follows:

$$Y = -12752256.17 + 36848162(A) + -225567.617(B) + -450397.234(C)$$

Where the variable A is an estimation parameter of Gini Ratio, B is an estimation parameter of GDRP and C is an estimation parameter of illiteracy using GWR modeling.

Based on the most recent data for 2017, Gini Ratio of DKI Jakarta is 0.409, GDRP is -6.183 and Illiteracy is 0.08. From the formula above, the estimation of internet users can be defined by plugging the value of each variable.

$$Y = -12752256.17 + 36848162(A) + -225567.617(B) + -450397.234(C)$$

$$Y = 12752256.17 + 36848162(0.409) + -225567.617(-6.183) + 450397.234(0.08)$$

$$Y = 12752256.17 + 15070898.26 + 13946845.576 - 36031.77872$$

$$Y = 41733968.23$$

Jawa island consists of six provinces: DKI Jakarta, Jawa Barat, DI Yogyakarta, Jawa Timur, Jawa Tengah, and Banten. In 2017, the statistic report has shown that there are about 86.3 million people who access the internet in Jawa island. From the formula above, the estimation of internet users in the next 10 years can for each province calculated by plugging the numbers of Gini Ration, GDRP, and illiteracy. The estimation is shown in the table below:

Table 4.2 Estimation Internet Users for Jawa Island

| Province | Estimation of Internet Users |
|-----------------------|------------------------------|
| DKI Jakarta | 41733968 |
| Jawa Barat | 28826422 |
| DI Yogyakarta | 28221221 |
| Jawa Timur | 30940032 |
| Jawa Tengah | 29900349 |
| Banten | 27914325 |
| Total estimated users | 187536317 |

The estimation calculation has shown that the estimation of users in Jawa Island will be double in upcoming years which indicates the development of ICT is rapidly developing in the future.

Conclusion and Future Work

The study objective was to find whether the correlation between defined three independent variables affect independent variables where there are spatial factors between those variables. The study result found that the most factor influences the development of ICT in Indonesia is economy prosperity that shown from the strong correlation of 0.5411 between the internet users in each province and economy prosperity. The R square shows a significant result that is 0.7. GWR is a predictive model for spatial case used in this research to estimate the correlation of technological developments for each province by three factors are education, economic prosperity and illiteracy. Each factor influences independent variable with different estimation of each province in Indonesia.

The future work for this research will be focus not only to Jawa island where the highest amount percentage of internet users located, but also to other 5 biggest island in Indonesia including Sumatera

Island, Kalimantan Island, Sulawesi Island, Bali and Nusa Island, and Papua and Maluku Island. The future work also will define more independent variables that might affect the development of ICT in Indonesia.



Bibliography

- [1] Ministry of Communication and Informatics (2014, November 2014). Pengguna Internet Indonesia Nomor Enam Dunia. Retrieved from https://kominfo.go.id/content/detail/4286/pengguna-internet-indonesia-nomor-enam-dunia/0/sorotan_media.
- [2] Asosiasi Penyelenggara Jasa Internet Indonesia (2016). Penetrasi dan Perilaku Pengguna Internet Indonesia. Retrieved from <https://apjii.or.id/download/file/surveipenetrasiinternet2016.pdf>.
- [3] Nasution, R.D. (2017). Effect of the development of communication information technology on local cultural existence. *Jurnal Penelitian Komunikasi Dan Opini Publik*, 21(June 2017), 1st ser,m 30-42.
- [4] Badan Pusat Statistik (2017). Jumlah Penduduk Miskin, Persentase Penduduk Miskin dan Garis Kemiskinan. Retrieved from <https://www.bps.go.id/statictable/2014/01/30/1494/jumlah-penduduk-miskin-persentase-penduduk-miskin-dan-garis-kemiskinan-1970-2017.html>.
- [5] Badan Pusat Statistik (2017). Percentage of Illiteracy. Retrieved from <https://www.bps.go.id/linkTableDinamis/view/id/1056>.
- [6] Badan Pusat Statistik (2014). Number of Villages Having Educational Facilities by Province and Education Level. Retrieved from <https://www.bps.go.id/linkTableDinamis/view/id/905>.
- [7] Schabenberger O, Gotway CA. 2005. *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC.
- [8] Fotheringham A S, Brunson C, Charlton M, 2002 *Geographically Weighted Regression—the Analysis of Spatially Varying Relationships*, Wiley, Chichester.
- [9] Badan Pusat Statistik (2016). Jakarta Dalam Angka 2016. Retrieved from https://jakarta.bps.go.id/backend/pdf_publicasi/Jakarta-Dalam-Angka-2016.pdf
- [10] Katadata (2018). Berapa Jumlah Penduduk Jakarta? Retrieved from <https://databoks.katadata.co.id/datapublish/2018/01/24/berapa-jumlah-penduduk-jakarta>

Author's Information



Dwi Puspita Sari
Master of Science in Computer
Information Systems Candidate
Metropolitan College, Boston University
[***dsardiyo@bu.edu***](mailto:dsardiyo@bu.edu)

| | |
|---|---|
|  | <p>Jamilatuzzahro Head of Research Development and Knowledge Management Data Science Indonesia <u>jamilatuzzahro@datascience.or.id</u></p> |
|  | <p>Vijay Kanabar Director of Project Management Programs and Associate Professor of Computer Science Metropolitan College, Boston University <u>kanabar@bu.edu</u></p> |

CASE STUDY: POWER BI VISUALIZATIONS APPLIED TO DIGITAL VIDEO PUBLISHER'S ADVERTISING SPACE

Joseph CHOMSKI

***Abstract:** As video viewing migrates from cable TV to increasingly fragmented platforms - web browsers, mobile phones, and over-the-top devices such as Roku and AppleTV – digital video publishers dependent upon advertising revenue face increasing operational challenges. Each platform has its own standards and data formats, creating multiple heterogeneous buckets of perishable advertising space. This case study presents how Microsoft Power BI visualizations were used to provide a holistic performance management view of advertising space, and helped optimize monetization of each platform by raising awareness of KPI's to operations managers.*

Keywords: Visualization, Microsoft Power BI, Digital Advertising, Ad Operations, Ad Inventory, Digital Media, Video Publisher

Introduction

This paper presents a case study of how Microsoft Power BI visualizations were used to help solve operational challenges in filling a digital video publisher's (DVP's) advertising space. The paper will show how consumer video viewing trends have resulted in a complex, fragmented environment for DVP's who depend on ad revenue. An analytics framework will be presented, highlighting important dimensions and metrics, as well as key performance indicators (KPI's). The reasons why Microsoft Power BI was selected will be covered, as well as how the self-service Business Intelligence (BI) approach was used. The Performance Management section of this paper will show how the analytics solution is integrated into decision-making processes, and how the solution has yielded key learnings and business value.

Digital Video Publishers Background

In the context of this paper, "digital" encompasses all platforms and devices other than traditional TV. DVP's can be of four main types: 1) traditional media companies such as Viacom and CBS having both TV and digital content containing advertisements; 2) Digital-only media companies such as Hulu and YouTube, containing limited advertisements; 3) Subscription-only digital media companies such as Netflix and Amazon Prime, containing no advertisements; 4) Social media companies such as Facebook and Snapchat with emerging ability to support advertisements. The focus of this paper will be primarily on the digital aspect of Type 1 DVP's, but is also applicable to Types 2 and 4 as they are all dependent upon Advertising Sales (Ad Sales) as a revenue source, i.e. paid ad campaigns sold to advertisers. Type 3 DVP's such as Netflix require only integration of video content delivery with new devices and platforms, but DVP types 1, 2, and 4 must also integrate ad technology (ad tech). They are thus more impacted by changing consumer preferences as to what devices they consume (i.e. watch) video content on.

Digital Video Publisher Industry Changes

In the past, most video consumption took place via traditional TV, and to a limited extent, desktop PC's via web browsers. Today, consumers – particularly younger audiences – are increasingly "cutting the cord" and doing without traditional TV (Perrin, 2017). Furthermore, digital video consumption has

fragmented across many devices. While desktops still have the richest ad targeting and measurement capabilities, more and more consumption takes place on mobile phones, tablets, and over-the-top (OTT) devices such as Roku and Apple TV (Wang, Scher, Yao, Rosenblum, & Dogaru, 2017).

Digital Video Viewing

The viewing process begins with a user initiating a session on their device of choice and requesting video content, launching a video content stream interspersed with ad breaks (Figure 1).

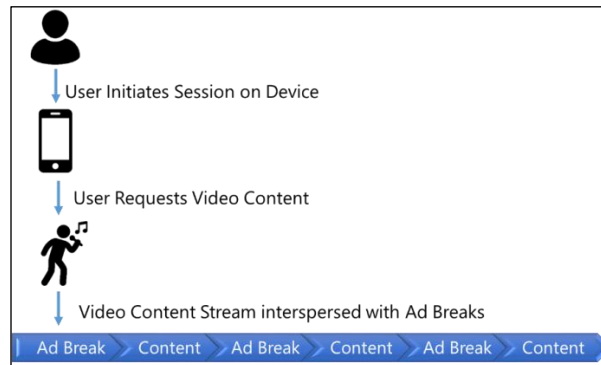


Figure 1. How Digital Video Viewing Works

Digital Advertising Space

Total advertising space is an aggregation of all the ad breaks generated by large numbers of user sessions on multiple platforms (Figure 2).

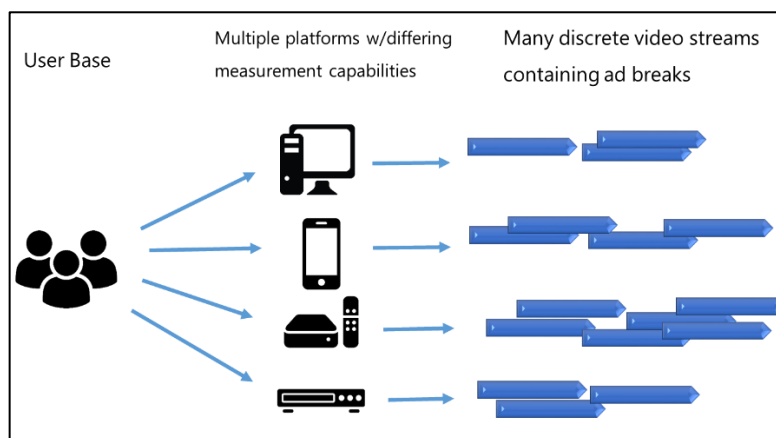


Figure 2. How Digital Advertising Space is Generated

Each digital video platform has its own targeting and measurement capabilities, requiring partnerships with multiple companies for operational integration. This heterogeneous environment creates significant complexity in monetizing the increasingly fragmented advertising space. DVP operational managers strive to keep all the “buckets” of advertising space as full as possible, because if advertising space is not filled with paid client ad campaigns, it is lost. The situation is similar to maximizing yield of hotel room inventory, but with rooms (analogous to advertising space) spread across multiple hotels (analogous to the various digital platforms), and with the hotels constantly changing in size (analogous to varying video consumption volumes).

Business Problem

Ad Sales’ overarching goal is to grow revenue. Advertising Operations contributes to this goal by optimizing the yield of available advertising space. The problem is twofold:

- 1) How do we fill all available advertising space?
- 2) How do we capitalize on changing user platform preferences?

Monitoring each bucket of advertising space requires a great deal of effort by operational managers. As total advertising space continues to fragment into smaller and smaller buckets, it is no longer practical to manage them individually via lengthy reports.

Analytics Framework

A better approach is to use an analytics framework to provide management with top-level views and the ability to quickly highlight problem areas. This answers the Descriptive Analytics question of “What happened?” and supports root cause analysis. The principal KPI is fill rate, defined as:

$$\text{Fill Rate} = \frac{\sum \text{Ad Impressions}}{\sum \text{Ad Slots}}. \quad (1)$$

Maximum fill rate is 100%, and the goal is to keep it as high as possible. Component measures of formula (1) are defined in Table 1.

Table 1. Component Measures of Fill Rate KPI

| Measures | Definition |
|----------------|---|
| Ad Impressions | A viewing of an ad by one person |
| Ad Slots | A portion of an Ad Break in which it is possible to display an ad. Ad Breaks usually contain multiple ad slots. This measure represents the maximum number of ads possible. |

The KPI component measures can be summed across several key dimensions listed in Table 2 below, corresponding to the ways advertising space can be bucketed/sliced.

Table 2. Dimensions of Fill Rate KPI

| Dimension | Definition | Example |
|--------------------|---|-----------------------------|
| Channel | Content Owner | Nickelodeon, Comedy Central |
| Platform | Type of device on which video is viewed | Desktop, Mobile App, OTT |
| Partner | Video distributor via which video is viewed by consumer. Has partner relationship with DVP. | Comcast, Youtube |
| Filled vs Unfilled | Indicates whether an ad slot was monetized e.g. by the sales force or via programmatic ad buying, or was not monetized. | Filled, Unfilled |

Presenting KPI's visually is an ideal way to monitor performance and flag issues, and provides situational awareness (Few, 2013). Given the analytical sophistication of operational users accustomed to Microsoft Excel, the challenge is to provide them with a visualization tool that allows them to quickly adapt the solution in response to changing business conditions, without intervention from the Information Technology (IT) department.

Solution: Microsoft Power BI Visualizations

Microsoft Power BI is an ideal visualization tool in this case, as users appreciate its similarity and common user interface with Microsoft Excel, which increases comfort level, accelerates the learning curve, and encourages adoption. It supports a self-service BI approach, with users creating their own dashboards, data models, and datasets. Its data preparation features allow users to merge multiple data sources together and perform basic extract, transform, and load (ETL) functions without requiring involvement by the IT department. It is cloud-based, with no servers required, and its cost is very low. Gartner Group ranked Power BI at the top of its 2018 "Magic Quadrant" review of Analytics and BI tools (see Figure 3).

**Figure 3.** Gartner 2018 Magic Quadrant for Analytics and BI Tools

Power BI Architecture

Power BI Architecture can be divided into On-Premises and Cloud Services components (see Figure 4). In this case, On-Premises components include the data sources, as well as the Power BI desktop client, which is used for data modeling and visualization authoring. The bulk of the data is sourced from a data warehouse, supplemented by Excel spreadsheets which users merge into the dataset. The Power BI cloud service is used for publishing and sharing of visualizations, with access provided via web browser and

mobile app. Access via mobile device helps achieve Pervasive Analytics, where managers can access actionable information from anywhere (Lachev, 2018), including in meetings and while traveling.

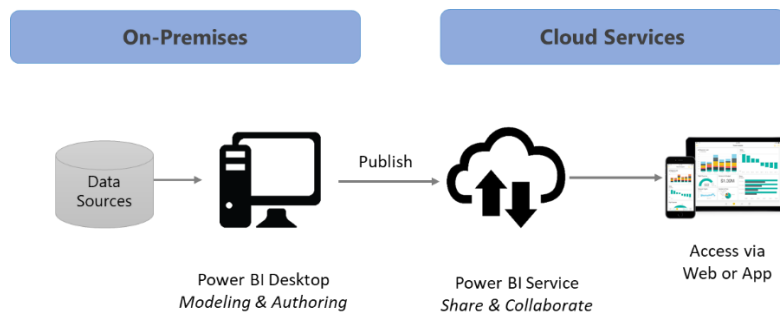


Image Credit: Microsoft

Figure 4. Power BI Architecture

Approach: Business-Led Self-Service BI

In the self-service BI approach, business users are responsible for all aspects of the solution, without requiring IT resources (Guillen & Coates, 2016). A current state assessment was performed, followed by the prototyping of initial seed datasets which were then published to operational users to design visualizations (see Figure 5). Because it is user-maintained, the solution provides agility and the ability to rapidly adapt to new business partnerships and platform integrations. Previously, this was achieved via Excel, but required considerable manual effort to refresh the data; the solution has cut refresh times by over 90%.

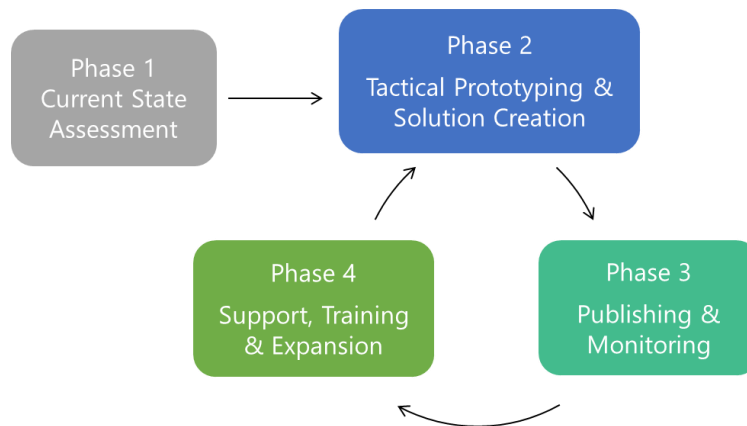


Image Credit: Microsoft

Figure 5. Self-Service BI Approach

Power BI Sample Visualization

Below is a sample visualization (see Figure 6) which illustrates how Power BI is used to quickly call attention to problem areas. The gauge at lower left shows that the overall fill rate is 71%, however, there is wide variation when fill rate is broken out by channel, platform, and partner, as shown in the horizontal bar graphs. Lower fill rates are colored red to draw attention. The volume tracker (lower right) shows daily ad impressions, which draws attention to sudden drops which could indicate a problem. The dashboard is interactive, so if the user clicks, say, on Channel 1, all the other visuals will automatically filter to show only Channel 1 data, which helps to drill down on a specific area.

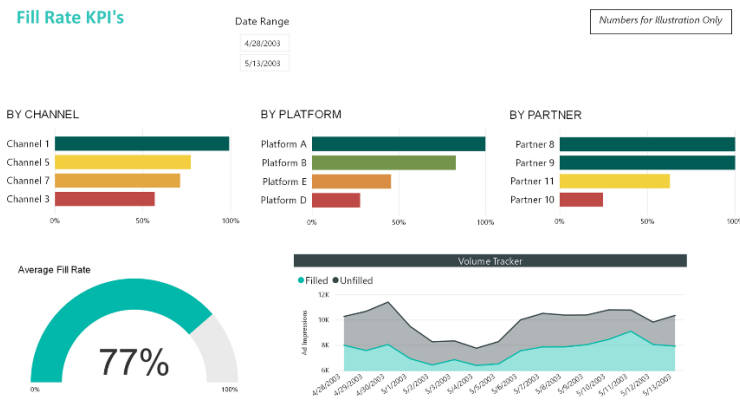


Figure 6. Power BI Sample Visualization

Maximizing the Data-Ink Ratio

The visualization above was designed so as to maximize the amount of space dedicated to conveying the data, i.e. the portions of the visual that change as the data changes, as defined by Tufte (2001):

$$\text{Data-ink ratio} = \frac{\text{data-ink}}{\text{total ink used to print the graphic}} \quad (2)$$

= proportion of a graphic's ink devoted to the non-redundant display of data-information

= 1.0 – proportion of a graphic that can be erased without loss of data-information

The data-ink ratio of the above visual is approximately 0.80.

Performance Management Process

The Power BI visualizations are integrated into operational decision-making processes. For example, if fill rate KPI is low for a given platform, managers may re-allocate client ad campaigns to that platform and/or try and persuade clients to buy more of it. If a given platform is always full, it may be oversold. If so, managers may reserve it for key client campaigns and/or bundle it with other platforms. If there is a sudden drop in fill rate or ad impressions, it could indicate an ad integration issue with a particular platform or partner. In these and other cases, visualization provides faster detection and response to problems.

Key Learnings

The system has resulted in increased focus on harder-to-fill platforms, which tend to be newer ones that clients may not be used to buying. Conversely, desktop is in relatively high demand because of its advanced measurement capabilities, and using it selectively leads to better platform diversification and increased ability to accept last-minute demand. Overall, the solution helps to shift emphasis from short term client campaign fulfillment to long term yield optimization of the entire advertising space.

Business Value of Power BI Analytics Solution

The solution has delivered significant value to stakeholders at multiple levels. Senior management now has a timely, objective view of fill rate across the entire business, and relies less on anecdotes. Middle managers are alerted earlier to problems, and can pinpoint exactly which buckets of advertising space have low fill rates. Front-line staff feels more empowered, as they have less manual work to do, and Power BI has given them an opportunity for creative work in designing their own visualizations. They are eager to enhance the system further.

Conclusion

This paper has focused on digital video publishers who depend on advertising as a source of revenue, and how changing consumer viewing habits create fragmented advertising space and operational challenges to fill it with client ad campaigns. The principal KPI is fill rate, which can be sliced by channel, platform, and partner. Microsoft Power BI offered several advantages, including increased user comfort level, cloud architecture, and a self-service BI approach. The solution provides operations managers with timely monitoring and raises awareness of issues, and also inspires creativity by front-line analysts, since they can adapt the solution themselves to changing business conditions. The solution's holistic view of advertising space has provided key insights, such as focusing sales efforts on filling newer platforms and shifting culture towards overall yield optimization.

Next Steps

Future enhancements include Predictive Analytics to project fill rate into the future, which will be helpful in assessing the likelihood of quarterly targets being achieved. Additionally, Prescriptive Analytics can help define what default actions should be taken in response to common exception conditions, and how demand should be proactively allocated to the various buckets of advertising space.

Bibliography

- Few, S. (2003). *Information dashboard design*. Burlingame, CA: Analytics Press
- Guillen, J. & Coates, M. (2016). Power BI governance and deployment approaches. Retrieved from: <http://go.microsoft.com/fwlink/?LinkId=785915&clcid=0x409>
- Lachev, T. (2018). *Applied Microsoft Power BI: Bring your data to life!* Prologika Press.
- Perrin, N. (2017). eMarketer lowers US TV ad spend estimate as cord-cutting accelerates. Retrieved from: <https://www.emarketer.com/Article/eMarketer-Lowers-US-TV-Ad-Spend-Estimate-Cord-Cutting-Accelerates/1016463>
- Tufte, E. (2001). *The visual display of quantitative information*. Cheshire, CT: Graphics Press
- Wang, Y., Scher, J., Yao, X., Rosenblum, S., & Dogaru, M. (2017). Freewheel Q1 2017 video monetization report. Retrieved from: <http://freewheel.tv/Insights/#video-monetization-report>

Author's Information



Joseph CHOMSKI, MSc student, Boston University, Department of Administrative Sciences. jchomski@bu.edu

Major Fields of Scientific Research: Business analytics, visualizations, business intelligence, performance measurement.

AN ONLINE PLATFORM FOR PROJECT BASED LEARNING - A PROPOSAL

Yuting Zhang

Abstract: *Project Based Learning (PBL) has become a widely used teaching method in K-12 education and spread to higher education. It has been used in many technical fields, particularly in Engineering, also increasingly now in Computer Science. PBL is a student centered pedagogy that integrates knowing and doing. Through real life projects,, students work as teams to form project ideas, design projects and implement projects. Research has shown that PBL can motivate learning, enhance learning quality, and help students learn not only technical skills but also soft skills .*

While PBL has a number of benefits, it is also quite challenging to implement PBL effectively. In this paper, we propose an online platform to prompt the implementation of authentic PBL at the program level across different courses. The purpose of this online platform is to build a community for students to find, share, implement and discuss projects, as well as facilitate better communication between students and faculty members across different courses.

Keywords: *Project Based Learning, Online Platform, Computer Science Programs*

ACM Classification Keywords: *A.0 General Literature - Conference proceedings*

Introduction

There is a famous proverb by Confucius: “*I hear and I forget. I see and I remember. I do and I understand*”. Students learn better through doing than hearing. However, in the traditional teaching method, instructors spend most of their time in preparing and delivering in-class lectures. Students work on their to-do list after the class, which may include homework, labs, projects, quizzes, and exams. Quizzes and exams are mainly used for assessment rather than learning itself. Though labs and assignments can help students learn domain specific knowledge, these knowledge may be forgotten by students and outdated quickly. More important, students need learn practical skills and further thinking models. These practical skills include computer skills, programming skills, problem solving skills, and soft skills such as communication, collaboration, presentation, management skills, etc. Furthermore, computational thinking, critical thinking, system thinking, scientific thinking should be developed, particularly in their graduate study. Simple assignments and labs are usually ineffective to help students learn in these areas. Rather, projects are better means through complex tasks based on challenging problems to involve students in design, problem solving and decision making, etc. Through doing projects, student can learn better.

Project Based Learning (PBL) is a student centered pedagogy that integrates knowing and doing [Thomas Markham, 2011]. It is a style of active and inquiry-based learning. The idea of using projects in education

can be dated as earlier as 16th century in architecture and engineering education in Italy. The United States pioneers such as John Dewey had advocated learning from projects rather than from isolated problems. Nowadays, PBL has become a very popular teaching method in K-12 education and spread to higher education. It has been used in many technical fields, particularly in Engineering, also now increasingly in Computer Science. Students usually work as teams to form project ideas, design projects and implement projects. Through real world project, students learn new knowledge and skills. Research has shown that PBL can motivate learning, enhance learning quality, and help students learn not only technical skills but also soft skills [Thomas, 2000].

While PBL has a number of benefits, it is also quite challenging to implement PBL effectively. In this paper, we propose an online platform to prompt the implementation of authentic PBL at the program level across different courses. The purpose of this online platform is to build a community for students to find, share, implement and discuss projects, as well as facilitate better communication between students and faculty members in different courses.

Challenges of Implementing PBL

Currently there is no standard model or theory to describe PBL. Synteta made a good synthesis of the features described in the literature in his master thesis as following [Synteta, 2001]:

- are central to curriculum,
- long-term (more than a couple of class days and up to semester),
- Interdisciplinary,
- have a driving question that is challenging and constructive,
- are student-centered and
- are based on collaborative or cooperative group learning,
- are integrated with real world issues and practices,
- have productive outcomes,
- have an impact on “life skills” like self-management, group process, and problem-solving skills,
- and use cognitive tools, usually technology-based

Our department (CS department at Boston University Metropolitan College) offers a number of courses with project components, some as semester long projects, others as final projects. For example, both CS673 (Software Engineering) and CS683 (Mobile Application Development with Android) courses that I am teaching feature semester long projects. CS673 features a semester long group project, consisting of 6-8 students developing a comprehensive real-world software project. CS683 also features a semester long project to develop a mobile app. In both courses, projects are counted about 50% or more in the final grade. Students have very positive feedback about the project components. Many students think that the project is the best part of the course.

While we offer a number of courses that have project components, the question is whether we can leverage PBL at the program level to enhance the quality of our programs? Given the above features and benefits, I think it will enhance our program and distinguish us from others if we can effectively implement PBL.

However, there are a number of challenges of effectively implementing authentic PBL at the program level across different courses.

- The design of projects central to curriculum: this is the key of PBL and also the most challenging task. The project is not just used as an example or a complementary exercise or an assessment mean as in most of our courses, but is the main teaching and learning strategy. The project should be well designed to drive students to learn the central concepts and principles. A variety of scaffolding content also need be provided to help students become proficient at conducting inquiry activities.

- The origin (generation) of projects: lots of students have difficulties to generate proper project ideas. Currently most software development projects in our courses are students self-defined projects and usually discarded after the semester ends. It is challenging to find real world clients to use applications after completion.
- The complexity and length of projects: it is very challenging to implement group based projects, particularly in online courses. Most of our course projects are individual projects. CS673 does feature large group projects of 5-8 students. Communication and management are always one of the most challenging aspects. Moreover, all course projects are within a single semester. Most students can barely finish basic features without rigorous evaluation or testing within a semester length.
- The isolation of projects: While there are a lot of overlap among projects in different courses, there are no effective way to promote collaboration. Projects are mostly isolated from each other. Faculty and students are not aware of projects in other classes even they may be related or similar. Projects are also constrained in a single course. This makes hard to implement big scale projects which usually involve inter-discipline and multiple domain knowledge. For example, a group of students in CS673 proposed to implement web based job analysis application. However, due to lack of time and expertise, the application only implements basic interface and very simple statistics analysis. If we can have a joint project with the data mining course and expand it to multiple semesters, a much more comprehensive application may be developed to be used in the real world.
- The feedback and advice of projects: Direct guidance, constant and customized feedback to projects from instructor is also a critical success factor. General graders or even facilitators may not have enough expertise to provide such advices. This will generate heavy workload, for instructors, particularly of large classes.
- The evaluation of projects: as projects are mostly open ended without predefined specific outcome, it is very challenging to provide fair and subjective assessment. It usually requires varies and frequent assessments, including instructor assessment, peer assessment, and self assessment. To maintain a uniform standard, it is also very important but challenging to develop a guideline of the evaluation process and rubric.

For a PBL central curriculum, we have to balance the breadth and depth. Sometime, a single project may go in depth in some particular topic, but lack of breadth. In this case, we may need integrate multiple projects, and with other components such as lab exercises with clear instruction, self test questions, written problem solving homework, etc.

The Proposed Online Platform - AllAboutProjects

This paper does not attempt to address all challenges listed above. Instead, we would like to propose an online platform AllAboutProjects to support the implementation of authentic PBL at the program level across different courses. This purpose of this online platform is to create a community for students to find, share, implement and discuss projects, as well as facilitate better communication between students and faculty members in different courses. This platform will help prompt our program and assist the program assessment as well.

The most important entity of this application is **Project**. Each project should have a unique id, a title and a brief description. In addition, a project can be associated with a number of keywords or tags used for categorization and search. The programming languages, frameworks, and any particular skill sets used in the projects are also specified. Each project is related to certain courses or programs in our college. Additional details about a project may be provided through a number attachments or links such as github links of the project source code, website links of the hosted web application, youtube links of the project presentations,

etc. A project may also receive attentions and comments from other users. Other quantifiable features such as the difficulty, the complexity, and the effort may also be collected.

Each project is associated with a number of **Users**. There are several types of users:

- Clients: any individual or organization who want to find students to work on a particular software / IT / data analysis project.
- Implementers: students who participate in the design and implementation of the project. There will be different roles for students.
- Advisors: faculty members who supervise the project and provide the feedback.
- Viewers: general users who simply view the project information and may provide some comments (only logged-in users can make comments)

The following components are proposed as part of this online platform:

Project Proposal Submission: Any faculty members, students, individuals, and organizations can submit project proposals through this platform. Without logging into the system, anyone can fill out a project proposal form, providing basic idea and requirements about the project, and the contact information. When a proposal is submitted, a faculty member who is teaching related courses or expertised in the related area will review the proposal and decide if the proposal is suitable for their courses or research. This can facilitate the generation of real world project ideas, and acquire more real customers in our project centered program.

Categorized project portal: Students or faculty members can publish completed and ongoing project information to share with others through this platform. It will be categorized based on the course numbers, programs, subjects, platforms, skill sets, etc. It will be searchable so that other people can easily find interesting projects and also post comments. This can help students and faculty members get to know projects in other courses, and also potentially find the collaboration opportunities.

To control the quality of the projects and prompt our program, all projects will have one or multiple faculty members as advisors. Only an advisor can create and delete projects, but all contributors can edit and modify project information.

Anyone can search the projects. This can not only help existing students, but also prospective students know more about our programs and courses. This is also a great resource to be used in marketing. However, for better privacy and security, only logged-in user can see more details and the contact information of project members, as well as make comments. The project members should also be able to configure which information to be made public.

An Online Discussion Board: The purpose of this platform is to create a community in our college to share and discuss project ideas and prompt collaboration. An online discussion board should be included to facilitate the discussion among faculty members and students to share the project ideas, technical tips and find possible collaborators.

Project Management Tool Suite: For any real world project that has certain complexity and involves multiple collaborators, project management is a very important aspect. Effective project management tools can help improve the project quality and success rate. This should included as part of this platform as well.

Project Support and Resources: Through this platform, we would also like to provide additional resources to help both faculty members design project centered curriculum and help students implement projects. A number of guidelines, standards, and tutorials will be collected and provided through this platform. It serves as a central point and a starting point for anyone who wants to start a project. The following list is just an exemplary list that we should develop and publish on this platform to support PBL.

- Guidelines on Project Centered Curriculum Design: This guideline may define the percentage of the project works in a program/ course, project duration, project complexity, project evaluation criteria, project requirements, learning objectives, prerequisites, context,
- Project document templates: A number of templates can be provided to students. For example, project proposal template, project management plan template, project midterm report template, project final report template, research paper template, project design and project testing document templates for a

software development project. In addition, APA and IEEE reference guidelines and bibliography template shall also be provided.

- Coding standards and other standards: for coding project, coding standards for different programming languages can be collected here for references.
- Tutorials: A number of tutorials can be provided to prepare students for their projects. This may include tutorials on Linux operating systems, tutorials on virtual machines systems, tutorials on git usage, tutorials on UML diagrams. We may also include basic tutorials on various programming languages and frameworks such as R, python. Moreover, tutorials on how to do literature research and how to write technical papers may also be provided. Since there are already tons of online resources on Internet. Instead of developing all these tutorials, we may provide just links to existing resource that we have collected and filtered.

In summary, with above features, this platform can help address the challenges listed in the previous section:

- A number of guidelines and standards will be developed and shared through this platform to help instructors develop project centered curriculum. Common project resources are provided as “scaffolding” materials to help student in their project based learning.
- The project proposal component can help generate real world project ideas and acquire real clients.
- Project portal and online discussion board break the project isolation and facilitate more communication, sharing and collaboration. It also helps students get feedback and advice from other people.
- Project management tools help manage complex group projects, making the implementation of PBL in online courses more possible.

The implementation of the platform

To make best use of the proposed platform, it may be integrated with our current websites and the blackboard system. This will give users seamless experience. Ideally, the authentication should be done through BU shibboleth, so that users can log into the system using their BU credentials. This will help us better control the access rights. Proper access control is very important to ensure the security and privacy of this platform. The following list show some examples of basic access control rules:

- Anyone can submit a project proposal through the platform.
- Anyone can search and view public information of public projects.
- Only a BU account can register on this platform.
- Only logged-in users can make comments to a project
- Only advisors can create/delete a project
- Only advisors and implementers can edit a project

Since this platform consists of several components, it is quite complex to build all of them from the scratch. Currently there are already a number of good tools available for use in various projects. For example, code repository such as Github and bitbucket are necessary tools for code sharing. Google doc can be used for documentation collaboration. Slack can be used for team communication. We may seek the solution to integrate the existing tools into our platform.

The highest priority item in the to-do list is the project portal where projects can be published and searchable. In CS673 S17 (Software Engineer Spring 2017) class, a group of students developed a prototype of the project portal using the MEAN stack. This prototype implements some basic functionalities such as basic login, create and delete projects, and basic searches. In CS683, we also implement a simple version of project portal as an Android application. While these prototypes may not be in production quality and used directly, they do give us better ideas about what and how to implement this component. Basic operations such as creation, modification, and deletion of projects are easy to implement. The user authentication and access control are more challenging, particularly if we want to integrate with BU login credentials. Also sustainable web hosting is another key factor that we will need help from the IT department.

On the other hand, we will need the collaboration of all faculty members to discuss and develop contents in the project support and resource component, including both program dependent and program independent materials.

In the previous CS673 classes, we have also developed a project management tool suite. With more testing and refining, this may be used directly. Alternatively, we may seek the integration of other existing project management tools and online discussion board application.

Conclusion

In summary, PBL is a student centered pedagogy that integrates knowing and doing. Research has shown that it has a number of benefits to promote active learning. We can leverage PBL to enhance the quality of our degree programs. However, there are a number of challenges to implement PBL at the program level. An online platform is thus proposed to address some of these challenges and support the implementation of PBL.

Bibliography

- [Markham, 2011]. Markham, T. . Project Based Learning. *Teacher Librarian*, 39(2), 38-42.
- [Synteta, P., 2001]. Synteta, P. . Design and Development of a Scaffolding Environment For Students Projects. 2001. Master thesis, University of Geneva, Geneva, Switzerland.
- [Thomas, 2000]. Thomas J. W. A review of research on project-based learning. 2000.
http://www.bobpearlman.org/BestPractices/PBL_Research.pdf
- [Fincher & Petre, 1998]. Fincher, S., & Petre, M., Project-based learning practices in computer science education. In proceedings of: IEEE Frontiers in Education Conference, Tempe, Arizona, USA, November, 1185-1191.

Authors' Information

Yuting ZHANG, Ph.D. Assistant Professor,
Computer Science Department, Boston University
Metropolitan College, danazh@bu.edu.
Major Fields of Scientific Research: system,
security, software engineering

STRIKING THE BALANCE: TEACHING DATA MINING WITH THE RIGHT MIXTURE OF DEPTH AND BREADTH

Gregory PAGE(1), Slav ANGELOV(2), Penko IVANOV(2), Vladimir ZLATEV(1)

***Abstract:** The purpose of this paper is to describe the methodology that was used to develop and teach AD699: Data Mining for Business Analytics at Boston University’s Metropolitan College in Spring 2018 (on campus) and in Summer I 2018 (online). Data mining for business analytics is a field that sits at the intersection of statistics, computer science, and business-specific domain knowledge.*

The goal in teaching the course was to prepare students for a “full-spectrum” approach to data mining. That is to say, graduates of AD699 should be able not only to interpret and assess the results of the data mining process, but also to build the models, know the underlying functions, and alter the code in the R programming environment as needed to adjust the models. That said, this was not a Computer Science course, and the students were not coming from a programming background -- chief among the challenges associated with delivering this course was the question of how much emphasis to place on proper syntax.

Where the course developers had to choose between breadth and depth, they opted for breadth; however, they remained careful at each step of the way to ensure that the course material remained substantive and challenging. By giving students a broad basis of exposure to topics such as data exploration and initial analysis, performance assessment, and implementation of both supervised and unsupervised learning models, the intention was to prepare them to have a reasonable level of fluency in whatever specific topic an employer wished them to drill down on at a deeper level in the workforce.

From a pedagogical perspective, the course developers took a wide-ranging approach. Among the teaching methods employed were traditional lectures, video-based tutorials, technology-enabled tutorials (such as the AD699-specific lessons built using the Swirl platform, and accessible through the BU Virtual Lab environment), assigned readings, individual assignments, online discussion boards, and a culminating exercise that required students to work in teams to use a real-world dataset based on Airbnb rentals to build and apply data mining models.

Keywords: data science, data mining, R language, data science education.

ACM Classification Keywords: Computer science education, Information systems education, Model curricula.

(1) Boston University Metropolitan College, Department of Administrative Sciences, Boston USA

(2) New Bulgarian University, Department Information Technology, Sofia Bulgaria

Introduction

Our group consisted of four members: Professor Vladimir Zlatev, a Professor at Boston University, Metropolitan College; Slav Angelov, a Ph.D. student at New Bulgarian University; Penko Ivanov, a full-time data analyst with GFK, a European-based company, as well as a part-time Ph.D. student at New Bulgarian University; and Greg Page, a Lecturer in Administrative Sciences at Boston University.

Professor Zlatev is an Associate Professor of the Practice of Administrative Sciences at BU, and the supervisor of the Applied Business Analytics program. He conceived the idea of introducing a rigorous, hands-on data mining course within the Administrative Science Department at MET several years prior to the launch of AD699. He also coordinated and led the overall effort as this team built AD699 from scratch during the winter months of 2017-18.

Mr. Angelov is an expert in statistics and regression models. He is also quite familiar with several machine learning algorithms. He developed slides for more than half of the in-class lectures, in addition to writing scenarios for in-class activities designed to give students hands-on exposure to the related concepts. For each of the class sessions that involved the use of any particular machine learning algorithm, he wrote two original use-cases: first, one in which the students were asked to simply enter a series of commands into their R consoles as they read the detailed notes that he included to explain the purpose of each command; second, he gave the students an original problem that required them to apply the things that they had just seen in a novel way, with a brand-new problem set.¹

Mr. Ivanov is highly accomplished both in industry and in academia. He works on large-scale data problems full-time for a multinational firm based in Europe. He is also very close to receiving his Ph.D., and has extensive teaching experience both in Bulgaria and in the United States, including several semesters of experience as a course facilitator at BU. A skilled programmer, he developed a series of lessons specifically for AD699 using a platform called Swirl, which offers in-console instruction in R.²

Lastly, Mr. Page is a generalist who was new to R and to many of the specific concepts covered in the course. However, he came to this project with an extensive teaching background, including teaching Computer Science at BU and three other schools, as well as tutoring in statistics and mathematics. He also has more than ten years of experience as a Military Intelligence Officer in the United States Army.

As a group, we met via semi-monthly Zoom sessions throughout November and December of 2017, and January through March of 2018, as coordinated and led by Professor Zlatev. The frequency of these meetings was effective, as two-week intervals enabled each of us to complete the various portions of the lectures before reconvening, reviewing, and moving to other content. At the conclusion of each of the meetings, we summarized the major points that we had covered during that particular meeting, and also outlined our deliverables for the following meeting.

The complementary skill sets of the group members proved to be a huge help in terms of developing and clarifying the final product. Among the members of the group, Professor Zlatev brought the most teaching experience to the table, and was therefore instrumental in determining what items were needed for this first course iteration. Mr. Ivanov brought the most familiarity with R, and with the deployment and use of data mining models. Mr. Angelov brought the most knowledge of regression, and was therefore instrumental in developing the content related to the task of continuous numerical variable prediction. Finally, Mr. Page, the course instructor, kept the content and discussion focused on answering the questions of “Will

¹ R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

² Sean Kross, Nick Carchedi, Bill Bauer and Gina Grdina (2017). swirl: Learn R, in R. R package version 2.4.3. <https://CRAN.R-project.org/package=swirl>

this work in the classroom?” and “Is this the right approach for our student population?” The various backgrounds, knowledge areas, and perspectives that each member brought to the team formed a huge component of the team’s success in forming the course.

Answering the “Depth vs. Breadth” Question

As AD699 began to take shape, one of the central questions that we wrestled with was that of depth vs. breadth. Of course, this was not a question that lent itself to an easy answer. Any one of the statistical methods covered in the course could have easily constituted an entire semester’s worth of material, if explored fully enough. At the same time, we wanted the students to see the “wow” factor that comes with learning about a new data mining algorithm and its applicability to solving real world problems. If we spent the entire semester on regression, for example, we would not have been able to show or explain how Pandora uses a proximity model similar to k-nearest neighbors to determine what song a particular user will hear. If we focused too much on proximity models, however, our students would not have become acquainted with the connection between naive bayes and spam filtration.

If we shot too far in the direction of breadth, however, the course could have become an overwhelming blizzard of terms, concepts, algorithms, and coding instructions that left students frustrated and unsure about what they had accomplished in the course’s fourteen weeks.

Ultimately, we aimed for a middle ground solution tilted in the direction of breadth, primarily for three reasons:

- (1) The student audience was mostly preparing themselves for business careers in management;
- (2) A wide exposure to many concepts would provide a “jumping off” point for any who wished to deep dive further into the topics on their own afterwards, and
- (3) Continual exposure to new ideas and concepts was the best way to maintain students’ interest throughout the semester.

Since the Administrative Science Department offerings are mainly business-oriented, and the students in the program are mainly focused on preparing for careers in management, it seemed fitting to deliver the content with the idea that these students would be more likely to be managing a department that included data scientists, rather than coding the models themselves. As such, a course such as this should envision its students as future generalist managers. To do, course graduates should be able to supervise a variety of potential data science projects, for which they need to have a general familiarity with several types of models and algorithms, but do not need detailed knowledge in all of these areas.

The second reason mentioned above is that the course can provide a launching pad for students who wish to delve deeper into the topics independently, after the conclusion of the course. If we can expose a student to more than a dozen different concepts, he or she might develop a couple of favorites during the term from among the overall group. A student interested in consumer segmentation and interest-based marketing, for example, might be particularly fascinated by clustering. Someone determined to begin a career on Wall Street might become hooked on predictive techniques such as multiple linear regression that can aim to predict the movement of particular securities or indexes. In either case, that student could receive the “spark” from AD699 and then carry it forward on his own, after the course’s conclusion, in order to develop true expertise. Others still may see the phenomenal data visualization capabilities offered by R, and decide to explore that field in greater depth in their professional lives. Because the students’ personal and professional interests are widely varied, we can’t choose just one or two of these concepts and limit the entire scope of the course to just that narrow slice of the data mining world. By casting a wide net, so to speak, we can reasonably and realistically aim to connect with each of our students’ fields of interest at least once or twice throughout the term.

With 14 scheduled class meetings throughout the semester, we were able to achieve a pace that meant one major topic (such as a new concept, or a new algorithm) would be introduced in each class session. In order to maintain strong student comprehension of previously-covered topics, our instructor began each class with an “opener” -- this was a review question based on material that the students had already seen. The course also included three quizzes, each of which covered either three or four class sessions’ worth of material. Homework assignments also helped with concept reinforcement (more about the assignments and assessments will be covered in a subsequent section).

The final project for the course also offered students a chance to revisit concepts from throughout the semester in a way that allowed them a considerable degree of leeway with respect to algorithm selection, variable selection, and model selection. For this project, students analyzed a real-world dataset that came from Airbnb. Students were divided into teams, with each team focusing on one of the six cities featured in the dataset. Together, they performed several exploration and visualization tasks (for the visualization and summary statistics portion, they had free reign to choose from among any of the methods that they had seen in the course, or even that they had explored independently). They also built a prediction model using a host’s review score as the outcome variable. They performed one neighborhood-based clustering task, and two classification tasks -- one that required the use of the k-nearest neighbors algorithm in order to determine whether a particular type of host would likely implement a cleaning fee, and another that asked students to bin the Airbnb rentals in different price groups, and then classify a particular apartment based on its likely group. For the price classification task, students were given the choice of using either naive Bayes or a classification tree.³

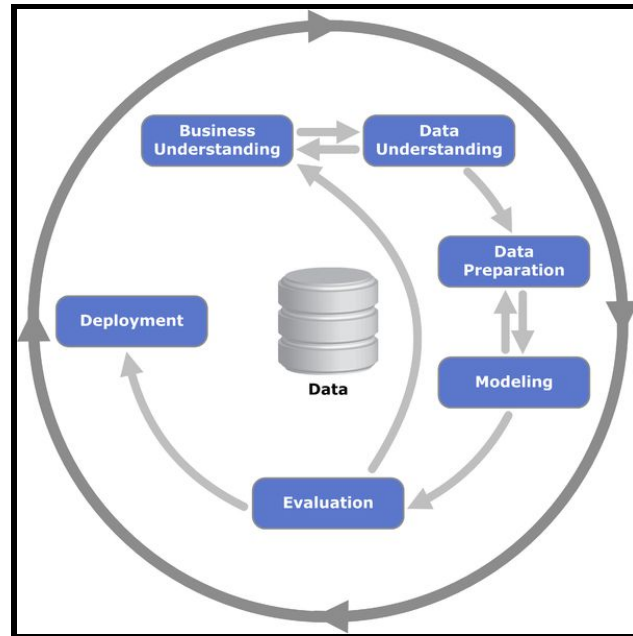
The Airbnb project included two deliverables -- a written report and a final presentation. The final products that we received were quite impressive -- they showed us that the student groups were able to synthesize very large data sets and apply the algorithms seen in AD699 in the appropriate situations. As we reviewed the project submissions from our students, taking note of the wide range of data mining tools that they employed, it became clear to us that emphasizing breadth was the right move for this course.

The Data Mining Process

To introduce the concept of data mining, we mainly follow the CRISP-DM model.⁴ CRISP-DM stands for Cross-Industry Standard Practices for Data Mining. As depicted in the figure below, it involves six main phases: Business understanding; Data Understanding; Data Preparation; Modeling; Evaluation; and Deployment.

³ Terry Therneau and Beth Atkinson (2018). *rpart*: Recursive Partitioning and Regression Trees. R package version 4.1-13. <https://CRAN.R-project.org/package=rpart>

⁴ [Kelleher, John D. and Brendan Tierney, 2018]. Kelleher, John D. and Brendan Tierney. *Data Science*. Massachusetts Institute of Technology, 2018.



One of the most important things to emphasize about the CRISP-DM model is the circular pattern that surrounds it. Constant iteration is a vital aspect of any good data mining model.

Image Source: Wikimedia Commons.

With our students, we emphasized the importance of the circular arrows along the periphery of the diagram. A data mining model could be excellent at the exact moment in which it was created; however, as the circumstances surrounding the model change, so must the model. For example, if a spam filtration model could identify spam-associated words in e-mails with nearly perfect accuracy, it would soon be obsolete -- once spammers figured out some aspects of the model, they could change the nature of spam messages in order to bypass this filter. Likewise, a model built to solve one of the classic consumer behavior questions, such as whether a mobile subscriber will move to another service, or whether a bank consumer will accept a loan offer, is only as good as the period in which it was built. As consumer tastes and habits shift over time, so must a model built around predicting their actions.

On that same note, we always sought to remind students that we would emphasize data mining that served a particular business purpose. With each concept that was introduced in the course, we focused on how a business could draw value from its use. For example, how could a company use k-nearest neighbors to determine which people were likely buyers of its products, and which were not? How could business travelers use a naive Bayes algorithm to estimate the likelihood that their flight would be delayed? How could a supermarket most effectively take advantage of a pattern it learned about purchase patterns and non-intuitive associations between types of goods?

However, big data projects place new challenges that are not fully considered by existing methodologies such as CRISP-DM. Consequently, we also introduced our students to some other big picture concepts to data modeling and to model deployment. In the final lecture of the course, we mentioned MapReduce and the Hadoop Distributed File System (HDFS). Although we did not get into any meaningful detail on this, it did offer the students some exposure to the way in which a massive, open-source system can be used along with parallel processing to handle data that is not necessarily obtained a "neat", flat

format that students are used to seeing in Excel spreadsheets and in dataframes in R. This final lecture also talked about proprietary systems such as Microsoft's SQL server and Azure offerings.

The Statistical Methods Selected for the Course

The class started with a unit on data exploration and data visualization. At the outset, we wanted students to quickly become comfortable with using R to derive summary statistics from a dataset, and to create basic visualizations. As predicted, that led to some questions along the lines of "What are we really doing here that we can't just do in Excel?" After cycling through the range of visualizations from the base package, however, we moved right into ggplot.⁵ A package that enables a wide range of impressive, multi-dimensional visualizations, ggplot opened our students' eyes to the rich graphing opportunities offered by R. Mastering most of the other material taught in AD699 requires detailed statistical knowledge, but mastery of ggplot is more a matter of simply learning the capabilities of the package, and how to use them. Developing ggplot familiarity is also a great example of a "quick win" that an AD699 student can achieve in order to have an immediate impact at his or her place of work.

After moving through visualization, exploration, and some fundamental terminology and concepts, we introduced linear regression. Starting with simple linear regression, we emphasized the business value of a model that can predict a continuous numerical outcome from a single input variable. We also covered the derivation of the best-fit line, interpretation of the regression equation, and analysis of the model. This background set the stage for multiple linear regression, which brought the additional complexity of requiring analysis of the various variable-to-variable relationships in order to build the best model possible, as well as an even larger range of diagnostic tools for examining the results and comparing various models.

Classification tasks covered in the course included k-nearest neighbors, naive Bayes, logistic regression, and classification trees. Along with k-nearest neighbors, we introduced three types of distance metrics used to determine the proximity of records in a dataset -- Euclidean distance, correlation distance, and hamming distance. Students applied classification methods independently as part of course homework assignments. For example, students used a naive Bayes model to analyze on-time and delayed flights, they used logistic regression to determine the factors that most influenced whether members of a certain book club would buy a particular title, and they used classification trees to analyze bank customers as being likely or unlikely to accept a loan offering.

Unsupervised learning topics covered in the course included clustering (both k-means and hierarchical), data visualization, and association rules. For association rules, we used the Groceries transactional database from the arules package in R.^{6,7} For clustering, we began with a k-means analysis of iris flower species, and then used a hierarchical clustering example involving utility companies, taken from the course textbook (*Data Mining for Business Analytics*, by Shmueli, Bruce,

⁵ H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

⁶ Michael Hahsler, Christian Buchta, Bettina Gruen and Kurt Hornik (2018). arules: Mining Association Rules and Frequent Itemsets. R package version 1.6-1. <https://CRAN.R-project.org/package=arules>

⁷ Michael Hahsler, Sudheer Chelluboina, Kurt Hornik, and Christian Buchta (2011), The arules R-package ecosystem: Analyzing interesting patterns from large transaction datasets. *Journal of Machine Learning Research*, 12:1977--1981. URL: <http://jmlr.csail.mit.edu/papers/v12/hahsler11a.html>.

⁸ Michael Hahsler, Bettina Gruen and Kurt Hornik (2005), arules - Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software* 14/15. URL <http://dx.doi.org/10.18637/jss.v014.i15>.

Yahev, Patel, and Lichtendahl).⁹ Students learned how to interpret dendrograms, as well as how to measure things such as in-cluster homogeneity and average weighted distance from within a cluster to records placed in other clusters.

For the most part, the conceptual basis for the concepts and algorithms noted here came from the course textbook. After hearing about the concepts from their instructor, the students were given access to PowerPoint slide decks prepared by Mr. Angelov. The format of these slide decks was as follows: first, they included a very high-level overview of a particular algorithm, followed by a series of slides that walked students through fully-coded example in R. Then, they included a challenge to the students along the lines of “Now that you’ve seen how to do this, can you build your own solution to the following question?” Finally, a sample solution to the challenge question, as built by Mr. Angelov, was shared with the students.

Along the way, students in our course received a solid grounding in the process of machine learning. From an early point in the course, they learned about the need for data partitions, and were exposed to several sets of functions that can be used in R to accomplish a data partition.

Handling Queries and Presenting Programming Material to Non-Technical Students

Working with a non-technical student audiences presents a unique set of challenges. For example, a Computer Science concept as basic as instantiating a variable before using it may not be obvious to an Administrative Sciences student; therefore, an instructor or facilitator who says “Use the attached code snippet” or “follow the example on page 174 in the textbook” cannot necessarily assume that the student will understand that the code cannot be implemented if run *literally* off the example without first importing certain libraries or instantiating certain variables.

One of the major ways that we sought to ease the transition for students was by offering an optional lab session each week to students who sought extra help with assignments or with general understanding of R syntax and course concepts. By identifying a regularly-held two-hour “drop-in” window for students, we aimed to ease the anxieties of some members of the course who felt intimidated by the task of interpreting or writing code. These sessions were sparsely attended at first, but the number of student visitors steadily increased throughout the term.

The Swirl lessons prepared by Mr. Ivanov were an enormous help in terms of acclimating the students to R. Swirl is an R instruction package that is deployed in the R console, and is completely compatible with RStudio. It begins with a series of prompts that ask students to complete various tasks with R; as students progress, the tasks become more complex. Mr. Ivanov built a series of AD699-specific R tutorials that simultaneously gave students exposure to the R environment and to the major concepts covered in the course.

Specifically, Swirl helped introduce students to the concept of disparate types of data, such as characters, integers, logical values, etc. Data types are another concept that does not necessarily register automatically with students who have not had any prior Computer Science education. For example, take a declaration such as “A matrix must contain data all of the same type, but a data frame can contain multiple types of data. The individual vectors within a data frame, however, must all be of the same type. If they are not originally inputted that way, type coercion will force the uniformity within each particular vector.” To someone who has taken an undergraduate computer science course, the previous quotation should make sense without any additional explanation or context; for someone who is completely new to the field, however, it could be a dizzying blur of jargon that requires several layers of unpacking. Swirl lessons that exposed students to such concepts in a hands-on way helped to bridge this divide.

One way that we will ease the transition into R in the future is by requiring students who enroll in AD699 to have previously completed ADR100: Introduction to R for Business, a free laboratory offered by BU and built on the edX platform.

⁹ Shmueli, Bruce, Yahav, et. al. *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. Wiley: Hoboken, NJ, 2018.

This introductory course is designed to de-mystify R -- it gets students to become familiar with the RStudio interface, and with using several important built-in R functions for data exploration and visualization. It also touches upon two themes from data mining -- regression and time-series analysis, but without going into considerable depth.¹⁰ Students who complete ADR100 should come away with a baseline of confidence in their abilities to use R and RStudio, in addition to a general sense of the language's functional uses.

Assessing Learning Outcomes

Homework assignments challenged students to explore these topics even further on their own. The assignments were all crafted so that no two students' submissions would be exactly alike. For example, an early assignment on data exploration and data visualization was based on a dataset from Major League Baseball. Along with the prompt for this assignment, the instructor posted a list of individual years, with each student asked to only analyze data from his or her particular year. Students were not prohibited from collaborating with one another; in fact, the instructor encouraged them to work together. However, wholesale copying was rendered impossible by this unique assignment system.

In other cases, students were asked to create something uniquely original and then assess it using one of the algorithms learned in the course. For example, one homework assignment asked students to design a unique type of candy bar, with binary attributes assigned to categories such as chocolate or peanut butter flavoring, and a continuous numerical score used to indicate sugariness. The students then ran the algorithm against a series of other candies that had been labeled as "winners" or "losers" based on an existing dataset that contained the results of head-to-head candy matchups. Again, the requirement that students generate something uniquely theirs, and then assess it, mitigated the copying/plagiarizing risk that might have tempted students who were overwhelmed by the challenges of the course and in search of an easy escape.

Another assessment method came in the form of discussion board posts that the students wrote in teams. In the second week of the course, the students were split into six teams: ALPHA, BRAVO, CHARLIE, DELTA, ECHO, and FOXTROT. Every other week during the term, each team was responsible for posting an answer to one problem from the textbook, and for responding to posts from two other teams. In some cases, these posts were useful learning aids; however, at other times, questions from the textbook were used as the questions, and our instructor and his students both learned together that certain textbook questions were far more challenging than the material contained within the pages of the corresponding chapter.

The quizzes were intentionally designed to not include any R code, but to instead gauge students' understanding of the underlying themes and statistical concepts from the course. At first, this was surprising -- and even confusing -- for some students. However, building the quizzes outside of R helped us to reinforce the theme about concepts, rather than syntax. For a topic such as association rules, for example, several quiz questions were built around a very tiny, hypothetical series of transactions. Using only a pen, their notepaper, and a basic calculator, students were asked to calculate association rules such as support, confidence, and lift. Students' ability to complete a task such as this demonstrated whether they truly understood the underlying concepts. Typing lines of code into a console is a far less meaningful skill than being able to understand a data mining algorithm well enough to apply it in a setting such as this.

Conclusions and Lessons Learned

The first iteration of the course came with its share of bumps in the road, but we feel quite confident that AD699 is built with the right framework for ensuring our graduates' success, even long after their time at BU MET has ended. At the end of the

¹⁰ RStudio Team (2015). *RStudio: Integrated Development for R*. RStudio, Inc., Boston, MA URL <http://www.rstudio.com/>

first iteration of the course, feedback from students included several comments that indicated how much the course exposed them to entirely new ideas and concepts.

Since the students cannot be 100 percent certain of their post-graduation work environments yet, we feel that exposing them to a variety of approaches and concepts builds a foundation that they can further develop independently, based on the specific needs and demands on their employers. To use an analogy from everyday life, we reasoned that if a person was preparing to travel through Europe for several months -- but did not yet know which countries he would visit -- he would be better off learning 50 words in 10 languages than learning 500 words in a single language. This would provide him with a baseline from which he could build, once his travel plans became clear. While 500 words in a single language might be more rewarding in some ways, it would effectively limit our traveler's mobility range before his trip had even begun.

To continue with that analogy, if an AD699 graduate were to join the workforce and be asked to build a k-nearest neighbors model to aid with a classification task, we wouldn't expect him or her to flawlessly build and execute such a model on the spot. However, that graduate's experience with the data mining concepts would mean that he or she would be able to place this request in context. By reviewing class notes, and taking a look at some of the source code in the book, that student would be able to effectively "dust off" his skill set in that area and get back up to speed shortly thereafter.

Besides, it is more likely that our graduates will be managers or supervisors; therefore, that baseline familiarity will enable them to better understand the work done by those who build such models on a full-time basis. (Though, of course, we would never discourage our graduates from seeking to work directly in the field of data science). AD699 *alone*, however, is not going to transform anyone into a data scientist, just as a single statistics course would not claim to make its graduates statisticians.

A message that our instructor delivered at several points throughout the semester was that the underlying concepts behind the algorithms are more important to students' understanding. The particular tools used in data mining will change. Whereas a course such as AD699 is taught today in R, it may have been taught several years ago in MATLAB. Ten years from now, it might be taught in Python, or something else entirely. The concepts, however, are enduring -- therefore, the most important thing that students can take away from the course is a deep appreciation of the general way in which data mining algorithms are built and employed, rather than the specific nuances of the underlying code.

Most educated people are aware of the concept of Big Data. On some level, they understand that their demographic profiles and their online actions impact the way that marketers attempt to reach them with messages about products. They understand that data collection occurs all the time, through everything from smart devices to passive sensors to harvesting of messages, e-mails, and web clicks. They do not, however, know about the ways in which the distance between two records can be measured in order to influence an advertising decision or product recommendation. They do not know about the mathematical equation that undergirds most e-mail spam filters. They do not know about the mathematical relationships that help businesses help to uncover purchase patterns, including non-intuitive associations among products. Graduates of AD699, however, know all of these things. They will still have far more work ahead of them, should they wish to delve deeper into the worlds of machine learning and data science, but they will see their business careers, and the world around them, in a whole new way after having taken this course.

Bibliography

Hahsler, Michael, Christian Buchta, Bettina Gruen and Kurt Hornik (2018). arules: Mining Association Rules and Frequent Itemsets. R package version 1.6-1. <https://CRAN.R-project.org/package=arules>

Hahsler, Michael, Bettina Gruen and Kurt Hornik (2005), arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software* 14/15. URL: <http://dx.doi.org/10.18637/jss.v014.i15>.

Hahsler, Michael, Sudheer Chelluboina, Kurt Hornik, and Christian Buchta (2011), The arules R-package ecosystem: Analyzing interesting patterns from large transaction datasets. *Journal of Machine Learning Research*, 12:1977–1981. URL: <http://jmlr.csail.mit.edu/papers/v12/hahsler11a.html>.

Kelleher, John D. and Brendan Tierney, 2018. Kelleher, John D. and Brendan Tierney. *Data Science*. Massachusetts Institute of Technology, 2018.

Kross, Sean, Nick Carchedi, Bill Bauer and Gina Grdina (2017). swirl: Learn R, in R. R package version 2.4.3. <https://CRAN.R-project.org/package=swirl>

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL <http://www.rstudio.com/>.

Shmueli, Bruce, Yahav, et. al. *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. Wiley: Hoboken, NJ, 2018.

Therneau, Terry and Beth Atkinson (2018). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-13. <https://CRAN.R-project.org/package=rpart>

Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Authors' Information

| | | | |
|--|---|---|---|
|  |  |  |  |
| Greg PAGE , Boston University. Research Interests: Data Science, Machine Learning, E-Commerce | Slav ANGELOV , New Bulgarian University. Research Interests: Predictive Analytics, Regression, Data Modeling | Penko IVANOV , New Bulgarian University. Research Interests: Big Data Analytics, Data Modeling, Machine Learning | Vladimir ZLATEV , Ph.D., Boston University. Research Interests: Econometric Analysis, Linear Optimization, Statistical Applications, System Dynamics |

A FRAMEWORK FOR MODELING IN SCALE: AN INTRODUCTION

Eric BRAUDE

Abstract: *UML class models, arranged in the plain in the traditional manner, are very useful. However, this usefulness degrades rapidly for large, realistic models. The paper introduces a disciplined layout format that addresses this limitation.*

Keywords: *UML, class models, modeling, scale-ability, design visualization, software modeling*

ACM Classification Keywords: *Software and its engineering → Software creation and management → Designing software*

Introduction

UML class models are very useful for understanding the relationships between classes. But this usefulness degrades with the number of classes and connections, and almost disappears for classes in the hundreds. Critically, the traditional way of laying these out on a plane becomes impractical and unhelpful for realistically-sized applications. In particular, it is hard to identify underlying structures, locate a given class, or trace relationships that intersect others.

We present a layout format that addresses this problem.

Existing Literature

[Chaudron, 2017] reported on the difficulty of assessing how industry uses UML, but observed that it is used significantly for communication during development. We infer that this refers to smaller-scale, perhaps abstract or informal, figures. One way to address the scaling problem is to observe layout practices, as suggested by [Schmid, 2009] and [Jünger, 2003]. [Sharif, 2011], and [Andriyevska, 2004], showed that layout has significant effect on comprehension, and can be arranged to emphasize architectural levels. Another approach is to introduce kinds of telescoping notations. This is followed, in part, by [Kagdi, 2007] and Maletic in [3]. However, these all assume the traditional layout representation of UML—though modified in [3]. The method presented here borrows much from the standards, but is otherwise quite different.

The Approach

Our approach is organized around class hierarchies, each of which contains a maximum number of classes subject to the constraints described below. In other words, inheritance is the main organizing principle. (For the sake of clarity, we will include interface implementation with the term “inheritance.”) The centrality of inheritance can be seen in the example in Figure 2.

These inheritance hierarchies are organized in two sets, each arranged lexicographically. The uppermost set consists of those classes whose ancestors inherit from no more than one class within the application.

The lower set consists of the remaining classes. The parents of each of the latter’s roots are shown by means of an inheritance channel on the left of the hierarchies, as in Figure 2. Inheritance is denoted with the usual “ Δ ” symbol, and “implements” with “ \wedge .”

Associations (including aggregations and compositions) are displayed by means of vertical channels on the right of the inheritance list. There is one channel for each one-to-many relationship and one for each

many-to-one relationship. Figure 2 uses “ \Leftarrow ” and “ \Rightarrow ” symbols to identify the “one” in each one/many channel. Further distinguishing between association types remains a work in progress. Labels on associations can easily be inserted via a column to the right of the hierarchies (and exposed/suppressed at will).

Example

A realistic example, consisting of at least a hundred classes, would show a barely comprehensible figure when using a traditional UML class model. We believe that a comprehensible one results when using the format introduced above.

As a still-comprehensible example for comparison, we have taken the Java concurrency API class model at <https://www.uml-diagrams.org/java-7-concurrent-uml-class-diagram-example.html>. Figure 1 shows the original published model.

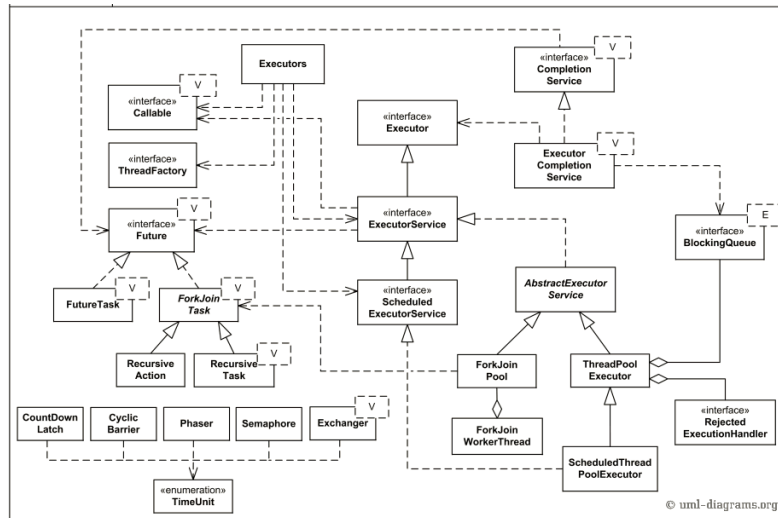


Figure 1: Java concurrency API Class Model

Figure 2 shows the form which we claim is scalable. In this example, there is only one interface (ScheduledThread PoolExecutor) that belongs to the lower set of hierarchies.

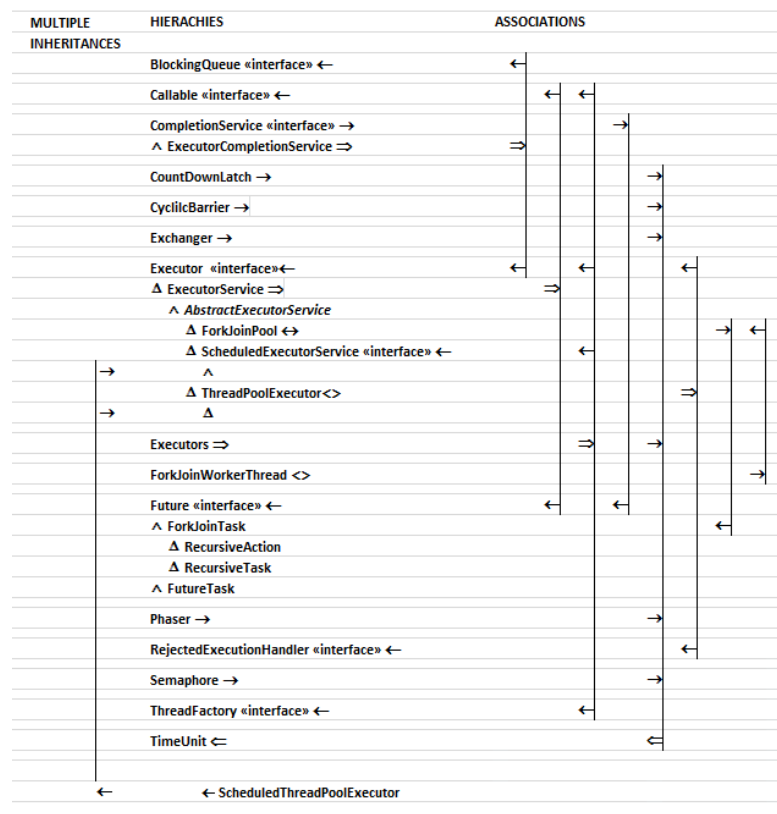


Figure 2. Java concurrency API in proposed format

Advantages

The main advantage is scalability. Standardization also allows the user to discern designs, freed from layout distractions. Classes can be easily searched. The alphabetical nature of the listing is familiar but is absent from traditional UML class model displays. Collapsing or expanding a given class to show attributes and methods (not shown in Figure 2) is trivial technically. To determine which classes are related to a given class is to identify the channels associated with it. Other slicings, such as displaying all classes related to a given class, are readily obtainable, either in isolated form or else highlighted in the context of the whole. The rectangular format also allows for hiding or showing the attributes and methods of selected classes (not shown in Figure 2) without significantly disturbing the rest of the diagram. The format also facilitates de-scaling such as replacing a hierarchy with the base class.

Reverse engineering applications to produce traditional UML class models produce potentially incomprehensible tangles, which limits its viability. We have not implemented reverse engineering yet for the format proposed but its disciplined nature promises untangled results.

Limitations

The technique described uses two categories of hierarchies; however, a more refined categorization may be appropriate. This has not yet been explored. Reverse engineering and the integration of packages have not yet been implemented.

The classical two-dimensional layout with boxes has the advantage that groupings can be shown without the strictures of a standard format. While the author believes that the format in Figure 2 is beneficial for large models, this hypothesis has yet to be verified by experiment. We are in the process of gathering informal feedback before finalizing the design of such an experiment.

Conclusion

We have presented a format for UML class models that we believe goes a long way to dealing with realistically-sized implementations. We are seeking feedback to be used for implementation and for designing experiments of effectiveness.

Bibliography

- [Andriyevska, 2004] Andriyevska, O., Dragan, N., Simoes, B., and Maletic, J. I., "Evaluating UML Class Diagram Layout Based on Architectural Importance," 3rd IEEE International Workshop on Visualizing Software for Understanding and Analysis, doi: 10.1109/VISSOF.2005.1684296
- [Chaudron, 2017] M. Chaudron. "Empirical studies into UML in practice: pitfalls and prospects, " Proceedings of the 9th International Workshop on Modelling in Software Engineering (2017) pages 3-4, doi: 10.1109/MiSE.2017..24
- [Jünger, 2003] M. Jünger, Karsten Klein, Joachim Kupke, Sebastian Leipert and Petra Mutzel "A new approach for visualizing UML class diagrams," Proceedings of the 2003 ACM symposium on Software visualization, pages 179-188, doi:10.1145/774833.774859
- [Kagdi, 2007] H. Kagdi and J. Maletic. "Onion Graphs for Focus+Context Views of UML Class Diagrams," 4th IEEE International Workshop on Visualizing Software for Understanding and Analysis, 2007, doi: 10.1109/VISSOF.2007.4290704
- [Sharif, 2011] B. Sharif "Empirical assessment of UML class diagram layouts based on architectural importance," 27th IEEE International Conference on Software Maintenance (2011), pages 544-549, doi: 10.1109/ICSM.2011.6080828
- [Schmid, 2009] K. Schmid. "Guidelines on the aesthetic quality of UML class diagrams," J. Information and Software Technology 51.12 (2009), pages 1686-1698, doi:10.1016/j.infsof.2009.04.008

Authors' Information



Eric BRAUDE, Ph. D., Associate Professor of Computer Science, Boston University MET College, 808 Commonwealth Ave, Boston MA 02215, ebraude@bu.edu.
Major Fields of Scientific Research: Software Engineering, Automated Theorem Proving

ON SPERNER'S THEOREM

Ivan Landjev

New Bulgarian University, Department of Informatics

Abstract: . Let R be a finite chain ring with $|R|=q^2$, $R/\text{rad}R=F_q$. Every submodule M of ${}_R R^n$ is isomorphic to a direct sum of cyclic modules:

$$M=R/N^{\lambda_1} + R/N^{\lambda_2} + \dots + R/N^{\lambda_n},$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ are integers and $N=\text{rad } R$. The n -tuple $(\lambda_1, \lambda_2, \dots, \lambda_n)$ is called the shape of M . We consider the partially ordered set P of all submodules contained in a module of shape $(\lambda_1, \lambda_2, \dots, \lambda_n)$. We prove an analogue of Sperner's theorem saying that the size of a maximal antichain in P is equal to

$$\sum_{\mu \leq \lambda} \begin{bmatrix} \lambda \\ \mu \end{bmatrix}$$

where the summand is defined as the number of all modules of shape μ contained in a module of shape λ and the the sum runs over all partitions $\mu=(\mu_1, \mu_2, \dots, \mu_n)$ with $\sum \mu_i = [(\sum \lambda_i)/2]$.

Keywords: finite chain rings, modules over finite chain rings, antichains, partially ordered sets, Sperner theorem

Sperner-type theorems answer the question about the maximal size of an antichain in a partially ordered set P . The classical Sperner theorem says that this size is $\binom{n}{\lfloor n/2 \rfloor} = \binom{n}{\lceil n/2 \rceil}$ for the partially ordered set (poset) of all subsets of an n -element set [S28]. A similar result holds for the poset of all vector spaces of F_q^n . The idea of one of the possible proofs is that the number of chains containing a fixed subspace depends only on its dimension [E97,LW92].

In this paper we consider the problem of finding the maximal size of an antichain in the lattice of all submodules of a finitely generated module over a chain ring R . In this case counting formulas are very complicated and do not seem to lead to a proof. Nevertheless a Sperner-type theorem can still be formulated.

In what follows R is a finite chain ring with $|R| = q^m$, and factor field isomorphic to F_q . We consider the free left module ${}_R R^n$ and denote by P_n the lattice of all submodules contained in ${}_R R^n$. Instrumental in our result will be the formula giving the number of submodules of given shape contained in a fixed module.

Lemma 1. Let ${}_R M$ be a module of shape $\lambda = (\lambda_1, \dots, \lambda_n)$ over the chain ring R . For every partition $\mu = (\mu_1, \dots, \mu_n) \leq \lambda$ the module ${}_R M$ has exactly

$$\begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \prod_{i=1}^n q^{\mu'_{i+1}(\lambda'_i - \mu'_i)} \begin{bmatrix} \lambda'_i - \mu'_{i+1} \\ \mu'_i - \mu'_{i+1} \end{bmatrix}$$

submodules of shape μ . Here λ', μ' are the conjugate partitions of λ and μ , respectively.

Let P be a poset. We call P a ranked poset if there exists a function $r: P \rightarrow N_0$ with $r(x) = 0$ for some minimal $x \in P$ and $r(y) = r(x) + 1$ for all y immediately preceding x . The maximal rank of an element in P is called the rank of P . The i -th level of a ranked poset P is defined by

$$L_i(P) = \{x \in P \mid r(x) = i\}.$$

The i -th Whitney number is the cardinality of L_i : $W_i = |L_i(P)|$. A graded poset is a ranked poset in which all elements have rank 0. We say that the element y of a poset P covers the element $x \in P$ if $x < y$ and if $x < y' \leq y$ then $y = y'$. In this case we say that x immediately precedes y and denote this by \prec . The Hasse diagram of a partially ordered set is a directed graph $H(p) = (P, E(P))$, where

$$E(P) = \{(x, y) \mid x \prec y\}.$$

The underlying non-directed graph is called the Hasse graph.

The lattice of all submodules of a finitely generated R -module is a graded poset. The rank function for a submodule of shape (μ_1, \dots, μ_n) is defined as

$r(M) = \sum_{i=1}^n \mu_i$. In case of $M = R^n$, we have $r(P_n) = mn$, where m is the nilpotency index of R . The k -th Whitney number is given by the following sum:

$$W_k(P_n) = \sum_{\mu} \begin{bmatrix} m^n \\ \mu \end{bmatrix}.$$

where the sum is over all shapes $\mu = (\mu_1, \dots, \mu_n)$ with $\sum_i \mu_i = k$.

It is said that level L_i can be matched into level L_j , where $j = i - 1$ or $i + 1$ if there is a matching of size W_i in the Hasse

graph defined on the elements of $L_i \cup L_j$. Our main tool is the following theorem which is a corollary of Dilworth's theorem.

Theorem 2. Let P be a graded poset. If there exists an index h such that L_i can be matched into L_{i+1} for all $i = 0, 1, \dots, h$, and L_i can be matched into L_{i-1} for all $i = h + 1, \dots, n$, then the size of the largest chain is $W_h = W_h(P)$.

Next we need a lemma on the existence of matching in certain bipartite graphs. A bipartite graph $G = (X \cup Y, E)$ is called piecewise regular if there exist partitions of X and Y

$$X = X_1 \cup \dots \cup X_s, \quad Y = Y_1 \cup \dots \cup Y_t,$$

Such that each vertex in X_i is adjacent to exactly x_{ij} vertices in Y_j and each vertex in Y_j is adjacent to exactly y_{ji} vertices in X_i .

Now a necessary and sufficient condition for the existence of a matching of size $|X|$ in G is given by the following lemma.

Lemma 3. Let $G = (X \cup Y, E)$ be a piecewise regular bipartite graph. A necessary and sufficient condition for the existence of a matching of size $|X|$ in G is given by the inequality

$$\sum_{i \in I} |X_i| \leq \sum_{j \in J(I)} |Y_j|$$

for every set of indices $I \subseteq \{1, \dots, s\}$. Here $J(I) = \{j \mid x_{ij} > 0, i \in I\}$.

Starting with this result we can prove our main theorem. It is based on the observation that every two consecutive levels in the Hasse graph form a piecewise bipartite graph.

Theorem 4. The size of a maximal chain in P_n is equal to

$$\sum_{\mu \leq \lambda} \binom{\lambda}{\mu}$$

where μ runs over all partitions $\mu = (\mu_1, \dots, \mu_n) \leq (m, \dots, m)$ with $\sum_{i=1}^n \mu_i = \lfloor km/2 \rfloor$.

Proof. Check the necessary and sufficient condition of Lemma 3 using the formulae from Lemma 1.

Acknowledgements. The research was supported by the Scientific Research Fund of Sofia University under Contract 80-10-51/17.04.2018.

References

- [E97] K. Engel, Sperner Theory, Encyclopaedia of Mathematics and its Applications 65, Cambridge University Press, 1997.
- [LW92] J. H. van Lint, R. M. Wilson, A Course in Combinatorics, Cambridge University Press, 1992.
- [S28] E. Sperner, Ein Satz ueber Untermengen einer endlichen Menge, Math. Zeitschrift 27(1928), 544-548.

Author's Information



Ivan Landjev, Prof. DSc

***New Bulgarian University, Department of
Computer Science***

***Major Fields of Scientific Research: finite
geometries, coding theory, combinatorics,
algebra***

i.landjev@nbu.bg

INTRODUCING AGILE CONCEPTS IN PROJECT MANAGEMENT AND SOFTWARE DEVELOPMENT COURSES

Vijay KANABAR, Kalinka KALOYANOVA

Abstract: *Software project management and software development courses have started introducing Agile Project Management theory and practice into their curriculum. Pedagogically very little has been written about their curriculum design, approaches or tools and techniques when attempting to embed agile topics into such courses. In this paper, we survey agile concepts and describe a curriculum of topics and learning outcomes which can be injected by any instructor seeking to enhance their course with agile theory and practice. We also map the PMBOK® standard with agile practice.*

Keywords: *Project Management Education, PMI Curriculum, Agile Project Management, Information Systems Education.*

ACM Classification Keywords: *K.3.2 Computer and Information Science Education, Curriculum, Information Systems education, K.6.1 Project and People Management, Management techniques.*

Introduction

In this section we provide a research overview for teaching both agile concepts and Project Management (PM) concepts in any Computer Science (CS), or Information Systems (IS) program. A research conducted by the leading Communications of ACM reveals the sad state of coverage of PM topics in many CS or IS curricula [Reif & Mitri, 2005]. Table 1 provides an overview of the level of PM coverage by typical courses one would find in a Computer Science department offering. It reveals that while courses titled with labels such as systems analysis or PM have good project management coverage, the other key courses don't appear to be doing justice to this important topic.

Table 1. Coverage of Project Management Literacy in Computer Science Courses (*Adapted from [Reif & Mitri, 2005]*)

| Course title | Breadth of PM Coverage |
|--|------------------------|
| Introduction to IS | Light |
| Advanced IS | Moderate |
| E-business | Moderate |
| Introductory Programming | Light |
| Advanced Programming & Web Development | Light |
| Networking | Light |
| Database | Light |
| System Analysis | Very heavy |
| Project Management | Very heavy |
| DSS/IA/Knowledge Management | Light |

| | |
|--|----------|
| IT Management or Business Process Design | Moderate |
|--|----------|

This situation, where inadequate coverage of project management topics exists in many programs today, has led to crisis of quality in the software industry where only 20% of large software systems are implemented on time and, of those, approximately two-thirds experience cost overruns approaching 100% [Reif & Mitri, 2005].

More recent research by the co-authors reveals that this trend is universal. Within project management, a key area of emphasis is developing communication and leadership skills. To address workforce expectations, instructors must focus not only on technical project management skills but also on the learning outcomes that develop soft skills or behavioral competencies [Kanabar & Kaloyanova, 2017]. Another research dealing with assessing mastery of Project Management core competency in an IT project management course arrived at several key conclusions:

- The students appeared to be competent learning and leveraging technical skills rather than non-technical skills.
- While the students from both USA and Bulgarian universities in the study demonstrated similar competence for technical project management knowledge areas, the students from USA showed better results in non-technical areas. [Kaloyanova & Kanabar, 2015]

IS curriculum developed by a consortium of industry and educational experts advocate increased emphasis on project management (PM) and even recommend an entire course emphasizing PM concepts and practice [Reif & Mitri, 2005].

While it is a challenge to introduce coverage of soft topics in many computer science programs due to limitations of credit hours, it is certainly possible to optimize coverage in a distributed manner across courses. For example, a database course could cover project life cycles within the context of database life cycles and the advanced programming course should involve teams working together on a project preferably involving external industry participants as sponsors. Such practicum approach for software project management education has proven to be effective according to a lot of research [Manzil-E-Maqsood & Javed, 2007]. In this manner referring to Table 1 again, a curriculum map of project management topics across the courses described should minimize redundancy and maximize of project management topics.

Project Management and Agile Concepts

IT projects are heavily represented in the ever-increasing annals of failed endeavors, with many spectacular examples portraying unfulfilled ambition and immaterialized promise, and one can conclude that software development is now considered to be among the most challenging and complex activities carried out by humans. [Dalcher, 2015]. Sometimes the reason behind the failure of a software project is having project managers with poor skills, however, not leveraging the correct methodology or frameworks for software development might also be a key cause. To mitigate this, educators, in addition to making sincere effort to provide students with Project Management literacy in their CS/IS courses, must focus on leading frameworks for managing projects and teams. The agile software development methodology has turned out to be one of the most universal pathways for developing software and managing projects. A key motivation early on for the agile movement was the complex structure of software and the frequent changes to requirements. Applying agile concepts to PM processes can result in delivery of quality software at a quicker pace. An analysis of the history of classical software development and agile development reveals that the agile practice for software development is mentioned by Basili as early as 1975, subsequently Gladden and Gilb proposed the practice of “delivering working software early” in the early 1980s to address the issue of late delivery of software products resulting in customer dissatisfaction [Jiang & Eberlein, 2009].

Early approaches to software development were highly plan based. Here the development team creates a complete project plan and executes it to completion without obtaining stakeholder feedback. Quite frequently, in almost all cases, this approach resulted in “shock and awe”. If agile concepts are applied, the entire development team is compelled to think of a project's scope in terms of larger business goals first and then at a tactical level. If upon early inspection of working software, the sponsor wishes to make strategic changes to the software, the entire project and rest of the product design must change. The key attributes of agile project management are adaptability and willingness to make changes.

Before we continue further, we need to answer the question whether agile approaches are for only software projects. This is a critical question, as the scope of our paper is wide – we hope to present an agile rich project management curriculum for industries spanning from manufacturing to life sciences. “The terms agile and agility can be traced back to the manufacturing industry in 1991 when lean development emerged in manufacturing with the aim of eliminating waste, amplifying learning, delivering as fast as possible and empowering teams and Youssef even coined the term agile manufacturing around that time” [Jiang & Eberlein, 2009].

It is evident therefore that agile methodologies can be traced back to traditional engineering disciplines possibly before the software sector. In a survey nearly half of the respondents reported that they used Agile tools for non-IT projects [Taiga,2015]. The news that there are several companies who are already reaping the benefits of Agile methodology in other industries is intriguing, the most useful features leveraged here appears to be workflow tracking - 90% respondents, story mapping and analytics - 89% and scrum boards and activity streams are third with 85% people's votes [Taiga,2015].

Based on interview discussions, the following strengths of agile software development were identified in a recent research paper - intense and good cooperation, simple work planning by chunking, possibility to correct mistakes quickly, and if preconditions are in place the quality is usually good. [Taymor, E, 2018]. Agile strengths can turn to weaknesses if agile principles are not followed actively. In a study conducted by [Gandomani et al.,2013], the obstacles and issues in agile software development are categorized under four themes; organizational and management related challenges, people related challenges, process related challenges and technology and tools related challenges. Many of the current challenges stem from the culture and structure of the organization which is serving needs of traditional methods.

Let us delve into agile frameworks next. We leveraged Scrum as the framework of choice to teach agile project management to students. According to the creators of the Scrum Guide Ken Schwaber and Jeff Sutherland, Scrum is a framework within which people can address complex adaptive problems, while productively and creatively delivering products of the highest possible value; scrum is designed as to be lightweight as such it is simple to understand and from our experience not too difficult to master [Schwaber, K. and Sutherland, J. 2017].

Why did we choose scrum as the avenue to provide experiential knowledge to our students? We believe it is its versatility as well in addition to the simplicity. To quote the architects Sutherland and Schwaber [Schwaber & Sutherland, 2017]: “Scrum has been used extensively, worldwide, to:

1. Research and identify viable markets, technologies, and product capabilities;
2. Develop products and enhancements;
3. Release products and enhancements, as frequently as many times per day;
4. Develop and sustain Cloud (online, secure, on-demand) and other operational environments for product use; and,
5. Sustain and renew products.”

Scrum has been used to develop software, hardware, embedded software, networks of interacting function, autonomous vehicles, schools, government, marketing, managing the operation of organizations and almost everything we use in our daily lives, as individuals and societies.

As technology, market, and environmental complexities and their interactions have rapidly increased, Scrum's utility in dealing with complexity is proven daily [Schwaber, K. and Sutherland, J. 2017].

Some of the popular agile frameworks are:

- Extreme Programming (XP) -- a software development methodology intended to improve software quality and responsiveness to changing customer requirements. Like Scrum it promotes frequent "releases" and readily adopts new requirements.
- Crystal -- The Crystal family of methodologies includes several different methodologies for selecting the most suitable methodology for each individual project. Each member of the Crystal family is marked with a color indicating the 'heaviness' of the methodology, i.e. the darker the color the heavier the methodology. This approach advocates more coordination and heavier methodologies for larger projects than the smaller ones.
- DSDM (Dynamic Systems Development Method) is a framework for developing software using clearly defined phases, sub-phases, roles and principles in order to enable development teams to work efficiently.
- Kanban is a lean software development methodology that focuses on just-in-time delivery of functionality and managing the amount of work in progress (WIP) very elegantly [Inflectra, 2018]. This aspect is popular with developers and Scrum teams frequently use it.
- LEAN - As noted earlier Lean was popular in the manufacturing sector. It promotes maximizing customer value, while minimizing waste and aims to create more value for the customer while using fewer resources.

Agile Learning Outcomes and Modules

In late 2012 the Project Management Institute (PMI) responded to the expressed need of educators interested by designing a series of workshops to collect more systematic information from faculty on current teaching in project management and input on how to respond to increasing demand for curriculum support [IEEE Guide--Adoption, 2011]. The project management curriculum guidelines provide a good opportunity to understand the core learning outcomes [Task Force on PM Curricula, 2015].

1. Distinguish the approaches, advantages, and disadvantages of both classic and agile project methodologies, assess the deliverables and contexts best suited to each method, and apply these principles to the development of an appropriate PM strategy.
2. Develop a workable PM approach that includes the typical steps, activities, and participant roles for an agile project, and evaluate how and when these agile characteristics can be integrated with steps from a traditional PM life cycle to achieve an effective hybrid approach.
3. Use appropriate tools and resources for agile projects, including specific or adapted metrics that can assist the project manager in defining, executing, and controlling projects that follow an agile, or hybrid, life cycle and methodology.

Key modules that address the above learning outcomes should include:

- History, principles, and values of agile PM and the Agile Manifesto
- Understanding agile: general practices
- How agile PM is similar and how it is different from traditional PM life cycles
- Strengths and weaknesses of the agile approach
- Agile frameworks – Focus on Scrum
- Design and Implement of an agile project

From our experience with teaching a course, we recommend five agile lectures that supplement a traditional project management course. The recommended lecture modules align very well in scope with a pure agile only course [Meyer B, 2018].

- Introduction: The Agile manifesto and the context of agile methods and key methodological ideas that underlie the agile movement
- Agile roles: Discuss how does agile redefine traditional software jobs and tasks, and what is the project manager's role. This is an issue of concern as pure Scrum adopters don't define any role for the project manager.
- Agile practices: what are the concrete techniques that agile teams use to apply these methods. This should introduce all the agile artifacts and the duration of the various timeboxes.
- Agile Tools: What practical tools are needed for agile developers. Tools that can be used range from Jira and Trello to VersionOne. Fortunately, a lot of vendors provide limited licenses for students at no cost.
- Agile Group Project: Forming student teams, assigning roles, setting expectations for deliverables. We strongly believe that the Scrum term project that students implement provides real-world agile experience.

Integrating PMBOK and AGILE

One might be wondering if a relationship exists between the Project Management Body of Knowledge (PMBOK) and agile processes. Several research papers compare PMBOK project risk management processes with an agile project management approach, for instance, Scrum.

To illustrate, the risk management knowledge area consists of processes such as plan risk management, identify risks, perform qualitative risk analysis, perform quantitative risk analysis, create risk response plan, and monitor and control risk.

We believe the same processes are adopted by agile developers. For instance, Scrum developers are actively identifying risks and mitigating them when they do a daily 15 minutes standup and conduct a sprint end retrospective.

So, while the standards and guidelines are different, it is possible for an instructor familiar with both PMBOK and Scrum to teach the same concepts in a course concurrently.

In our group project we try to make sure that students understand both PMBOK and Agile as shown in Table 2. This is important knowledge as students must be versatile with different approaches since we are unsure which organization will be hiring them.

Table 2. Agile Group Project

| Phase | Key tasks | Learning Outcome |
|-------------------------|---|---|
| Teams | Roles of Product Owner, Scrum Master, and Development teams are assigned. | Students can redefine traditional project management jobs and tasks to agile |
| Product Being Developed | Identify key features | Students map traditional business analysis and requirements management to agile and map the Initiating phase of PMBOK to Scrum |
| Scrumming | Students implement the project | Students understand traditional PMBOK processes of Planning, Execution, Monitor & Control, and Closing. They map differences with Scrum practice. |

In the Table 3 below we describe the parallels between PMBOK and Agile for the Initiating phase.

Table 3. Initiating Phase

| PMBOK - Initiating | AGILE - Envisioning |
|---|--|
| Project Sponsor is identified Strategic Goals are defined High Level Scope is presented Assumptions are made High Level Budget is calculated Statement of Work is done High Level Risks are identified Project Key Success Factors are fixed | Product Owners and Stakeholders communication is available Backlog Item/User Story identification is started Product Owner monitors the links with the Strategic Goals |

Next we discuss the planning phase. Some key points to note here are: backlog items/user stories can be scaled to have no dependencies, all stakeholder communications occur through product owner, risk management occurs during the agile sprints, and impediments are identified at the team level and removed during daily standups, resource management is set with dedicated teams, costs are evenly spread and predicted based on dedicated teams and steady team velocity. The overall planning focus is on product and this is different from traditional projects which are “project based”. Quality and testing aspects are addressed in each sprint by the team as they know the product very well. Product owner is responsible for both internal and external stakeholders management.

Table 4. Planning Phase

| PMBOK - Planning | Agile - Speculating |
|--|---|
| Scope is defined Requirements are detailed Initial Design is done Project Schedule is prepared A set of plans are ready: <ul style="list-style-type: none"> • Risk Mitigation Plan • Communication Plan • Resource Management Plan • Cost Management Plan • Quality Management Plan • Stakeholder Management Plan • Procurement Management Plan | Extensive conversations between Product Owners and Scrum Team Extensive work on Backlog Item/User Story List Preliminary Design is made in parallel with Backlog preparing Project Schedule is detailed based on the Backlog and Sprint schedule |

We do not delve into the parallels of the execution and controlling phase of the PMBOK guide due to constraints but note that these stages translate to Exploring and Adapting aspects of Agile approaches respectively.

The Closing phase of the PMBOK guideline translates to closing phase of the agile project as well. However, note that with approaches such as Scrum many aspects of project closure such as “Lessons Learned” are done at the end of a Sprint within the context of “Retrospective”.

Conclusion

In this paper we have tried to introduce the reader to the merits of agile approach. We described various models, that one could consider. We described the essential learning outcomes and course modules or topics that one should embed in a traditional project management course. Finally, an instructor teaching such a course should be able to map traditional plan-based approach for doing projects with agile approach. We have described a few parallels between the PMBOK guide and Agile approach.

In a future research, we plan to collect data of our course executions for the “hybrid” curriculum and interpret student satisfaction and learning. While the hybrid approach teaches two approaches, it is quite possible that students might learn more optimally from a pure Agile delivery of the project management course.

Bibliography

- [Abrahamson, P., Salo, O., Ronkainen, J. and Warsta, J., 2002]. Agile Software Development Methods. [ebook] VTT Publications. Retrieved from: <https://www.vtt.fi/inf/pdf/publications/2002/P478.pdf>.
- [Broman, S., 2017]. Agile Pitfalls. Masters. Helsinki Metropolia University of Applied Sciences.
- [Dalcher, D., 2015]. Introduction to Software Project Management. Project Management Journal, 46(4).
- [Gandomani et al, 2013] Obstacles in Moving to Agile Software Development Methods; at a Glance, Journal of Computer Science , Volume 9, Issue 5, pp. 620-625
- [IEEE Guide--Adoption, 2011]. IEEE Guide--Adoption of the Project Management Institute (PMI(R)) Standard A Guide to the Project Management Body of Knowledge (PMBOK(R) Guide)--Fourth Edition, 2011.
- [Inflectra., 2018]. What is Agile Kanban Methodology? Retrieved from: <https://www.inflectra.com/methodologies/kanban.aspx>.
- [Jiang, L., & Eberlein, A., 2009]. An analysis of the history of classical software development and agile development. In (pp. 3733-3738).
- [Kaloyanova & Kanabar, 2015] Assessing Mastery of Project Management Core Competency in an IT Project Management Course, Proceedings of the 11th International Conference Computer Science and Education in Computer Science (CSECS), 4-7 June, 2015, Boston, USA, pp. 19-25.
- [Kanabar, V & Kaloyanova, K., 2017] Identifying and Embedding Behavioral Competencies in Information Systems Courses". In Proceedings of the 25th European Conference on Information Systems, Guimarães, Portugal, 2017, pp. 3115-3122.
- [Kanabar, V & Messikomer, C., 2015] Design and Implementation of an Adaptive Curriculum Framework for Project Management Education, Proceedings of IRNOP 2015, International Research Network on Organizing by Projects, London, UK, 2015.
- [Manzil-e Maqsood & Javed T., 2007] Practicum in software project management: an endeavor to effective and pragmatic software project management education, In Proceeding of The 6th Joint Meeting on European software engineering conference and the ACM SIGSOFT symposium on the foundations of software engineering: companion papers, Dubrovnik, Croatia, 2007 pp. 471-479
- [Meyer B., 2018] Agile Software Development. Retrieved from <https://www.edx.org/course/agile-software-development>

- [PMBOK, 2017] A guide to the project management body of knowledge (PMBOK guide). Newtown Square, Pa: PMI, 6th Edition, 2017.
- [Reif, H. L., & Mitri, M., 2005]. How university professors teach project management for information systems. Communications of the ACM, 48(8), pp. 134-136.
- [Schwaber & Sutherland, 2017] The Scrum Guide, 2017, Retrieved from: <https://www.scrumguides.org/docs/scrumguide/v2017/2017-Scrum-Guide-US.pdf#zoom=100>
- [Task Force on PM Curricula, 2015]. PM Curriculum and Resources. Newtown Square, Pa: Project Management Institute, 2015.
- [Schwaber, K. & Sutherland, J., 2017]. The Scrum Guide: The Definitive Guide to Scrum: The Rules of the Game. Creative Commons.
- [Taiga, 2013] Agile as a management tool for non-IT industry: an insight, Retrieved from: https://blog.taiga.io/agile_as_management_tool_for_non_IT.html
- [Taymor, E., 2018]. Agile Handbook. [ebook] Philosophie. Available at: <http://agilehandbook.com/agile-handbook.pdf>

Acknowledgments

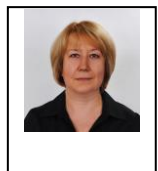
Kaloyanova: This paper is partially supported by Sofia University "St Kl. Ohridski" SRF under the contract 80-10-143/2018.

Authors' Information



Vijay KANABAR, PhD, PMP, CSM, Boston University, Metropolitan College, Associate Professor. Email: kanabar@bu.edu

Major Fields of Scientific Research: IT Project Management, Curriculum Design and Development, Online education. Web application Development. Web Analytics.



Kalinka KALOYANOVA, Professor, PhD, University of Sofia, Faculty of Mathematics and Informatics and BAS – Institute of Mathematics and Informatics, kkaloyanova@fmi.uni-sofia.bg.

Major Fields of Scientific Research: Database Systems, Information Systems, Big Data, Software Engineering, Project Management, Service Management.

CSECS 2018, pp. 181 - 191

Computer Science and Education in Computer Science
14th Annual International Conference
ISSN 2603-4794

June 29 – 30, 2018, Boston, USA

TEST DESIGN PROCESS IMPROVEMENT BY SIX SIGMA (DMAIC) AND R

Dobromir Dinev

***Abstract:** The world we live in is characterized by certain dynamism and vitality, effortlessly recognizable by all the happenings around us. The organizations that make business in such environment, should endeavor to become effective and efficient; and at the same time to be profitable. There is methodology that helps those organizations eager to achieve such uneasy task (to have better services and products); often the same has been referred to the Swiss Army knife of process improvement; this is Six Sigma. The article below is presenting a project that has been conducted by following the DMAIC trail of Six Sigma; the project itself managed an improvement of 37% of the test design process. For the Six Sigma statistics during the project it has been used the R programming language. Only results from the executed code will be presented in the article to follow.*

***Keywords:** Six Sigma, DMAIC, Kaizen, Test design process, improvement*

***ACM Classification Keywords:** Software and its engineering*

Introduction

The project presented in the current article has been selected among list of projects in the testing function for software delivery of greatly customized end user software solutions in the financial field. Each of the future Six Sigma projects (from the list prior selecting the presented in this article project) was supposedly expected to improve different part of the testing process itself, e.g. it was addressing diverse kind of issues in the process. The main reason for the current project to be selected was the impact on quality over the deliverables in the next process's stages, where the manual cases are being executed. It was well known fact by time that the rework in terms of number of reworked and dropped test case scenarios during the test execution is very high -

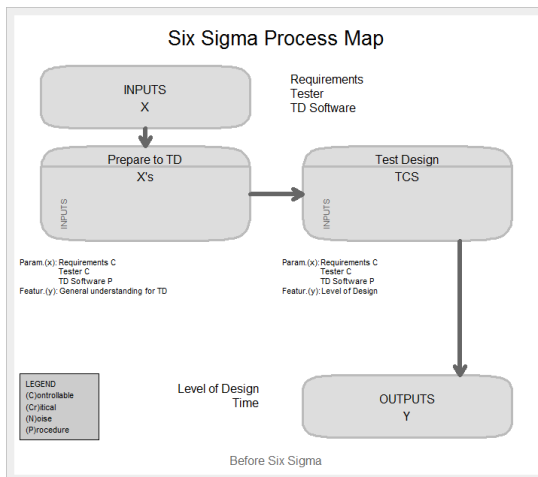


Figure.1. Preliminary process Kaizen DMAIC project, R

57% from the testing scope, across the projects. That was asking for a short Kaizen DMAIC project to be run in search of the root causes for the current situation; this short version of the typical DMAIC process is especially appropriate if the root cause is obvious [George, et al., 2005], as it was in the current situation.

Fig.1. is presenting the process and its state by the time. From the preliminary analysis of the process was clear that: there was no review by the test team members of the test scenarios (which is considered a best practice), often change of the specifications. The kaizen DMAIC process has been not conducted till completion as all participants in the project agreed, at the analysis stage of the project, that few changes should be implemented immediately to remediate the current situation. Next changes were implemented immediately in the test process: 1) internal review of written scenarios from senior testers, 2) formal questions/answers sessions between the testers and the business analysts. In parallel to this decision it has been decided by the senior management a full Six Sigma DMAIC process to be established; its primary goal: To improve the test design process, in a ways to prohibit such high number of reworked/dropped test scenarios in the execution phase of the testing scope.

Project charter, problem statement, and metrics (Define)

At the time the full Six Sigma project started, the results from the initial changes made by the kaizen DMAIC project were visible. The average number of reworked test cases has dropped by 20% to 37% across the projects. This was the main reason in the project charter the problem statement is phrased as follows: "There is very high level of rework (avg. 37%) from the test design process in test executing phase of the testing function." Further, the objective of the project has been set to: "To decrease the level of waste and to organize the process in a way it prohibits future growing in the primary metric." The main metric has been, expectantly, set to be DPO < 10 %. During the kaizen DMAIC project it has been identified that the best time to spend in writing particular test scenario is between 9 and 15 minutes. Based on that finding it has been formulated the first in row, secondary metric: The test case design to be 11 +/- 4 min. Additionally, to the secondary metrics were added next two: "Level of Specification description between "very good", "good", and "fair", and "Refactoring test cases - 0 min, no more than 5 min,

for the primary metric is kept”. The process map after the kaizen project is presented in Fig.2. This was the initial process map. At that stage five, different in size, projects were selected and used for the statistics (Table 1.).

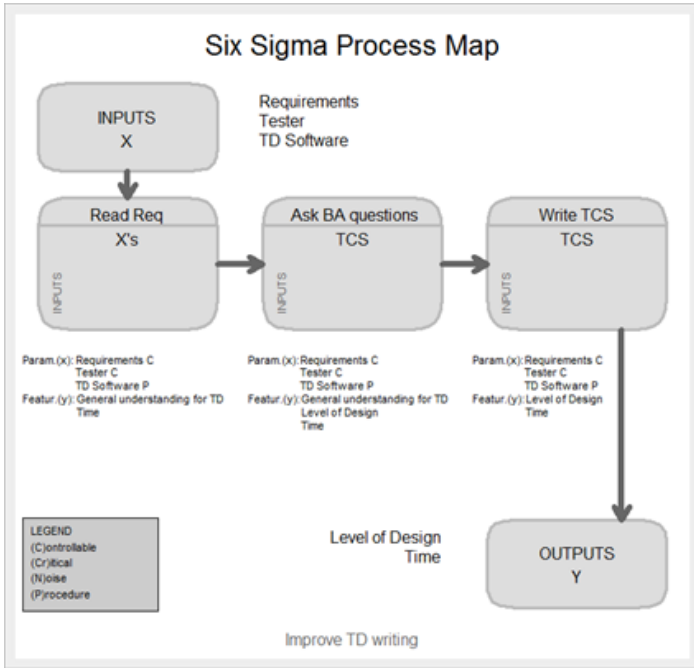


Figure 2. Process map from R, after kaizen DMAIC project

| Project | Category | Number of scenarios |
|---------|----------|---------------------|
| P1 | Big | 1944 |
| P2 | Small | 111 |
| P3 | Medium | 186 |
| P4 | Medium | 272 |
| P5 | Small | 57 |

Table 1. Selected Software projects in the Six Sigma Project

Waste identification

For the identification of the waste generators a series of a brainstorm meetings were conducted. The test team has been divided in two groups. To the second group of test specialists as an addition to them were invaded and present representatives from the business analysis, the development and management. To both groups the same questions has been given to brainstorm on: “What are the reasons, by your opinion, for low quality test cases to be identified during the test execution? What may cause the

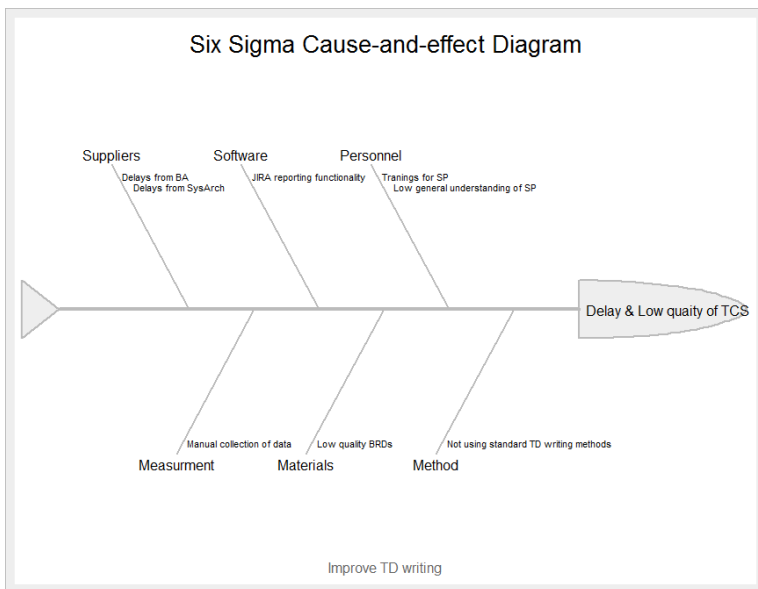


Figure 3. What are the reasons for low quality test cases to be identified during the test execution, as well wat cause delays

test team to count delay during the same phases?"; The findings from those sessions are presented in Fig.3. Consequently, as main CTQs were identified 1) the maturity level of the specification given to the test team to write scenarios, 2) delays from the

previous functions: business analysis, technical architecture, and etc. 3) time spent in reading the specifications, 4) time spent in writing the particular scenario 5) experience level of the test case creator, 6) manual collection of KPI data. 7) specifics of the software used for test case creation, 8) missing knowledge for the software solution to be tested and for the business. By multi-voting [1] next few CTQs were prioritized over the other, and used in the project: Quality of the specification given to the test team to write scenarios (Maturity_spec); Test case maturity level (TCS_level); time spend in creation for the particular test cases scenario. The waste indicator has been used in a way its value to show root cause for its presence (ch_indicator); Table 2, is holding information for the discrete variables collected for the five projects, and their corresponding values.

| Variable | Values |
|------------------|--|
| TCS_level | Very good Description |
| | Good Description |
| | Fair Description |
| | Poor Description |
| | Very poor Description |
| Maturity_spec | Very good |
| | Good |
| | Fair |
| | Poor |
| | Very poor |
| ch_indicator, | None refactored |
| | Dropped from execution |
| | Added |
| | Refactored |
| | Other |
| Reason_indicator | String saying what was the reason for the change |

Table.2. Collected information for each of the projects

Apart those listed in the table, information for the creator’s seniority has been collected. For each test case we also gathered data in minutes for the time needed the scenario to be written, and for the time needed for the specific remediation of the “waste” situation around each test case. For instance, if scenario is added, the time spent in writing is consequently considered a waste, or the time needed to amend other scenario to reflect new requirements, again stored and afterwards taken as waste to the test process.

Initial measure of the main metric in measure phase

The initial measures regarding the main metric was showing values across the selected projects between 30.88% and 36.55% with avg. 34.22%.(Table 3.)

| Project | TCS N | Dropped | Added | Re-work | FTY | De-fects | DPMO | DFO % |
|---------|-------|---------|-------|---------|-------|----------|--------|-------|
| p1 | 1944 | 30 | 220 | 424 | 65.32 | 685 | 352366 | 35.23 |
| p2 | 111 | 0 | 15 | 24 | 64.86 | 39 | 351351 | 35.13 |
| p3 | 186 | 0 | 15 | 53 | 63.44 | 68 | 365591 | 36.55 |
| p4 | 272 | 0 | 15 | 67 | 69.85 | 84 | 308823 | 30.88 |
| p5 | 57 | 0 | 9 | 10 | 66.66 | 19 | 333333 | 33.33 |

Table 3. Main metric in measure phase, for each selected

Analyze – main findings confirmed

During the Analyze phase the main expectation for the maturity of the specification is having direct impact on the quality of the written scenarios and the presence of waste has been proven. Still the results were showing weak association by the Goodman Kruskal gamma statistics. (Fig.4.) Further, examination (*measurement system analysis*) has been done for the possible cause of those results.

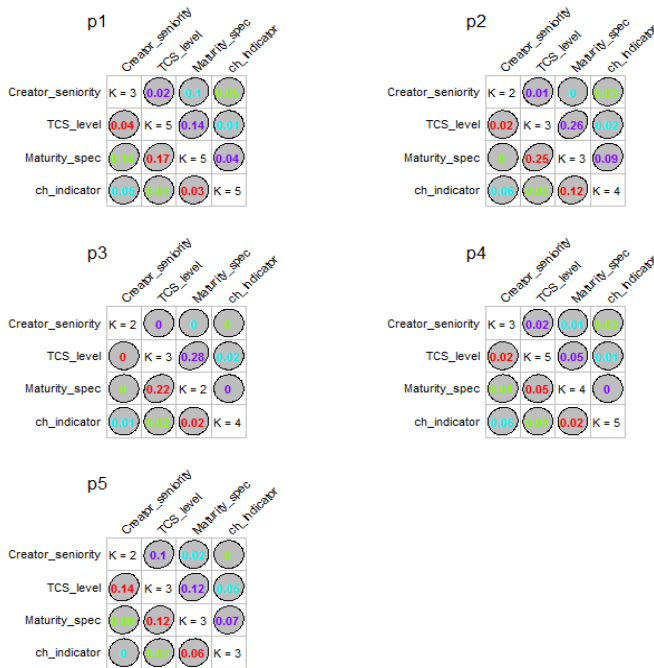


Figure 4. Association between the variables

During the evaluation process of the test case maturity (TCS_level) and the quality of the specification given to the testers to write their scenarios (Maturity_spec) each of the domain experts was having different criteria to categorize particular scenario in given category, same was present for the quality of the specification provided. After the source of the variance has been identified a corrective actions were taken the measurement system to the changed in a way the unwanted variance is eliminated. Those actions were mainly in two categories: a set of rules for evaluation the test scenarios and the specifications were created, and KPIs indicating the quality of the written specifications, were

introduced, as well this second category is used in one of the improvements implemented to minimize the waste.

Improve

After the Analysis phase it comes the improve, here next few improvements were suggested and adopted by the organization: 1) introducing toll gating system, to guarantee the provided specification documents are with the minimum required quality, 2) meeting(s) to introduce the test team to the required and specified software solution to be developed and tested, 3) introduce internal for the test team review meetings, following strict procedure for conducting the review in order to confirm the relevance and the correctness of the written scenarios, 4) introducing external reviews of the test scope (test scenarios) by the business analysts, developers, and if considered needed by the client, 5) creation of a training program to confirm the skills of the test experts are at the same level, as well an refresh program for the same skills, 6) automatic collection of metrics data.

Control

To achieve the purpose of the control phase and confirm sustainability of the process was put in place an automatic dashboard showing all the projects, serviced by the test team group. The dashboard itself is presenting a u-chart (following the standard formulas) per project for the reworked scenarios, by category, as well it presents information for the main metric (and average for all the projects) and for all the supplementary metrics from the Six Sigma process. The dashboard is always up-to-date, which is allowing decisions to be made during the project life-span and on time corrective actions to be taken. Table 4. Is containing information for the primary metric and the general Six Sigma process measure, by the figures below it is visible the process is still not stable, and there is room for even better sigma level, which at the moment is between 2 and 3.

| Project | TCS N | Dropped | Added | Re-work | Defects | FTY | DPO % | DPMO |
|---------|-------|---------|-------|---------|---------|-------|-------|--------|
| p1 | 2200 | 50 | 156 | 192 | 407 | 81.5 | 18.5 | 185000 |
| p2 | 863 | 150 | 5 | 27 | 182 | 78.91 | 21.08 | 210892 |
| p3 | 562 | 19 | 32 | 60 | 112 | 80.07 | 19.92 | 199288 |
| p4 | 128 | 6 | 9 | 19 | 34 | 73.43 | 26.56 | 265625 |
| p5 | 491 | 11 | 9 | 59 | 81 | 83.50 | 16.49 | 164969 |
| p6 | 56 | 1 | 5 | 4 | 10 | 82.14 | 17.85 | 178571 |
| p7 | 92 | 9 | 20 | 2 | 31 | 66.30 | 33.69 | 336956 |
| p8 | 168 | 15 | 10 | 9 | 34 | 79.76 | 20.23 | 202380 |
| p9 | 255 | 10 | 13 | 16 | 44 | 82.74 | 17.25 | 172549 |
| p10 | 1560 | 36 | 98 | 102 | 240 | 84.61 | 15.38 | 153846 |

Table.4. Results across the projects after the improvements has been implemented

Conclusion

The applied Six Sigma methodology – DMAIC, brought an average improvement of the main metric by 37%. During the project it has been identified that despite the changes made in the measurements system there is still huge amount of variance sourced by the fact: The test design process is a creative (intellectual) process and the involved people in it, those people are in different mental and psychological condition over the project life-span. It will be beneficiary to see the impact of this condition over the quality of the work provided, as well is there relation between the provided test cases and the Belbin's team roles [Belbin, 2010].

Bibliography

- [George, et al., 2005] The lean Six Sigma pocket tool book, M.L. George, D. Rowlands, M. Price, J. Maxey, McGraw-Hill, 2005
[Belbin, 2010] Team roles at Work, M. Belbin, Routledge, 2010

Authors' Information



*Dobromir DINEV, PhD student, New Bulgarian University, Sofia,
21 Montevideo St.,
Corpus II, Room 611,
dobromir@gmail.com*

Major Fields of Scientific Research: Software Quality Assurance, Artificial Intelligence

GENERATION OF VIRTUAL ANNOTATED CORPORA

Mariyana Raykova, Valentina Ivanova, Hristina Kostadinova
New Bulgarian University, Department of Informatics

***Abstract:** Course designers and instructors do their best to provide well-structured and motivating e-learning process by choosing the most appropriate e-materials and to create variety of digital activities of different types. One possible approach to reduce their efforts in the process of creating e-courses is presented in this paper. It is based on the idea of applying structured annotation of text corpora, using previously defined annotation elements, divided into four groups. The annotation is done by experts in linguistics and thousands of annotated text corpora is stored and analyzed. The set of annotated text corpora can be used to provide two important features: 1. to generate new text corpora, called virtual corpora, which can be used as learning materials in e-course and 2. to generate new test items of different types, which can be used to enrich question banks and to create online quizzes. The first feature will improve the quality of the e-course content and the second will enhance learners' assessment and self-assessment. Web-based annotation software system called MorphAnalyzer with separate modules for virtual corpora and test items generation is created and integrated into Moodle learning management system. The annotation system is tested and approbated in the foreign languages subject area.*

Keywords: *e-materials, e-assessment, question items autogeneration, text annotation, texts generation, corpora generation.*

ACM Classification Keywords: *A.0 General Literature - Conference proceedings (This is just an example, please use the correct category and subject descriptors for your submission. The ACM Computing Classification Scheme: <https://dl.acm.org/ccs/ccs.cfm?id=0&lid=0>*

Introduction

E-course consists of sequence of e-learning objects: materials (text, audio, video etc.) and digital activities (quizzes, participation in discussions, creating wikis etc.). Course designers put a lot of effort to provide well-structured and motivating e-learning by choosing the most appropriate e-materials and to create variety of digital activities of different types. Adding metadata to digital objects can increase efficiency in the process of selecting the most suitable ones according to the learning objectives and learner's preferences. Another important aspect in the process of creating well-designed digital courses is to provide clear rules for learners' assessment, including means to give students detailed feedback, according to their results. One of the most popular means for assessment is an e-learning environment is quiz which consists of questions of different types. The process of creating quizzes includes several basic steps: creating a question bank filled with test items of different types, difficulty levels and categories (in different topics).

Annotation is the process of adding notes to text document. Generally, annotated texts contain additional information about the meaning of the text, its paragraphs and sections. This additional data is in the form of comments of the text and gives information about the main topics covered in the text, including definitions of concepts, description of facts etc. It can be successfully used in the learning process, because most of the learning materials are in text form. Collected notes about the e-learning materials in text format can be used to improve e-

learning materials included in the e-course and increase motivation of the learners by constructing the e-content according to the learners' needs.

The annotation process cannot be conducted without appropriate software tools, which provide facilities for graphical notation of the text's parts (words, sentences, paragraphs, sections etc.). There are two main groups of annotation tools, which can be used to successfully to add comments to digital content, including texts. The tools are divided based on the annotation types:

Type 1. Unstructured annotation type: in this type of annotation the text comments are freeform and together with highlights of the text are placed on texts' parts. There are no rules for the comments types and there are no means to connect the commented text's parts based on the comments content. Most of the annotation tools provide interface for creating custom comments on web pages and imported text document formatted files [annotate 2018], [diigo 2018], [hypothesis 2018], [Scribble 2018]. This type of annotation is used in most of the learning management systems (LMS) as part of the learning process in order to do collaborative activities using annotation, or in the grading process when the teacher gives learners understandable feedback on the results of their assignments and other assessment activities [moodle annotation tool 2018], [blackboard inline grading 2018].

Type 2. Structured annotation type: this type of annotation is based on fixed form of the notes. The rules, which are used for the form of the comments are predefined and commented texts' parts can be connected to other texts' sections depending on the comment. This type of annotation is used for automation of the language processing and automatic interpretation of the text. The tools which are used to conduct the process of creating structured annotation provide means for creating list of comments and associate more than one text segment with

given comments which leads to associative relation between text's parts [brat 2018], [gate 2018].

In this paper we are focused on the second type of annotation and its usage in the field of e-learning process. An approach for automation of the process of creating digital materials in text format and test items of different types is presented below. This approach is based on idea of the text corpora structured annotation. A model and implementation of a web-based software system called MorphAnalyzer for text corpora annotation, which is used to provide the process of adding notes to text corpora is described in the next sections. The system is integrated in Moodle LMS to use it in the process of providing efficient learning in different subject areas, including foreign languages, and subjects learned in foreign language.

The research tend to achieve the following goals:

- Perform functionality for text annotation to develop a database of metadata for each text corpora.
- Provide means for generation of new text materials or sets of texts called corpora, which can be used as online materials in e-course in different subject areas.
- Provide means for generation of huge sets of question items for e-assessment, to perform accurate assessment.
- Provide means for generation of question banks from different type of test items (fill in the blanks, matching texts, matching images, reordering) in different subject areas.

Annotating Text Corpora

The annotation process is provided by a group of experts in linguistics by applying the structured annotation approach. Each corpus has four main annotation groups (created by experts in linguistics and discussed in [Stambolieva 2017a, Stambolieva 2017b]): 1. File options, 2. Part of speech, 3. Sentence categories and 4. Lemmas. These four groups consist of lists of elements of different types which are summarized in Table 1: The annotation

is conducted by choosing the text corpora and continuously selecting the elements listed in Table 1 and pointing part of the text corpora, which is an example of the selected item. For example, in the sentence “The book is about a young boy named Harry Potter.”, “book” is a noun, which is countable and neutral.

| Group 1. File Options | |
|--|---|
| 1. Corpus name; 2. File name; 3. Corpus description; 4. Text author; 5. Author gender; 6. Author status - author, coauthor, creator; 7. Date of generation; 8. Date from 9. Date to; 10. Domain (Subject Area) 11. Keywords or tags; 12. Language (Bulgarian and English are available); 13. Publication type (book, part of book, publication); 14. Number of pages; 15. Publisher; 16. Start page; 17. End page; 18. File type; 19. File level (A1, A2, etc.) | |
| Group 2. Part of speech, Prefix and Suffix | |
| Noun | Type: Common, Definite Article, Indefinite Article, Proper, Countability, Countable, Sg tantum, Pl. tantum, Gender, Masc., Fem., Neut. |
| Verb | Type: (FV), (LV), (ModV), (DelexV), (AuxV), /Aspect, (Perf), (Imperf), (State), (Process), (Event), /Transitivity, (Trans), (DiTrans), (Intrans), (CausTrans), (Erg), /Finiteness, (Finite), (PresPart), (PastPart), (AdvPart), (PassPart), (Inf), (Ger); |
| Adjective | Type: Qualitative, Relative, Num.&Quant, Degree, Positive Degree, Comparative Degree, Super; |
| Adverb | Type: Quantitative, Qualitative, Circumstance-Place, Circumstance-Time, Degree: Pos., Comp., Super. |

| | | | | | | |
|-----------------------------------|---|---------|---------|--------------|----------|---------|
| Preposition | Conjunction | Pronoun | Numeral | Interjection | Particle | Article |
| Group 3. Sentence Category | | | | | | |
| By structure | Simple, Complex, Compound, Complex-compound, Declarative, Interrogative, Imperative, Exclamatory. | | | | | |
| | Clause category, Syntactic phrase, part category, morphology category, No-subject, Term included, Term with certain definition included | | | | | |
| Group 4. Lemmas | | | | | | |

Table 1. Annotation Groups of Elements

The annotation is done by using a web-based software system, which is described in detail in the next sections of this work. Annotated text corpora can be used in two main fields, :1.to generate new text corpora, the so called **virtual corpora** which can be included in the e-course as learning materials and 2. as a **template for test items generation** which can be used in e-course assessment. These two applications are described in the next two sections of the paper. MorphAnalyzer is accomplished with different modules, which provide the generation of the virtual corpora and test items.

Virtual Corpora

As a result of the annotation of the different learning materials in text format, a set of over 150 corpora annotated and semantically analyzed corpora is collected. This set can be used to generate virtual corpora, according to different criteria. The purpose is to

collect already analyzed texts from different subject areas, and to be able to generate new texts on their bases.

After the process of the annotation is complete, each text corpus has a set of collected metadata and information about its segments (sections, paragraphs, sentenced, words etc.) following the above described four groups of annotation elements. This additional data attached to the corpus can be used to create new corpora. Course designer or an instructor in the e-course can select previously annotated corpora, based on different rules. The annotated corpora can be chosen by their keywords, subject area, difficulty level etc. and when they are combined all the metadata from the original text corpora will be attached to the new one. The new text corpus, which is created by the combination of the original ones is called **virtual corpus**. The annotation software system includes a module for virtual corpus creation, which provides means for selection of the annotated corpora, filtering them by different criteria. The following options are available when the virtual corpus is created: search for already generated virtual corpora, check the list of all virtual corpora, view, edit, delete, copy a selected virtual corpus or generate a new one by different criteria. The process of creating virtual corpora consists of several steps, which are possible to be done by the means of the annotation software system:

Step 1. Select generate new virtual corpus.

Step 2. List the annotated corpora.

Step 3. Filter annotated corpora, according to the different criteria by using appropriate search method to list only the most appropriate annotated corpora. MorphAnalyzer provide means to use search by the four groups and their elements in the annotation structure, described in Table 1:

- Search by file options – all sentences with the selected options will be listed and can be chosen.
- Search by Lemma (form of a set of words) – by choosing a word, all the sentences which contain its forms will be listed.
- Search by part of speech - search for sentences, that contain terms together with their definitions.

- Search by sentence category – category of the sentence can be one of the listed in group 3 in Table 1.

Two options are available in the annotation system: all criteria together (AND search) or any of the criteria (OR search). If any value for some of all subforms' fields is presented, only the sentences that cover this rule will be listed. If nothing is filled, then all sentences from all files and corpora will be listed.

The result list of found sentences is presented by sentence identification number (id), sentence title, file identification title, that contains that sentence, file name, which is a link to the file full content. The list is sortable by any of the shown fields.

Step 4. Select the corpora from the list to include them in the virtual corpus.

Step 5. Edit virtual corpus content, if needed: add, remove or edit sentences to the virtual corpus.

Step 6. Edit metadata about the virtual corpus, if needed

Step 7. Save the content of the virtual corpus.

On Figure 1. is presented a screenshot of the virtual corpora module, where a list of the annotated corpora, together with the different options for filtering them is displayed.

After the process of creating virtual corpora is complete, the new text with its metadata and segments' annotation can be used in the process of e-learning. It can be included in e-course to improve the digital content. Its "quality" can be measured by analyzing results of the learners who examined this material with the results of the learners who did not. The above described method of generating virtual corpora will support course designers and instructors to easily improve e-course content. The annotation system together with the virtual corpora module is efficiently used in teaching foreign languages and different subject areas such as geography, history, biology etc. in foreign language. The structured annotation rules give a framework which can be enlarged in order to be developed in other subject areas.

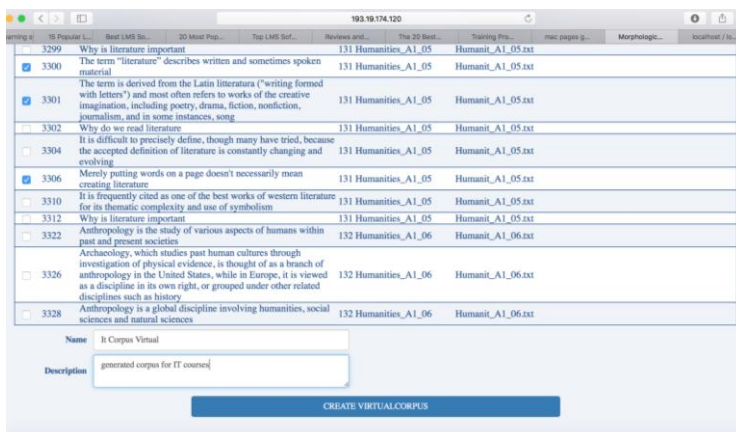


Figure 1. Creating Virtual Corpus

Auto Generating Test Items of Different Types

In the previous section is described how virtual corpus (combination of annotated text corpora) can be created and used in the process of e-learning. Another important usage of the set of stored virtual corpora, together with their annotation, is the possibility to generate test items of different types. This will lead to enriching and enlarging the question banks which are used for the creation of the quizzes in the e-course. Generated test items can be exported Moodle e-Learning system, where they will be included in question banks and quizzes.

MorphAnalyzer's test items generation module can automatically create three types of test items: fill in the blanks with variations, matching and reordering.

Creating Fill in the Blanks Test Item

Course designer can use the software module for test item generation to generate a new fill in the blanks question type. The following steps must be done:

- Step 1. Search for corpus, including virtual corpora (virtual corpora are labeled).

Step 2. Choose the corpus. MorphAnalyzer loads all sentences or paragraphs from the selected corpus (Figure 2).

Step 3. Select sentences. MorphAnalyzer displays the sentences with their annotation, lemmatization and analysis marked on them (Figure 3).

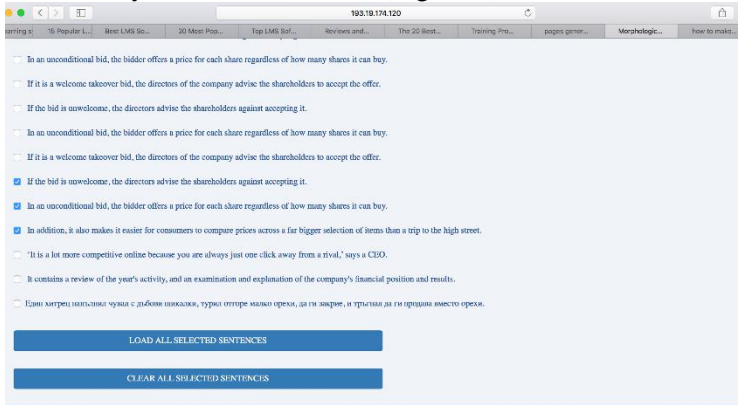


Figure 2. Selecting sentences for test item generation

Step 4. Set name and category of the question, if category is not set, LMS will automatically create a new one with the same name and will store the question there.

Step 5. Hide different part of speech, which are selected from a list with all parts. Which occurrences of the found words will be hidden or whether to hide prefixes or suffixes, is a matter of decision.

Step 6. Add more distractors to the list of autogenerated ones.

Step 7. Use only the base forms of the words, or the form they have in the text. Only part of speech and occurrence options are obligatory.



Figure 3. Annotated sentences for test item generation

On Fig. 4 is displayed generated fill in the blanks test item. The words in the brackets [] will be hidden from the learners, the words in the parentheses () will be used as help words. They serve as base form of the blanks. The creator of the question bank can select whether to generate fill in the blanks question, that can be answered with drag and drop functionality. The correct word can be dragged and dropped to the correct place.

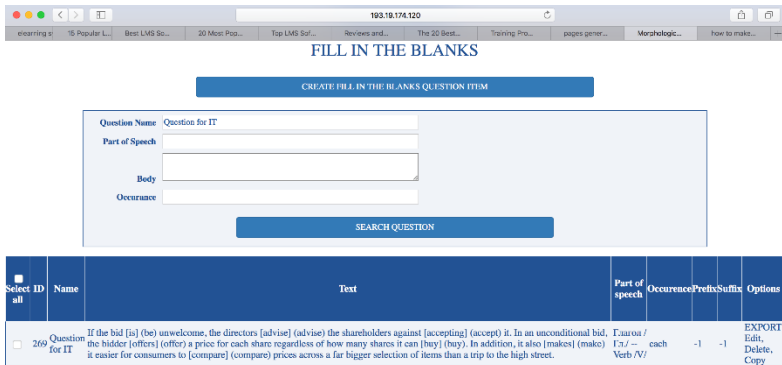


Figure 4. Fill in the Blanks Test Item Generation

Other two types of question items, can be generated analogously by using appropriate graphic interface, depending on each of their specific characteristics. MorphAnalyzer supports matching and reordering question types.

For the question type matching, it is possible to match terms with their definitions or image with a text. Again, after a search in the set of corpora or texts, the expert can select terms that he wants to be shown in the question. The definitions are extracted from the database after the annotation process. The expert has the following possibilities for the type – matching images with definitions, images with terms and term with definition.

The type reordering is working on the base of sentences or paragraphs. The experts select corpora or text files after applying filters and after that the system lists the paragraph/sentences (depending on a selection), from which the expert is making a subset. The generated subset is used in the question.

MorphAnalyzer – System Model

MorphAnalyzer is designed and developed using the LAMP stack because this stack is a set of open-source software technologies that are widely used for creation of web-based software and applications. In that stack a CodeIgniter [CodeIgniter] framework was chosen as very light and easy to use framework, which supports the MVC model according to which the system was designed. Virtual corpora generation part of MorphAnalyzer’s model is shown on Fig. 5. All the annotated corpora with its elements in different subject areas are stored in the database and can be used for virtual corpora generation.

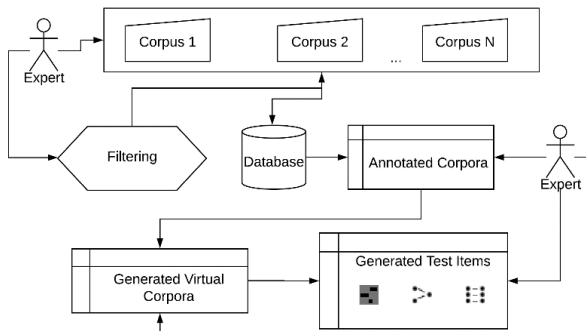


Figure 5. MorphAnalyzer: Virtual Corpus Generator Model

A virtual corpus will be used in the system as any other corpus and there is no need the user to know which corpus is virtual and which was created by the user himself. That is why after generation of such a corpus it should be usable everywhere in MorphAnalyzer. That means that these corpora can still be more deeply analyzed or according to the context their annotation properties and values can be changed. The implementation is done in a form of a plugin which has the following classes:

- class Filters generates the possible filter that a user can apply to restrict the set of text items that will be used to construct the new corpus. That class is responsible for the generation of web form in which the user will set values for the restriction. The class will be fed from a config file where all possible type of annotations for a text/corpus are listed. Each element from the annotation is described with label and type. The possible types are – input field, select, radio, checkbox, text area, predefined list of values, module from the database.
- class TextItemSets gets as an argument a set of all selected or entered data from the previous class. It will be used to search in the database for all sentences (the smallest item in a corpus) and to select only these of them which meet the search criteria from class Filters.
- class VirtualCorpus extends Corpus is a derived of the base class Corpus and it will extend the functionality of a corpus by adding member functions which will generate a corpus from a set of sentences. This class will provide facilities to integrate the virtual corpus as an ordinary corpus in the test item generation module. The expert should not find any difference between the two types of corpora – virtual or ordinary. This class is like a “shield” in the system, as its design is developed without changing the existing modules in MorphAnalyzer.

One additional table is added into the database, because most of the work is done by the VirtualCorpus class. This new table is virtual_corpora table, which stores the ids and types of all

elements that construct a virtual corpus. Except that, we need all the data for a corpus that is used for the filtering. That is why after a generation of a virtual corpus the expert will be able to add some additional annotation information, which is for the ordinary corpora. The new module will add automatically a new line in the corpus table and additional information will be added into the virtual_corpora – with which corpus id from the corpus table is connected the virtual corpus. Each virtual corpus is treated as an ordinary corpus, but there is additional information stored about it. Each corpus is constructed from one or more files, the system will generate a new text file that contains all texts selected by the expert.

The MorphAnalyzer, is designed as an autonomous web based system independent from any other systems. MorphAnalyzer is used by experts in the area of foreign languages, who will annotated different text for different languages and the system will be able to generate question items from different types. MorphAnalyzer does not support any modules for assessment. What the system supports is generation of question items/s or question banks. That is why we designed an export functionality in order to feed the generated question items or question banks into different LMS systems. As a standard solution for the integration between MorphAnalyzer and other LMS, we choose the export in XML files, which later can be used to be imported where the user needs. The system supports two formats of XML export – Moodle and according to the IMS QTI standard [IMS QTI 2018].

MorphAnalyzer Integration with LMS

The web-based annotation system is fully integrated and connected with the LMS Moodle, as it is one of the most popular open-source LMS. The module for the test items generation is integrated into the Moodle's Quiz module and the user will continue to use it without any change.

Several important features can be used within the new test item generation:

- Select questions to import.
- Download export xml file from the MorphAnalyzer.
- Import the file in Moodle question bank.
- Edit the questions if needed.
- Use the questions in a test via Moodle's modules for question banks and test or quiz creation.

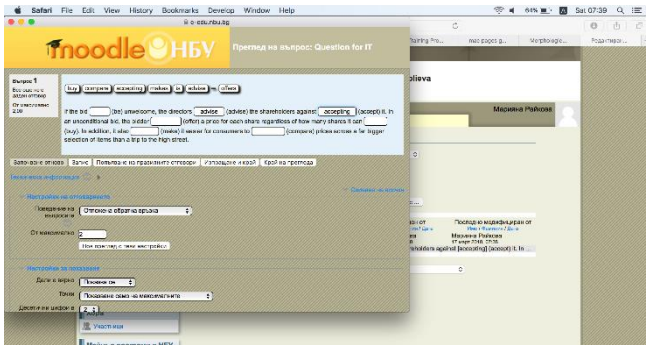


Figure 5. Test Item Generator Integrated in Moodle LMS

On Fig. 5 is shown a real question item that is generated from MorphAnalyzer and is imported in Moodle. The subject area for that particular question is Foreign Languages for students from different programs.

Conclusion

In this paper is presented an approach to improve the e-course content and automate the process of creating question banks, used in the quiz assessment activity. The method is based on the idea of applying structured annotation of text corpora, using previously defined annotation elements, divided into four groups. The annotation is done by experts in linguistics and thousands of annotated text corpora is stored and analyzed. The set of annotated text corpora can be used to provide two important features: 1. to generate new text corpora, called virtual corpora,

which can be used as learning materials in e-course and 2. to generate new test items of different types, which can be used to enrich question banks and to create online quizzes. The first feature will improve the quality of the e-course content and the second will enhance learners' assessment and self-assessment. Web-based annotation software system called MorphAnalyzer with separate modules for virtual corpora and test items generation is created and integrated into Moodle learning management system. The annotation system is tested and approved in the foreign languages subject area.

Still in that approach we have a broad field of researches. Possible developments of new functionality, which will lead to a system with better quality and more options for personalization and automation are:

- Integration of deep data methods.
- Integration of neural nets.
- Integration of computer linguistic methods.
- Development of plugin for connect maps.
- Integration with WordNet, FrameNet and VerbNet systems.
- Generation of question items from different types - true/false, essay, short answer with possibilities for auto assessment, number, compound etc.
- Development of algorithms to measure the quality of the generated question items.
- Integration of different methods and algorithms for auto determination of the knowledge level of each question item according to Bloom's taxonomy.

Bibliography

[annotate 2018] <https://annotate.net/> last seen on the 6th of June 2018

[blackboard inline grading 2018]

- https://help.blackboard.com/Learn/Instructor/Assignments/Grade_Assignments/Assignment_Inline_Grading
- [brat 2018] <http://brat.nlplab.org/> last seen on the 6th of June 2018
- [CodeIgniter] CodeIgniter Web Framework,
<https://codeigniter.com/> last seen on the 11th of June 2018
- [diigo 2018] <https://www.diigo.com/> last seen on the 6th of June 2018
- [hypothesis 2018] <https://web.hypothes.is/> last seen on the 6th of June 2018
- [IMS QTI] IMS Question & Test Interoperability Specification Overview | IMS Global Learning Consortium,
<https://www.imsglobal.org/question/index.html>, last seen on the 13th of June 2018
- [gate 2018] <https://gate.ac.uk/> last seen on the 6th of June 2018
- [moodle annotation tool 2018]
<https://www.annotate.co/moodle.html>
- [scribble 2018] <https://www.scribble.com/> last seen on the 6th of June 2018
- [Stambolieva 2017a] Maria Stambolieva, Milka Hadjikoteva, Mariya Neykova, Mariyana Raykova, Valentina Ivanova, The Nbu E-Platform In Teaching Foreign Languages For Specific Purposes , 13th Annual International Conference On Computer Science And Education In Computer Science 2017, June 30 to July 3, 2017, p. 203-218, Albena Bulgaria
- [Stambolieva 2017b] Maria Stambolieva, Milka Hadjikoteva, Mariya Neykova, Mariyana Raykova, Valentina Ivanova, Language Technologies in Teaching Bulgarian at Primary and Secondary School Level: the NBU Platform for Language Teaching (PLT), Language Technology for Digital Humanities in Central and (South-) Eastern Europe (LT4DH-CEE) Sept, September 8th, 2017, Varna, Bulgaria
- [Raykova 2017a] Mariyana Raykova, Valentina Ivanova, Nbu Eplatform In Teaching Foreign Languages For Specific Purposes – Software Engineering Challenges, 13th Annual International Conference On Computer Science And Education In Computer Science 2017, June 30 to July 3, 2017, p.21-36, Albena, Bulgaria

- [Raykova 2016] Mariyana Raykova, Hristina Kostadinova, George Totkov, Towards Automated e-Course Creation, 6th National Conference on e-Learning in Higher Education Institution, p. 216-225, 2-5 June 2016, Kiten, Bulgaria (in Bulgarian)
- [Raykova 2015] Mariyana Raykova, George Totkov, Automated Test Items Generation based on e-Learning Materials, Scientific Works of Union of Bulgarian Scientists – Plovdiv, p. 62-66, Union of Bulgarian Scientists, 5 - 6 November 2015

Authors' Information



Mariyana RAYKOVA, Ph.D. Chief Assistant Professor, New Bulgarian University, Informatics Department, 21 Montevideo Str., 1618 Sofia, Bulgaria, mariana_sokolova@yahoo.com.

Major Fields of Scientific Research: e-Learning, automatic test generation, programming

Valentina Ivanova, Ph.D. Chief Assistant Professor, New Bulgarian University, Informatics Department, 21 Montevideo Str., 1618 Sofia, Bulgaria, v.ivanova@gmail.com.

Major Fields of Scientific Research: PM, computer linguistic, ...



Hristina Kostadinova, Ph.D. Chief Assistant Professor, New Bulgarian University, Informatics Department, 21 Montevideo Str., 1618 Sofia, Bulgaria, hkostadinova@gmail.com.

Major Fields of Scientific Research: e-Learning, adaptive learning, adaptive test systems

CSECS 2018, pp. 211 - 228

Computer Science and Education in Computer Science
14th Annual International Conference
ISSN 2603-4794

June 29 – 30, 2018, Boston, USA

HOW TO IMPROVE TEACHING IN DISCRETE MATHEMATICS VIA PROGRAMMING AND VICE VERSA

Mariyana Raykova, Stoyan Boev

New Bulgarian University, Department of Informatics

***Abstract:** Nowadays learning mathematics in the university seems like a big challenge, taking in mind the decreasing level of mathematical skills from secondary education and low study motivation. But there is no doubt that learning Computer Science is going hand by hand with learning mathematics generally and Discrete mathematics in particular. In this article we try to propose a possible solution of that problem. As students that are studying Informatics at New Bulgarian University constantly ask as the question “Why should we learn mathematics?”, we came to the idea of putting Discrete Mathematics and Programming together in order to show them some of the reason. The most interesting point in that the students pretend to learn more programming on account of maths is that, their results in maths are much better, than in programming. So we had one more problem to solve. The first problem was how to increase the interest in learning mathematics, and the second was how to improve their programming skills. In order to solve these two problems (and not only them), we proposed a new course in our programs of Informatics and Information Technologies called*

“Computer labs in discrete mathematics”. In that course we are teaching some basic discrete structures and one of the most popular algorithms from discrete mathematics through programming with C++.

Keywords: *discrete mathematics, programming, algorithms, data structures, discrete structures, sets, number theory, modular arithmetic, combinatorics.*

ACM Classification Keywords: • *Mathematics of computing~Permutations and combinations*

Introduction

Nowadays learning mathematics in the university seems like a big challenge, taking in mind the decreasing level of mathematical skills from secondary education and low study motivation. In New Bulgarian University (NBU), students have mathematics in their schedule if they study Economics or Informatics. What are the problems we have faced during teaching of discrete mathematics and programming in Informatics? First of all following the program scheme in the university, we have 30 learning hours in class for each of these two courses, but they are extremely not enough, for all the concepts that they need to study in class. The main problem comes from that these courses are in the schedule for the first year students and first term. These students are still pupils. They do not have the abilities to learn in university community. They still tend to wait to learn all the materials in class and something more to make all the exercises in class together with the teacher. This is not the reality in a university. According to European Credit Transfer and Accumulation System (ECTS), for each 30 hours in class each

student should spend 60 hours exercising by him/herself without a teacher.

The set of concepts that should be learned by the students according to the course plans are too ambitious, the concepts and relations are too many as numbers, and the teacher is obligate to switch from one to another, without spending enough time for the different concepts and the most important thing - the relations between the concepts.

But there is no doubt that learning Computer Science is going hand by hand with learning mathematics generally and discrete mathematics in particular. In this article we try to propose a possible solution of that problem by putting them together. The lack of enough learning hours in class leads to the problem that the teacher has no possibility to show ~~his/her~~ students different algorithms from discrete mathematics that are good examples of real application for solving common problems from the real world. That leads to a problem that students are not able to understand the benefits and consistency of these courses.

Another problem for the courses in discrete mathematics and programming is that they cover completely new set of concepts that students have not met till that moment. The students (in very few exceptions) do not possess the skills to think discretely - step by step. They do not know how to solve problems in that way, but here it is very important to be able to. For example: to divide a problem to subproblems, to sort that set of problems, and to start the solution from the most trivial one. The main issue here is that it is very hard for the student to reset their way of thinking. Discrete mathematics will help them to start thinking in discrete terms and structures, which is fundamental for programming.

The most interesting point in that the students pretend to learn more programming on account of maths is that, their results in maths are much better, than in programming. So the first problem is how to increase the interest in learning mathematics, and the second is how to improve their programming. In order to solve these two problems (and not only them), we proposed a new course in the programs of Informatics and Information Technologies called “Computer labs in discrete mathematics”. In that course we are teaching some basic discrete structures and one of the most popular algorithms from discrete mathematics through programming with C++, in order to get rid of the problems that were listed, or just to some of them.

Course design

In the scheme for first grade students first term they standardly have 30 hours course for basic mathematics and 30 hours course for programming. As we already mentioned this workload is completely not enough to cover all the material that is needed.

For example in the separate course “Discrete mathematics” they learn concepts like sets, permutations, combinations, etc., they learn how to operate with them, but they do not have enough time to learn how to present a set in the memory of a computer, how to effectively perform operations with sets, what is the connection between sets and boolean functions or how to generate permutations and combinations. Also students are not able to transfer that knowledge in real life and to start using it for solving real problems. For example if they have a problem like this:

“At the end of the first semester, first grade students in Informatics at New Bulgarian University have gone to the

mountain. We know that the number of all students in the group is X , the number of students for skiing is Y (possibly along with snowboarding), the number of students for snowboarding is Z (possibly along with skiing), and the number of students for both is Q . There are students in the group that can neither skiing nor snowboarding. The values for X , Y , Z and Q are entered from the console by the user, and the program should check if $X > Y + (Z - Q)$ is true.

Write a program that asks the user to enter values for X , Y , Z and Q . The program should find how many of the students are able only to ski. Define all possible functions for the program realization.

Input: $X = 24$, $Y = 17$, $Z = 9$, $Q = 5$

Output: 12“

The students find difficulty in creating mathematical model of the problem - that we have one main set X , two subsets Y and Z , intersection of these subsets that is marked like Q . The students just need to find the cardinality of the result set that is the difference between Y and Q . A trivial problem for set operations.

Something more, when they learn discrete mathematics and programming they learn set operations and bitwise operations, but don't realize deeply the connection between them, based on the abstract discrete structure - Boolean algebra. In fact the operations over sets like union, intersection and complement and the boolean operations like disjunction, conjunction and negation are equivalent with respect to the boolean algebra. This problem floats on the ground with more power when they have to learn database operations and SQL language in higher semesters of

their study. So if we face such problems in lower semesters, the students will be more motivated and will learn easily later.

Some other examples are listed below:

- The concept for permutation is studied beyond the concept of the discrete structure called symmetric group S_n . So the students know what a permutation is and how to count permutations, but can not operate with, generate and use them.
- The concept of operations by modulus n is studied beyond the concept of the discrete structure called ring of integers modulo n . So the students know how to find the remainder using the quotient remainder theorem, how to operate in the field of real numbers R – addition, subtraction, multiplication and division, but could not apply the same operations in the field of integers modulo a prime number F_p (a fundamental knowledge for cryptography algorithms).

At the beginning of their study, students think that mathematics is hard for studying. In fact they face more problems in programming. As in first term they should be able to develop skills for applying information in practice, not only for remembering it. Also they should be able to apply information from more than one subject area. Our goal in first semester is just to teach them the syntax of a programming language, together with applying this syntax with discrete mathematics. The problems that we face generally are in some of the base concepts for:

- Bitwise operators – switching to calculations from decimal to binary values is an issue;

- Efficient code development – students tend to try just to write the program that, they are asked for. They do not search for second, third possible solutions, that will be more memory or time efficient;
- Algorithms to solve common tasks – there is a lack of knowledge of different standard algorithms for solving common problems;
- Data structures – most of them have never studied informatics at school, so main concepts like that are unknown, which is an issue for some of the algorithms;
- Set operations and Boolean algebra are used in a lot of the algorithms and in database courses;
- Performing operations in the memory is a big issue for students, as in C++, they have to allocate and release memory by themselves. This leads to a lot of source code that is not secure, memory leak or unknown exceptions;
- Solving real problems with programming – development of models from real life that are appropriate to be programmed;
- Conditional operators and Boolean expressions;
- Loops – nested loops, different kind of loops, using endless loops, loops for checking user data, etc.;
- Recursion – understanding the mechanism of self-call of a function and wrapping and unwrapping function frame stack, etc.

In order to solve these problems we propose a new course in the program scheme for the first year, first semester. The course name is “Computer labs in discrete mathematics and programming”. The course is conducted in computer labs and have 30 hours workload.

The purpose of the course is to:

- increase the workload in mathematics and programming;
- to show the students the real appliance of discrete mathematics in programming and real life;
- to help them with more exercises in both courses;
- to increase their interest in mathematics and to exercise their knowledge in programming;
- to increase the grades in discrete mathematics and programming;
- to help them to be able to think algorithmically;
- to be able to write efficient and optimized source code.

It is very important for the new course, the same teachers that lead the separate courses to lead that course too. This is so important, because it is needed to synchronize the material that is taken in both courses separately and together in the new one. This somehow will help the work in both separate courses and the new one too.

The new course will help teachers to find out all the problems that they have in their separate courses. These problems are not only connected with the students themselves, but also with the teachers. For example when in programming they have to work

with bitwise operations and mask they have difficulties and are not able to apply their knowledge from mathematics.

As in Informatics department at New Bulgarian University there are two bachelor programs learning programming - Informatics and Information Technologies there are two main teams one for Informatics and one for Information Technologies. There is no standard in the university for teaching a certain course in the different programs, so we are going to prepare the new course for the students studying Informatics only.

The issue when creating the course passport were not only the problems that the teacher and students have listed above, but also connected with the content. When we started to outline the different themes week by week, we came to the point that when they need to study recursion algorithms in discrete mathematics, they hadn't learned functions in the programming course. When they need to save different data structure in the memory for representing sets and operations over them, they still haven't learned arrays, two dimensional arrays, dynamic arrays, or structures and classes.

So the main issue was to find a way to prepare the course content in a way that it is possible everything learned in discrete maths to be exercised in computer labs via programming. So for each learned algorithm, we should write a program.

After taking into account the problems and syncing the materials for both of the courses we came to the following plan:

Part I – Number theory (6 hours math labs + 8 hours programming labs):

- Division – the quotient remainder theorem, numeral systems, prime numbers, fundamental theorem of arithmetic.
- The greatest common divisor (GCD) and the least common multiple (LCM). The extended Euclidean algorithm.
- Modular arithmetic – basic properties of congruence relation, operations in the field of integers modulo a prime number F_p .
- Classical applications of modular arithmetic in coding theory – affine ciphers, ISBN code.

Part II – Combinatorics (6 hours math labs + 8 hours programming labs):

- Subsets – representing via bitmasks, set operations.
- Subsets - counting, generating, coding.
- Permutations – counting, generating, coding.
- Combinations – counting, generating, coding.

Covering the course the students should know:

- Base programming concepts with C/C++ languages: input/output operations, base data types and operations, structures, expressions, pointers, arrays and user defined functions;
- Basic procedural programming techniques: iteration, recursion, and dynamic memory;
- Basic applications of modular arithmetic in computer science and coding theory.

- Basic combinatorial configurations and their implementation on programming language C

Covering the course the students should be able to implement effective algorithms for :

- finding GCD and LCM of two integers;
- prime number check and generation of prime numbers;
- number system conversion;
- representing and operating with sets;
- generating and coding of subsets;
- generating and coding of permutations;
- generating and coding of combinations.

It is a mandatory for every student to be subscribed and visit at the same time with the new course, the courses “Discrete mathematics” and “Programming”, together with “Computer labs in programming”.

Conducting the course

In NBU there is a standard for course schedule. Each semester is lasting 15 weeks. Through out the semester the students should have at least two current controls. The grades from the current control can be used as final grade of the course, if they are decided as enough by the teacher. For this course we put the minimal grade 4 (grade scheme from 2 to 6) for passing the course with current control. We had two current controls. In order to perform the final grade we gave the students additional half

grade if they had visited the course regularly, and one more time additional half grade if they had done their homework.

One of the main reasons for part of the groups to be a little bit late with the schedule, is that they were studding programming with different teacher, and using different materials, so we had lack of sync for them.

In order to sync the materials in both separate courses as far as possible, we made the following plan:

1. Three weeks for labs in discrete mathematics in order to be introduced different algorithms from the first part of the course - number theory;
2. Four weeks for programming labs, in order to get the appropriate material in programming, so they should be able in that stage to implement the algorithms that they have learned there weeks ago;
3. One additional lecture (out of the schedule) for preparation for the first control.
4. First control was an auto graded online test. The students did that test from home, and by this way we had one more additional week for learning new material.
5. Three weeks for labs in discrete mathematics in order to be introduced different algorithms from the second part of the course - combinatory;
6. Three weeks for programming labs in order to be implemented the algorithms from this part of the course;
7. The last week was the second control, which was online attendance test.

Both tests had some practical problems from discrete mathematics, and some problems that need programming skills,

but always the content of the tasks is was connected with the subject area of discrete mathematics.

The first test is was purely online, and it contained 15 only multiple answer questions - 10 from maths, 5 from programming. The mathematical questions concerned the corresponding discrete algorithms from number theory like the extended Euclidean algorithm, prime factorization or ISBN code. The questions for programming were needed to be read with understanding the proposed code and to be predicted what will be the result after execution of the code. Some sample questions are given below:

1. What is the next term in the following sequence: 10, 11, 101, 111, 1011, 1101, ...?
2. How many are the positive divisors of 450?
3. How many steps are needed for Euclidean algorithm to determine the GCD of 15 and 69?
4. What is the missing digit in the following ISBN-10 (International Standard Book Number) 0- 19-85□803-0?
5. Determine the result of the following code snippet:

```
unsigned int n=10, k=1, p=k, s=0;
    for (int i=3; i<=n; i++) {
        s = p+k;
        k = p;
        p = s;
    }
    cout << s << endl;
```

6. Find the errors in the code that solves the Eratosthenic Solution problem:

```

int n = 20;
int arr[] = {0};
int i = 0, j;
for(;i<n; i)
    if(arr[i]==0){
        arr[i] = i;
        for (j = i*i;j<=n; j+=i)
            arr[j] = -1;
    }

```

The second test was online attendance one, and it contained 5 mathematical questions, which concerned the corresponding discrete algorithms for set operations using bitmask, generating of permutations and combinations. The problems for programming were 5 also. Some sample questions are given below:

1. Let $U = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ be an universal set, $A = \{1, 3, 5, 7\}$ and $B = \{2, 3, 5, 6, 8\}$. Determine the bitmask of the set $(A \setminus B) \cup (A \cap B)$.
2. Write a C ++ function that takes as arguments an universal set of symbols and a random set of symbols. The function determines how many deletions of the elements of the second set should be made to reduce it to a subset of the universal one.
 Input:
 a b c e // universal set
 a x f b c d a r w h //random set
 Output: 7
3. Write C ++ functions that collect and subtract fractions by presenting the result in the form of a simple fraction.
 Input: $1/2 + 1/3$ Output: $5/6$; Input: $2/3 - 4/5$ Output: $-2/15$

It's very important to say that one of our goals was to show students different ways for solving a problem - effective one in regard to time and memory and easy one in terms of coding. Very typical examples are primary check, set operations and generating of k-combinations.

Results

We had already one term for conducting the course. In that course we had 122 students from Informatics program. The diagram on Figure 1 represents the comparison between the student results on the first and second test respectively.

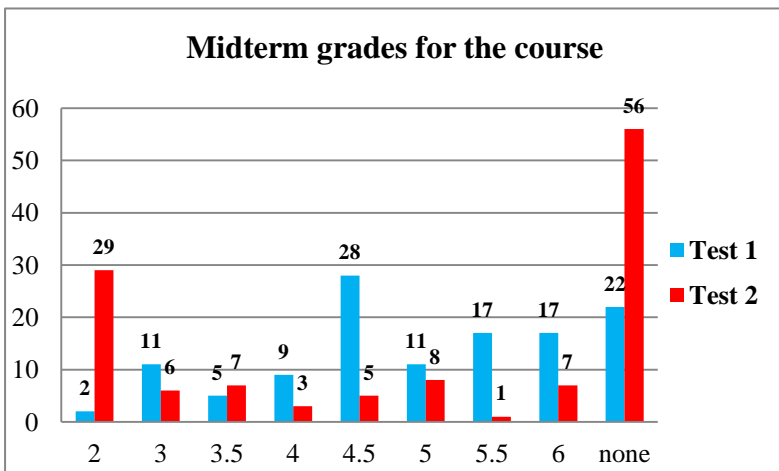


Figure 1. Midterm grades for Test 1 and Test 2

The first online test was done from 82% of the students with average grade of 4.67 (out of 6) while the second online attendance test was done from about 54% with average grade if 3.37 (out of 6) (Fig. 1). For the first test the distribution obviously is much better than on the second test (in respect to normal distribution). Here are some of the possible reasons for that:

- the first test checks for more theoretical knowledge of some basic discrete algorithms than programming skills. It confirms our initial expectations that the students find programming much more difficult than mathematics.
- the first test was online and this allows the students to work in groups although the test questions were generated automatically.

The following diagram shows the final grades for the course compared with the student attendance to the course.

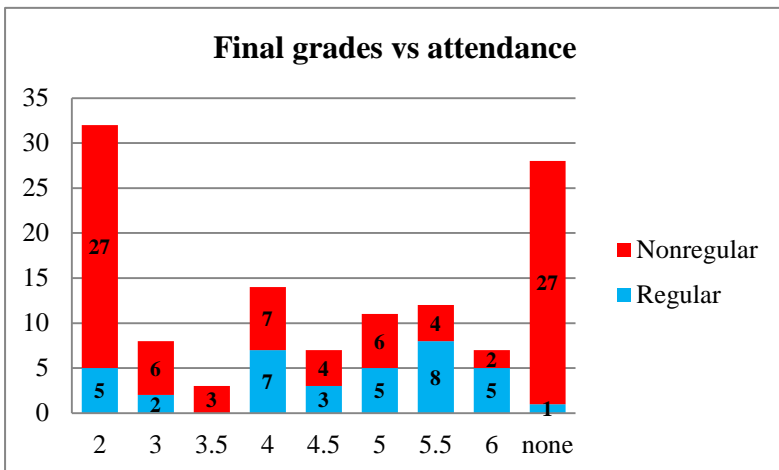


Figure 2. Final grades vs attendance

We see that exactly 50% of the students have passed with average result 4.62 (Fig. 2). The students with regular attendance to lectures and labs are about 30% of all and 95% of them have passed with average result 4.70.

Conclusion and future work

It's very difficult to measure the usefulness of this interdisciplinary approach but some very sure things could be said:

1. This course **increase the learning hours** with 12 for discrete mathematics and 18 for programming
2. It **enhances understanding** in the subject area of discrete mathematics and programming via considering and implementing appropriate algorithms.
3. The students are **more interested in mathematics**, as they are able to see different aspects of applications.
4. The students start to **think discretely and abstractly** by getting familiar with different discrete structures and applying the same operations on different type of sets
5. The students start to think not only for implementation but for **effective ways to code and apply algorithms and operations**, including source code optimisation, memory usage optimisation, time consuming optimisation;
6. Teaching the material in this way we **remove the borders** between the different courses, and show the students that there is visible connection between the different subject areas;

7. Finding the optimal way to implement and operate with different discrete structure is very important for our team for **competition programming**.
8. Last but not least having more knowledge and deeper understanding the students could be able to **increase their grades in both subject areas**.

Although we increase the workload of the course discrete mathematics with 12 hours and the course programming with 18 hours, it is **still not enough and some core themes were not taken**.

If the course is in two terms or it is with double workload, it will be good idea to include topics for automata, syntax trees, graphs, hash functions, etc.

Authors' Information



Mariyana RAYKOVA, Ph.D. Chief Assistant Professor, New Bulgarian University, Informatics Department, 21 Montevideo Str., 1618 Sofia, Bulgaria, mariana_sokolova@yahoo.com. Major Fields of Scientific Research: e-Learning, automatic test generation, programming



Stoyan Boev, Ph.D. Chief Assistant Professor, New Bulgarian University, Informatics Department, 21 Montevideo Str., 1618 Sofia, Bulgaria, stoyan@nbu.bg. Major Fields of Scientific Research: Coding Theory, Automata and Computability, Synthetic Geometry

CSECS 2018, pp. 229 - 249

Computer Science and Education in Computer Science
14th Annual International Conference
ISSN 2603-4794

June 29 – 30, 2018, Boston, USA

EARLY DETECTION OF FOREST FIRES - STANDARD INTERFACES AND PROTOCOLS AT SENSOR NETWORK AND CLOUD LEVEL DEFINITION

**Jugoslav Achkoski, Nikola Kletnikov, Nevena Serafimova,
Igorce Karafilovski, Rossitza Goleva, Katerina Zlatanovska**

***Abstract:** In this paper we presented full design of the system for monitoring forest which consists of cloud platform, sensor networks and mobile (drone) technologies for data collection and cameras. We first present the advanced design and structural model of an advanced system for monitoring of forest area. This model integrate sensor networks and mobile (drone) technologies for data collection and acquisition of those data at existing Crisis Management Information Systems (CMIS). Then we demonstrate the possibility to map different technological solutions and the main result was the definition of the set of standard interfaces and protocols for network interoperability.*

***Keywords:** Wireless Sensor Networks, Forest Fire Detection, standard interfaces, Communication protocol*

1. Introduction

A couple of years ago, the framework of “integrated system for prevention and early detection of forest fires”, “Macedonian Forest Fire Information system – MKFFIS”, was established in FYROM. But an integrated wildfire prevention and management system should provide instant access to functionalities such as active fire detection, fire danger forecasts, fire behaviour prediction, access to historical fire data and fire damage assessment and MKFFIS does not live up to these expectations.

With that being said, the goal of the ASPires project is to develop advanced concepts for early detection systems of forest. The integration of ASPires with MKFFIS will allow the users of MKFFIS to be able to access the active alarms and read out their values and therefore be able to react accordingly. Through MKFFIS all of this data will be practically distributed to all the institutions that form the General Crisis Management System which enables each of the to react in the limits of their authority.

Early detection of forest fires using wireless sensor networks (WSNs) should integrate sensor networks and mobile (drone) technologies for data collection and acquisition of those data in existing Crisis Management Information Systems (CMIS). The mobile (drone) technologies will allow to cover much larger areas to raise the percentage of forest fire detections in areas of importance, to monitor areas with high fire weather index, and to

monitor areas already affected by forest fires. Moreover, the use of wireless sensor networks distinguishes itself as the method that provides the best results in relation to the general costs of the equipment used, and even more importantly, it is the most reliable method of locating a forest fire in real time with the smallest variations in distance from the site of the fire.

For make this detection system functional standard interfaces and protocols at sensor network and cloud level are defined and presented in this paper.

2. 2. Early detection of forest fires using wireless sensor networks

The design of the forest fire detection is defined as a comprehensive solution to environmental measurement, based on information gathered by a wireless sensor network, allows a detailed and centralized analysis of possible fires from starting in the area to be monitored. The key features associated with the service are:

- Capturing information from sensors deployed depends on the line of sight.
- Information sent through the local wireless network using ZigBee.
- Received information processing based on models of prediction and approximation algorithms.
- Presentation centralized alarms, captured data and predictions obtained from each region through a base station will be located where a park ranger will report at

the time there is a fire, through its expertise in radio to handle this event.

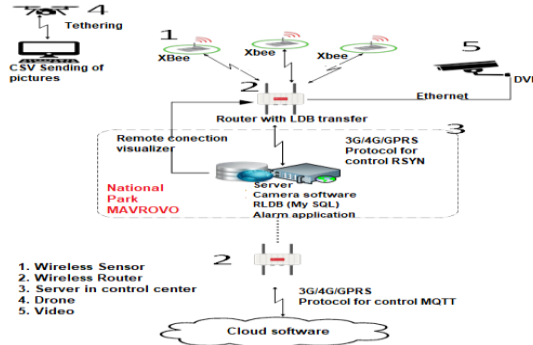


Figure 1. Design of an advanced system for monitoring of forest area

The primary product or module of ASPires (Figure 1.), includes the placement of sensors at previously specified locations (pointed out by experts coming from the National Parks (NP)) which are used for constant monitoring of the forest areas of interest, which in term are paired with a working algorithm that measures and compares the critical parameter values, and upon the exceeded thresholds of some or all of these values, reacts by activating the initial fire alarm. Depending on the location of the sensors, a flag (or some other predefined sign) is used to represent where the fire is taking place, either NP Mavrovo or NP Pelister. In addition to the changes in parameter values, the user will also know the coordinates to the fire affected area and the time of the activation of the initial and secondary alarm.

Following the successful extinguishing of the fire, the event will be documented using the purposefully made reports for such instances and uploaded in the MKFFIS archives. These reports, as

part of a summative report on a weekly or a monthly basis will be forwarded to European Forest Fire Information System EFFIS) in an XML file.

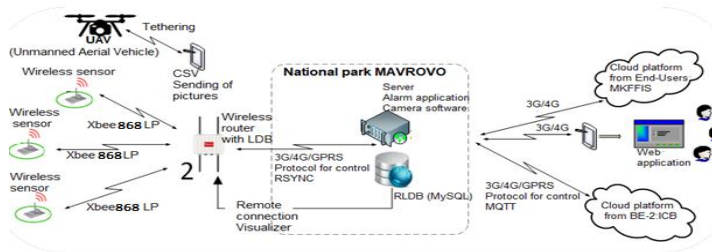


Figure 2. Structural model of ASPires in the NP Mavrovo

The structural models and their main building blocks contain three sensors, as they will be installed in the NPs Mavrovo (Figure 2) and Pelister (Figure 3). They collect information using 3 sensor probes, for temperature, humidity and pressure; carbon dioxide (CO₂) and carbon monoxide (CO).

Each sensor collects the data in specific time intervals. The data collected will be sent from the sensors to a field gateway. The data is then stored in a local MySQL database located in the NPs Mavrovo and Pelister with raw values for each sensor.

Then, the data is replicated to a server which is in the NP Mavrovo the data is converted to a PostgreSQL database.

The application in the control centre has the exact location of each sensor and the information it collects. This is how the operator gets a general idea of what is happening in the field. Furthermore at the user's disposal is the control over a camera or an aerial drone which can be used to get a live feed from the situation at hand. The application can also generate alarms in case the sensor values pass their assigned threshold.

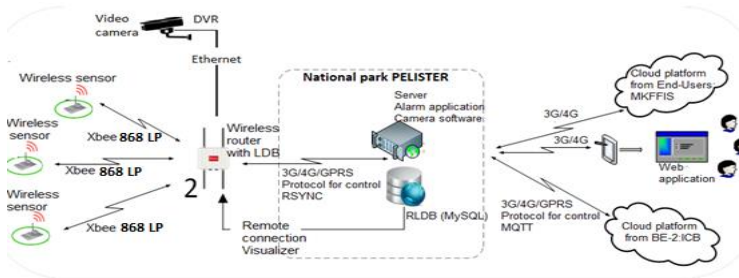


Figure 3. Structural model of ASPires in the NP Pelister

There are areas in the NPs which are difficult or even impossible to access by man. The access there can be only achieved by small aerial drones. In case of a fire alarm, the operators can check the situation on the spot using a drone in the NP Mavrovo and take pictures using a tablet that stores the images and coordinates in its own SQLite database and then send the data to the server in the NP Mavrovo.

In the case of the NP Pelister, a camera will be installed on a surveillance tower.

The data from the server in the control centre in Mavrovo is also sent to the Macedonian Forest Fires Information System (MKFFIS) server located in the CMC in Skopje. MKFFIS can read out the data for the alarms from the database or a service from the Web Server in National Park Mavrovo.

Also, all the data is sent to the cloud platform made by ICB, Bulgaria as a backup solution.

3. Standard interfaces and protocols at sensor network

3.1. Communication protocol used to transfer data from sensors to field gateway/ router

The sensors collect environmental data using 3 sensor probes for the following critical parameters:

- Temperature, humidity, and atmospheric pressure;
- Carbon Dioxide (CO₂); and
- Carbon monoxide (CO).

Waspmotes are deployed in strategic locations. Those sensors are connected to Wasp mote through the Gases Board, which contains the electronics needed to implement an easy hardware integration of a lot of different gas sensors.

Each sensor collects the data in specific time intervals. The data collected will be sent from the sensors to a field gateway, „Meshlium“.

The data will be sent via supported XBee 868LP (Figure 4) module designed to provide a high-performance, low-power module at an extremely competitive price point. ZigBee is based on the IEEE 802.15.4 standard.

The XBee 868LP module can run either a proprietary DigiMesh® or point-to-multipoint networking protocol utilizing a low-power.

XBee 868LP
Basic technical specifications for XBee 868LP

| Parameter | XBee 868LP |
|--|------------------------------|
| Frequency band | 863 to 870 MHz |
| Tx power | 14 dBm (software selectable) |
| Tx Current | 48 mA typical at 3.3 V |
| RF data rate | 10 kb/s |
| Rx sensitivity | -101 dBm |
| Max range, indoors/urban | up to 112 m |
| Max range, outdoors (line of sight) with ~2.1dBi antenna | up to 8.4 km |
| Regulatory approvals | Europe |

figure 19. Technical parametrs of XBee 868LP



figure 20. available bands XBee 868LP

Figure 4. Basic technical specifications for XBee 868LP

The XBee 868LP operates between 863-870 MHz, making it deployable in approved European countries by utilizing a software selectable channel masking feature. The frequency used for this system is the 868 MHz band, using 30 software selectable channels. Channels are spaced 100 kHz apart. The transmission rate is 10 kbps.

If the interface fails, the data will be sent by LoRaWAN or another interface.

LoRaWAN network can be used for communication between sensors. LoRaWAN represent a Low Power Wide Area Network (LPWAN) specification intended for wireless battery-operated devices. This network standard will provide good interoperability among the sensors without the need of complex local installations.

LoRaWAN network architecture is based on star-of-stars topology where gateways are transparent bridge relaying messages between end-devices and a central network server in the back-end. Gateways are connected to the network server via

standard IP connections, while end-devices use single-hop wireless communication to one or many gateways.

„Meshlium“ router will receive the sensor data sent by Plug and Sense using the RF radio (868LP) and it will store the frames in a local MySQL database located in the NPs Mavrovo and Pelister with raw values for each sensor. That can be done in an automatic way, thanks to the Sensor Parser1a software system which is able to do the following tasks in an easy and transparent way: receive frames from XBee modules (with the Data Frame format) and stores the data in the local database.

„Meshlium“ can perform two different storage options with the frames captured (Figure 5):

- Local database
- External database.

All the data is stored in the local database in the first place, then it can be synchronized to an external database as per user needs.

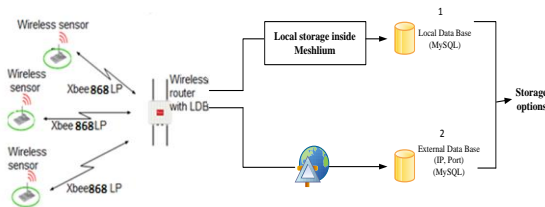


Figure 5. Meshlium storage option

3.2. Communication protocol used to transfer data from drone to field gateway/ router

In case of a fire alarm, the operators can check the situation on the spot using drone in the NP Mavrovo and take pictures using a tablet that stores the images and coordinates in its own SQLite database and then send the data to the server in the NP Mavrovo.

The drone interface using network communication via TCP/IP protocol, transceiver transfer of remote image data from the drone

through the net export to the protocol conversion computer, image data transfer to fire monitoring software via switch or fibre-optical network is presented in Figure 6 The data receiving module receives the video and telemetry data from the drone and stores them in its SQLite database.

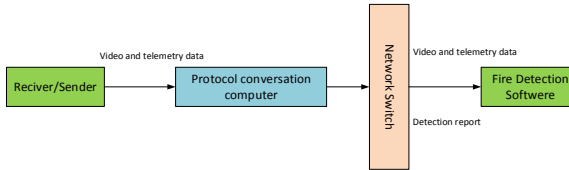


Figure 6. Software interface relationship

To support the data transfer, „Meshlium“ devices are being added to collect information over wireless fidelity (Wi-Fi) It is the only Multiprotocol router capable of interconnecting with 6 technologies: wireless sensor network (WSN): 802.15.4 / ZigBee, Wi-Fi: 2.4 GHz or 5 GHz at high or low power, GPRS: quad band, Bluetooth: communication with mobile phones or personal digital assistant (PDAs), GPS and Ethernet.

Wi-Fi is a wireless local area network (WLAN) that utilizes the IEEE 802.11 standard through 2.4 GHz UHF (radio frequencies in the range between 300 MHz and 3 GHz,) and 5 GHz ISM (2.4-GHz or the 900-Mhz frequencies) frequencies. Wi-Fi provides Internet access to devices that are within the range (about 20 m from access point).

3.3. Communication protocol used to tranfer data camera to field gateway/ router

In NP Pelister a camera will be installed on a surveillance tower. In case of a fire alarm, the operators can check the situation on the

spot using this camera and take pictures that are sent to the Web Server in the NPPelister and stored in the NPs database.

Towers that are equipped with internet connection could be used for information download and sending. In this case the HD quality images could be sent to the local server working at dew computing level. The data could be also sent directly to the ASPires cloud that is working at cloud computing level in real-time and be processed there. The use of the cloud computing allows virtualisation of the process of gathering information and the process of processing the gathered information regardless of time and space.

3.4. Communication protocol used to transfer data from field gateway/router to a database server

„Meshlium“ is a Linux router which works as the Gateway of the Waspnote Sensor Networks. As we said above It can contain 5 different radio interfaces. „Meshlium“ comes with the Manager System, a web application which allows to control quickly and easily the WiFi, XBee/LoRa, Bluetooth and 3G/GPRS configurations along with the storage options of the sensor data received. „Meshlium“ receives the sensor data sent by Waspnote using the XBee- PRO, LoRa, GPRS, 3G or WiFi radios. Then 4 possible actions can be performed :

1. Store the sensor data in the Meshlium Local Data Base (MySQL);
2. Store the sensor data in an External Data Base (MySQL);
3. Send the information to the Internet using the Ethernet or WiFi connection;
4. Send the information to the Internet using the 3G/GPRS connection.

Then, the data is replicated to a server which is in the NPMavrovo via 3G/4G/GPRS protocol and the data is converted to a PostgreSQL database.

PostgreSQL is an object-relational database management system (ORDBMS) with an emphasis on extensibility and standards compliance. As a database server, its primary functions are to store data securely and return that data in response to requests from other software applications. It can handle workloads ranging from small single-machine applications to large Internet-facing applications (or for data warehousing) with many concurrent users. PostgreSQL database boasts sophisticated features such as tablespaces, point in time recovery, Multi-Version Concurrency Control (MVCC), online/hot backups, asynchronous replication, nested transactions (save points), a sophisticated query planner/optimizer, and write ahead logging for fault tolerance. It supports international character sets, multibyte character encodings, Unicode, locale-awareness for sorting, case-sensitivity, and formatting. It is highly scalable in both ways: quantity of data it can manage and in the number of concurrent users.

Data integrity features include primary keys, foreign keys with restricting and cascading updates/deletes, check constraints, unique constraints, and not null constraints.

3.5. Desktop application and its connection with database

The application in the control centre has the exact location of each sensor and the information it collects. This is how the operator gets a general idea of what is happening in the field. Furthermore at the user's disposal is the control over a camera or

an aerial drone which can be used to get a live feed from the situation at hand.

The application, which in the NP Mavrovo is connected by WiFi and in NP Pelister is connected by 3G/4G reads out the information from the server in Mavrovo and the data is sent to MKFFIS in Skopje.

This communication module is specially oriented to work with Internet servers, implementing internally several application layer protocols, which make easier to send the information to the cloud. We can make HTTP navigation, downloading, and uploading content to a web server. We can also set secure connections using SSL certificates and setting TCP/IP private sockets. In the same way, the FTP protocol is also available which is useful when your application requires handling files.

The 4G module offers the maximum performance of the 4G network as it uses 2 different antennas (normal and diversity) for reception (MIMO DL 2x2), choosing the best received signal at any time and getting a maximum download speed of 100 Mbps.

4. Cloud level definition

4.1. Communication protocol used to transfer data from database server to Cloud Server

This project approach adds a LoRaWAN or IoT or other gateway between the wireless router of XBee network and the local server in the national park in Republic of Macedonia that is capable to timestamp the data and transfer it to the local server in the control centre and to the cloud using multicasting. The multicasting will use two parallel point-to-point sessions.

The connection to the cloud may use MQTT or HTTP(s) protocols or similar via 3G, 4G or Internet connection.

This approach (Figure 7) requires deployment of a gateway locally in the NPs. The delay in getting the data locally and in the cloud, is minimal. The added value of the solution is allowing the experts in the cloud to use other artificial intelligence solutions for data analysis that will enrich the forest fire prevention procedures used locally. There is a need to consider further how the local authorities in the NP will benefit from the expertise provided by the cloud back.

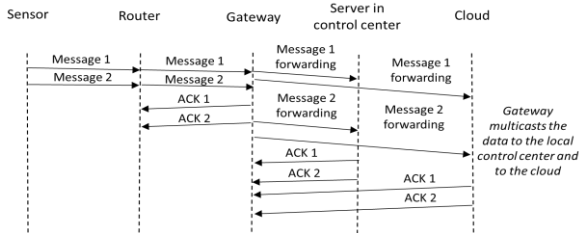


Figure 7. Connection scenario between local centre in Republic of Macedonia and the cloud solution

MQTT (Message Queuing Telemetry Transport) is publish/subscribe messaging protocol specially designed for constrained low-bandwidth, high-latency or unreliable networks. MQTT yet provides sufficient multi-layer security features, each designated to prevent from a specific attack:

- Network security – by using VPN as a foundation between server broker and clients, the protocol provides a trustworthy connection.

- Transport security – SSL/TLS is used for transport level encryption to provide confidentiality.
 - Application security – Client ID and user-password pair could be used by the client to authenticate the client to the server.
- MQTT protocol is often called “the messaging protocol for Internet of Things” and covers out of the box a multitude of failure scenarios to assure reliable communication.

4.2. Communication protocol used to transfer data from National Parks to MKFFIS in Skopje

The data from the server in the control centre in NP Mavrovo is also sent to Macedonian Forest Fires Information System (MKFFIS) server which is in Crisis Management Centre (CMC) in Skopje using 4G/3G/2G or GPRS (Depending on Base station in NP Mavrovo) communication module. The advantages of the communication module was explained in section 3.5..

MKFFIS can read out the data for the alarms from the database or a service from the Web Server in National Park Mavrovo and replicate it to its own PostgreSQL database. Also, all the data is sent to the cloud platform made by ICB, Bulgaria as a backup solution.

5. Conclusion

The aim of ASPires is to speed up the process of fire detection at the local, regional, national and international level as well as to support the better coordination between authorities and improve the existing guidelines on fire prevention using the fire weather index, risk analysis and using data from terrestrial sensors.

Application of a WSN in forest fire detection allows remote monitoring of various locations of interest through different types of sensors, which are communicating information with the control stations. The result is an increased coverage area, with quicker and safer responses.

In this paper we presented full design of the system for monitoring forest is described which consists of cloud platform, sensor networks and mobile (drone) technologies for data collection and cameras.

This paper aims to develop and show in more details one of the concepts that are based on sensor networks and cloud computing. It is demonstrating the possibility to map different technological solutions at different levels of the forest fire detection systems like access fixed and mobile network, edge solutions using gateways and scalable cloud computing pure software solutions. The main result was the definition of the set of standard interfaces and protocols for network interoperability.

6. Bibliography

- [Alkhatib, 2014] A. A. A. Alkhatib, A Review on Forest Fire Detection Techniques, 2014.
- [Bulusu; Heidemann; Estrin, 2000] N. Bulusu, J. Heidemann, and D. Estrin. GPS-less low cost outdoor localization for very small devices. IEEE Personal Communications, 2000.
- [Berni; Caramona; Martínez; Rodríguez, 2012] Jorge Fernández-Berni; Ricardo Carmona-Galán, Juan F. Martínez-Carmona;

Ángel Rodríguez-Vázquez, Early forest fire detection by vision-enabled wireless sensor networks, 2012.

- [Chen et al., 2010] Chen, Yihua, Mingming Zhang, Xin Yang, and Yongjin Xu. "The research of forest fire monitoring application." In *Geoinformatics, 2010 18th International Conference on*, pp. 1-5. IEEE, 2010.
- [Chenm; Yin; Huang; Ye, 2006] Chen T, Yin Y, Huang S, Ye Y Smoke detection for early fire-alarming system based on 12 video processing, 2006.
- [Jadhav; Deshmukh, 2012] P. Jadhav and V. Deshmukh., Forest Fire Monitoring System Based On ZIG-BEE Wireless Sensor Network, 2012.
- [Lozano, Rodriguez, 2007] C. Lozano,O. Rodriguez., Design of Forest Fire Early Detection System Using Wireless Sensor Networks, 2007.
- [Lora-Aliance, 2017] <https://www.lora-alliance.org/technology> (accessed 23 May 2018).
- [Martinez-de Dios et al., 2008] Martinez-de Dios, J. R., Begoña C. Arrue, Aníbal Ollero, Luis Merino, and F. GómezRodríguez. "Computer vision techniques for forest fire perception." *Image and vision computing*26, no. 4 (2008): 550-562.
- [Meshlium Technical Guide, 2017] Meshilium Tehnical Guide, 2017
<http://www.libelium.com/development/meshlium/documentation/meshliumtechnical-guide>.
- [Paneque-Gálvez; McCall; Napoletano; Wich; Pin Koh. 2014] K. M. Michael, M. N. Brian, A. W. Serge, P. K. Lian and J. Paneque-Gálvez, Small Drones for Community-Based Forest

Monitoring: An Assessment of Their Feasibility and Potential in Tropical Areas, 2014.

[PostgreSQL, 2017] <https://www.postgresql.org/about/> (accessed 23 May 2018).

[Rodoaplu; Meng, 1999] Rodoaplu V, Meng TH. Minimum Energy Mobile Wirele Networks [J]. IEEE J.Select.Areas Communications, 1999.

[Sabri et al., 2013] Y. Sabri, N.El Kamoun, V. Gramoli, R. Guerraoui, Rachid. Forest Fire Detection and Localization with Wireless Sensor Networks. In Networked Systems: First International Conference, NETYS 2013, Marrakech, Morocco, May 2-4, 2013, Springer Berlin Heidelberg, Berlin, Heidelberg.

[Solobera, 2010] Javier Solobera-Detecting Forest Fires using wireless Sensor Networks with wasp mote, Libelium Comunicaciones Distribuidas, 2010. http://www.libelium.com/wireless_sensor_networks_to_detect_forest_fires.

[Shannon, 1949] C.E.Shannon. The Mathematical theory of communication. In: The Mathematical Theory of Communication. Ed. C.E.Shannon and W.Weaver. University of Illinois Press, Urbana, 1949.

[Tagarakis et al, 2011] Tagarakis, A., V. Liakos, L. Perlepes, S. Fountas, and T. Gemtos. "Wireless Sensor Network for Precision Agriculture." In Informatics (PCI), 2011 15th Panhellenic Conference on, IEEE, 2011.

[Zhang; Wang; Peng; Li; Lu Guo, 2015] Lan Zhang, Bing Wang, Weilong Peng, Chao Li, Zeping Lu; Yan Guo, Forest Fire Detection Solution Based on UAV Aerial Data 2015.

7. Authors' Information



Jugoslav ACKOSKI, PhD, Assistant Professor, Military Academy „General Mihailo Apostolski“

jugoslav.ackoski@ugd.edu.mk.

Major Fields of Scientific Research:

system development in emergency medicine, improving detection and monitoring forest fires in the National Parks, published more than 50 (fifty) journal's articles and conference papers. He is author on a more than 5 (five) Book's Chapters and has been part of 2(two) Short Term Scientific Missions (STSM), which have been funded by European Cooperation in Science and Technology (COST) actions.



Nikola Kletnikov, MsC, Military Academy „General Mihailo Apostolski“ Skopje, nikola.kletnikov@ugd.edu.mk.

Major Fields of Scientific Research: Crisis management, Operation Planning, improving monitoring of forest area and detection of forest fires in the National Parks. Published more than 15 (fifteen) journal's articles and conference papers



***Nevena SERAFIMOVA, PhD, Assistant Professor, Military Academy „General Mihailo Apostolski“ –
nevena.serafimova@ugd.edu.mk.***

Major Fields of Scientific Research: mathematical modeling, game theory, algorithms for detection of forest fires. She is author of conference papers and articles in theoretical and applied mathematics that have been published in domestic and international journals.



***Igorce Karafilovski, MsC, Crisis Management Center “Skopje,
igorce.karafilovski@cuk.gov.mk.mk.***

Major Fields of Scientific Research: Information Technology, Computer Science, Geographic Information Systems, Project Management, Project Planning and Risk Management, Cyber Security. Published more then 10 (then) journal’s article and conference papers.



**Rossitza Ivanova Goleva, Ph. D.,
Assistant-Professor, Departement of
Informatics, New Bulgarian University,
Sofia, Bulgaria**

*Major Fields of Scientific Research:
Distributed Networks, Cloud/ Fog/ Dew/
Smart Dust Computing, Communication
Networks, Performance Analyse.*



**Katerina ZLATANOVSKA MSc.,
Ministry of Defence, Macedonia,
k.zlatanovska@yahoo.com.**

*Major Fields of Scientific Research: cyber
security, cyber defense and information
and communication systems, improving
detection and monitoring forest fires in
the National Parks. She is author Book's
Chapter „Hacking and Hacktivism as an
Information Communication System
Threat“, from Handbook of Research on
Civil Society and National Security in the
Era of Cyber Warfare.*

CSECS 2018, pp. 251 - 269

Computer Science and Education in Computer Science
14th Annual International Conference
ISSN 2603-4794

June 29 – 30, 2018, Boston, USA

TEACHING DATA MINING TECHNIQUES TO APPLIED BUSINESS ANALYTICS STUDENTS WITH THE HELP OF INTERACTIVE HANDS-ON TUTORIALS

Penko Ivanov

New Bulgarian University, Department of Computer Science

Abstract: *This paper discusses a contemporary educational approach for teaching data mining techniques to applied business analytics students with the help of interactive hands-on tutorials. This approach provides for different levels of guidance, depending on the complexity of the task. The most complex technical steps of the data mining process can be automated, allowing students with non-technical backgrounds to understand and appreciate the conceptual purposes without needing to dive into the minutiae of detailed optimization algorithms. The proposed method leverages advanced technologies to enable the delivery of a complete educational product that resonates with students of different backgrounds, and with various preferred learning styles.*

The author presents his work on the development of interactive hands-on tutorials for teaching data mining techniques directly in the R console for the purposes of the “AD699 Data Mining for Business Analytics” – a graduate course being delivered both

online and on-campus by Boston University / Metropolitan College, Department of Administrative Sciences.

Keywords: *Data Mining; Data Mining Techniques; Business Analytics; R; Python; Advanced Data Visualization; Statistical Modeling; Text Mining.*

Introduction

The role of business analytics is to generate insights that convert the potential value of data into tangible business value. Business analytics practitioners seek opportunities to apply descriptive, predictive, and prescriptive analytics both to solve business problems and to present related insights in an understandable format for key decision-makers. Data mining is the process of finding, extracting, visualizing, and reporting useful information from data. Naturally connected with computer science, data mining is an integral part of business analytics. Consequently, applied business analytics students should attain a comfortable level of fluency with data mining techniques, along with familiarity with relevant software tools and programming languages, such as R and Python.

The challenges of teaching technical topics to students with non-technical backgrounds were faced during the development of “AD699 Data Mining for Business Analytics”¹ graduate course, which is being delivered both online and on-campus by Boston University / Metropolitan College, Department of Administrative Sciences as an integral part of its Applied Business Analytics (ABA) program. The course development team needed to

incorporate into the course curricula tools for data analysis, statistics and machine learning. My main focus as a course co-developer was on finding an efficient method to teach data mining techniques to applied business analytics students.

Incorporating R into BU MET AD699 Data Mining for Business Analytics course

R is among the most popular analytics instruments across many industries. It is a programming language, but it was created for statisticians, data scientists and business analysts. The use of R does not necessarily require an understanding of the underlying computing logic – e.g. the paradigms of object-oriented programming, or the intricacies of software design patterns.

In comparison with a general purpose programming language like Python, which, in recent years, has become very popular for data manipulation and especially for machine learning, R is perceived as more comprehensible for non-technical users. For instance R allows a user to add functions to a single vector (a basic data type in R) without coding a loop, which is not the case in Python where to operate on lists is more complex.

Strong advantages of R are its widely-available extension packages and its powerful graphic capabilities. Furthermore R is platform-independent, interpreted language with an interactive console.

Last but not least, R is free to install, and can be used without the need to purchase a license.

The above listed arguments were rationale for us to incorporate R into BU MET AD699 Data Mining for Business Analytics course. Nevertheless we encourage our students to study also Python, which could give them additional capabilities once they master the fundamental data mining techniques.

The selection of R for the purposes of our course targets focus on the business perspective and on the value of data mining rather than on programming. However R has its specifics making it not trivial for learning. The students with non-technical backgrounds often need additional guidance when running their code in order to become comfortable with particular technique.

My approach to meeting this need was by means of interactive hands-on tutorials. This approach provides for different levels of guidance and it's equally applicable in both online and on-campus delivery of the course.

Selecting a platform for interactive R courses/lessons

The first step towards the development of the interactive hands-on tutorials was the selection of a technological platform suitable for our needs.

Popular platforms for interactive R courses are the commercial online platforms provided by third parties - companies such as DataCamp² (Figure 1). The strongest advantage of these platform is that they are available online and don't need local R environment to run R code. The online platforms provide both end-to-end solutions for interactive R courses accommodating the educational content and web services which could be integrated into educational provider's applications or sites.

Although they are very convenient, the online platforms have some drawbacks. The learning environment is simplified and doesn't look much like a real R environment. This is not exactly a

true disadvantage, but in our case we aim to also prepare our students for working with industry common tools and instruments like RStudio – an integrated development environment for R.

Another disadvantage of the online platforms provided by third parties is that the content developer (the educational provider respectively) is dependent on technology and/or policy changes done by the platform providers.

And again, last but not least, the commercial online platforms are not free of charge.

These reasons led me to searching for an alternative solution, which I found in swirl³. The swirl package is our selected platform for interactive R courses and lessons, which I'm using for the development of the hands-on tutorials for BU MET AD699 Data Mining for Business Analytics graduate course.

Swirl is a free R package “that turns the R console into an interactive learning environment. Students are guided through R programming exercises where they can answer questions in the console”⁴ and thus are being assisted to run their code applying different data mining techniques.

Swirl is developed with the educational purposes in mind. The content developer has full control on his courses and lessons (courses consist of one or more lessons). He is independent of any technology and/or policy changes done by the platform developers.

Also swirl is compatible with Git and GitHub, which could make swirl courses available in the cloud together with all needed sample datasets and other additional files. The swirl developers have also provided a GitHub course repository⁵ with free recommended courses covering the basics of the R programming.

DataCamp

EXERCISE

Course Outline

Build a classification tree

Let's get started and build our first classification tree. A *classification tree* is a decision tree that performs a classification (vs regression) task.

You will train a decision tree model to understand which loan applications are at higher risk of default using a subset of the **German Credit Dataset**. The response variable, called "default", indicates whether the loan went into a *default* or *not*, which means this is a binary classification problem (there are just two classes).

You will use the **rpart** package to fit the decision tree and the **rpart.plot** package to visualize the tree.

```

1 # Look at the data
2 str(creditsub)
3
4 # Create the model
5 credit_model <- rpart(formula = default ~ .,
6   data = ---,
7   method = "Class")
8
9 # Display the results
10 rpart.plot(x = ---, yesno = 2, type = 0, extra = 0)

```

INSTRUCTIONS 100 XP

The data frame `creditsub` is in the workspace. This data frame is a subset of the original German Credit Dataset, which we will use to train our first classification tree model.

- Take a look at the data using the `str()` function.
- In R, formulas are used to model the response as a function of some set of predictors, so the formula here is `default ~ .`.

R CONSOLE SLIDES

Run Code Submit Answer

Figure 1. DataCamp – online platform offering interactive R courses on topics in data science, statistics, and machine learning

Developing interactive hands-on tutorials

For development of new custom swirl courses and lessons, the swirl developers have provided another free R package called swirlify⁶. With the help of swirlify instructors are able to create their own interactive content.

The swirl lessons are written in yaml format (Figure 2). Lesson's yaml document contains five main types of classes: meta, text, cmd_question, multi_question and figure. Other class types are video, exact_question, text_question and script. One lesson.yaml file structures the content of each lesson – what the student sees inside the R console while taking the lesson.

The class meta is being used to add meta information about the lesson: course name, lesson name, author name, version, etc.

The class text is in place for displaying text information. It is useful for adding introductions, theoretical explanations and additional details to lesson's topics.

The class cmd_question is essential for the interactive lesson. A cmd_question requests the student to enter a line of R code according to instructions provided by the lesson author (instructor) - Figure 3. The custom instructions can provide the student with different level of guidance decided by the instructor depending on the complexity of the task.

```

-- Class: meta
Course: Rtf699 Data Mining for Business Analytics
Lesson: Evaluating a Linear Regression Model
Author: Penko Ivanov
Type: Standard
Organization: BU MET
Version: 2.4.3

-- Class: text
Output: "This lesson will guide you through Individual Exercise 6: Evaluating a Linear Regression Model. Following its instructions you will get familiar with the Linear Regression Model."

-- Class: text
Output: "For your convenience the data from the file Advertising.csv was loaded into an R data frame named 'data' at the beginning of the lesson."

-- Class: cmd_question
Output: "Use the head() function to display the header row of your data and it's first few observations."
CorrectAnswer: head(data)
AnswerTests: omitTest(correctExpr='head(data)')
Hint: "Enter head(data)"

-- Class: text
Output: "Before we start building our model, let's do some data exploration creating a matrix plot that visualizes the correlations between the 'TV', 'Radio', and 'Newspaper' advertising channels."

-- Class: text
Output: "As you have already learned, a good way to do so in R is with the help of the ggally package."

-- Class: cmd_question
Output: "Enter require(GGally) to make sure that you have loaded the libraries required for the next functions which you are going to use."
CorrectAnswer: require(GGally)
AnswerTests: omitTest(correctExpr='require(GGally)')
Hint: "Enter require(GGally)"

-- Class: cmd_question
Output: "Now enter ggpairs(data[, c(2,3,4,5)]). Note that we use a subset of your data to exclude the first column containing the variable 'X', which is the response variable."
CorrectAnswer: ggpairs(data[, c(2,3,4,5)])
AnswerTests: omitTest(correctExpr='ggpairs(data[, c(2,3,4,5)])')
Hint: "Enter ggpairs(data[, c(2,3,4,5)])"

-- Class: mult_question
Output: "Look at the visual. Is there a relationship between ads and sales?"
AnswerChoices: Yes, there is.,No, there is not.,It is not clear from this visual.
CorrectAnswer: Yes, there is.
AnswerTests: omitTest(correctVal= 'Yes, there is.')
Hint: "Look at the visual again!"

```

Figure 2. The swirl courses' .yaml format

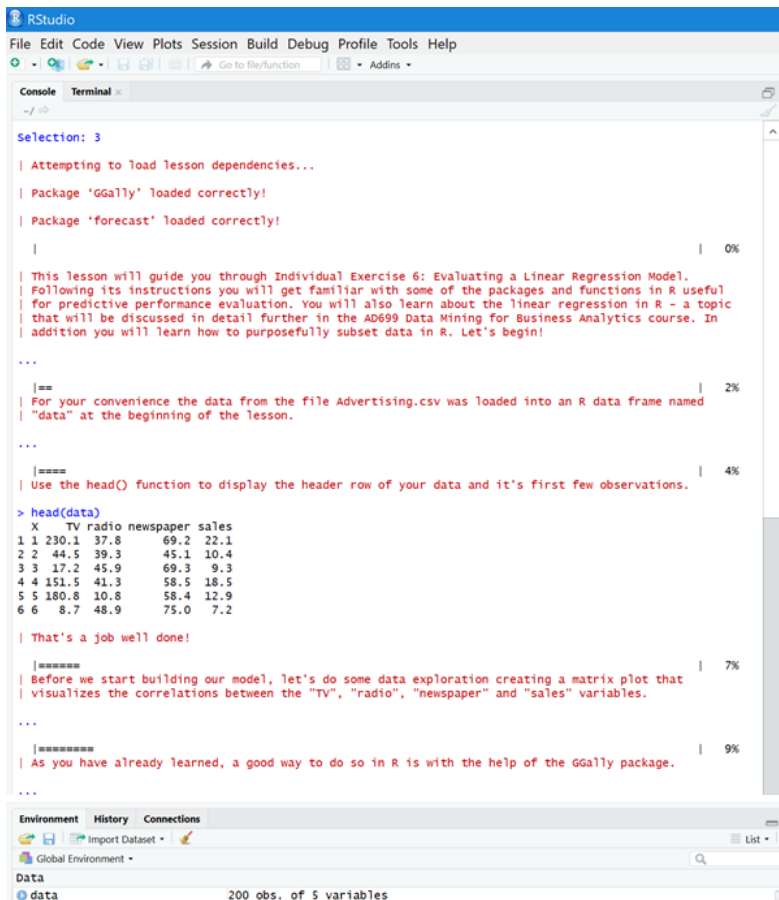


Figure 3. “Answering” a command question

Once the student has entered the required R code, the code is being automatically checked for correctness. Not only for syntax correctness, but also whether this is the correct code for completion of the required operation. If the entered code is correct, then it's being executed, the student receives feedback

and the lesson continues with the next class in its structure. If the code is incorrect, then the student again receives a feedback and a hint on how to fix the error. The hint is custom specified by the instructor as well.

That way, following the instructions organized in the lesson's structure, the student executes an R program and receives knowledge achieving the lesson's objectives.

In case of inability to answer any of the questions, the student has the opportunity to skip this question by entering `skip()` in the console. In that case the required code will be executed automatically and the lesson will continue with the next question (or with the next class in its structure). By entering `play()` the student is able to experiment in the console within the current environment, but outside the lesson's context until he enter `nxt()`.

The class `multi_question` gives the ability for integration of multiple choice questions into the swirl lesson (Figure 4). The multiple choice questions are efficient way to check the student's understanding of a topic by presenting him a selection of options. These options are presented in a different (random) order every time the question is displayed. Like in `cmd_question`, here is an option for providing a hint in case of wrong answer.

The class `figure` automatically executes a piece of predefined R code within the current environment inside the swirl lesson context. Common application of this class is for visualizing graphics for the purposes of theoretical explanations or multiple choice questions. The class `figure` calls a separate `.R` file where the predefined code is stored.

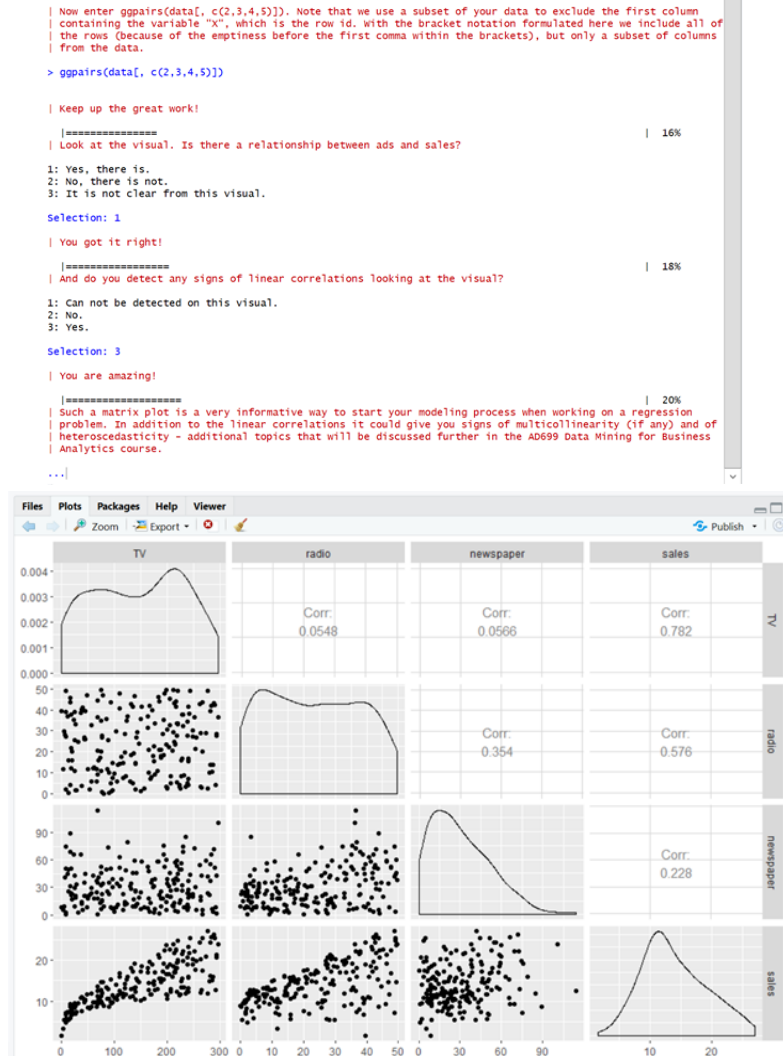


Figure 4. Answering a multiple choice question

In our case, for the purposes of BU MET AD699 Data Mining for Business Analytics graduate course, another application of the

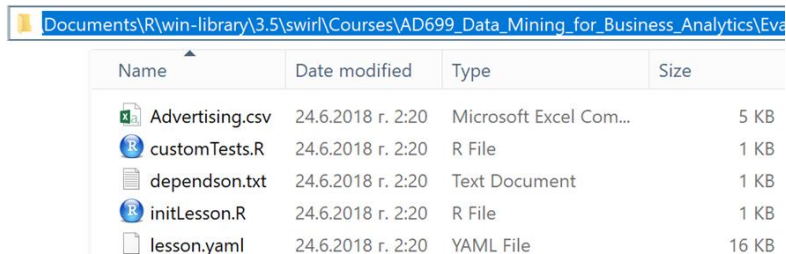
class figure is to automate the most complex technical steps of the data mining process, allowing students with non-technical backgrounds to understand and appreciate the conceptual purposes without needing to dive into the minutiae of detailed optimization algorithms.

The class video provides students the opportunity to open a URL in a web browser.

The `exact_question` and `text_question` classes are other options for testing the students' understanding and knowledge.

The class script is probably the most complex one. It involves writing of custom answer tests and allows to evaluate the correctness of a script that a student has written. This class is not used for the purposes of our administrative science course so far. The class script would be more suitable for computer science courses.

The swirl course package contains folders for course's lessons. Each lesson's folder contains all the files required for this lesson (Figure 5).



| Name | Date modified | Type | Size |
|-----------------|-------------------|------------------------|-------|
| Advertising.csv | 24.6.2018 r. 2:20 | Microsoft Excel Com... | 5 KB |
| customTests.R | 24.6.2018 r. 2:20 | R File | 1 KB |
| dependson.txt | 24.6.2018 r. 2:20 | Text Document | 1 KB |
| initLesson.R | 24.6.2018 r. 2:20 | R File | 1 KB |
| lesson.yaml | 24.6.2018 r. 2:20 | YAML File | 16 KB |

Figure 5. The swirl lesson folder structure

Every lesson/folder contains one `lesson.yaml` file. The file `dependson.txt` contains list of R packages to be installed and/or loaded at the beginning of the lesson. The file `initLesson.R`. This file contains an R script, which is executed automatically at the

beginning of the lesson, to prepare the R environment for the lesson - by loading a sample data set for example. The file named customTests.R is for the custom answer tests if any.

The swirl lesson's folder could also contain various file types with sample data sets for the lesson.

After a swirl course is created, and his lessons are developed, it could be packed in a single file packaging (.swc file) for distribution. This is done with the help of the swirlify package's function pack_course().

For the purposes of BU MET AD699 Data Mining for Business Analytics graduate course I developed a proprietary swirl course. Each lesson of this course is an interactive hands-on tutorial covering a particular topic from AD699.

AD699_Data_Mining_for_Business_Analytics.swc is available on the course page. The students can download it and simply install it in their local R environments with the help of the swirl package's function install_course()(Figure 6). The download and installation process is described in the "Short introduction to swirl - Tutorial", which is part of the course's teaching materials.

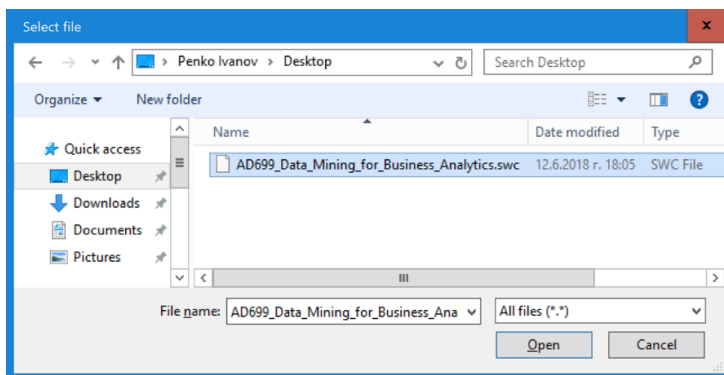
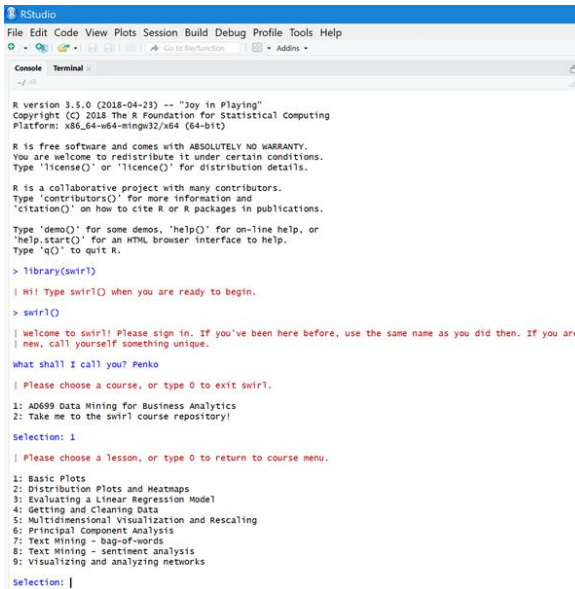


Figure 6. Installing MET BU AD699 swirl course

Once the course is installed, the student can run swirl, start the course, and select a particular lesson/interactive hands-on tutorial as shown on the Figure 7 below:



```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal
~/R

R version 3.5.0 (2018-04-23) -- "Joy in Playing"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(swirl)
| Hi! Type swirl() when you are ready to begin.
> swirl()
| Welcome to swirl! Please sign in. If you've been here before, use the same name as you did then. If you are
| new, call yourself something unique.
what shall I call you? Penko
| Please choose a course, or type 0 to exit swirl.
1: AD699 Data Mining for Business Analytics
2: Take me to the swirl course repository!
Selection: 1
| Please choose a lesson, or type 0 to return to course menu.
1: Basic Plots
2: Distribution Plots and Heatmaps
3: Evaluating a Linear Regression Model
4: Getting and Cleaning Data
5: Multidimensional Visualization and Rescaling
6: Principal Component Analysis
7: Text Mining - Bag-of-words
8: Text Mining - sentiment analysis
9: Visualizing and analyzing networks
Selection: |

```

Figure 7. Running swirl and starting the BU MET AD699 swirl course

There are many benefits of R, but to ensure smooth learning process its disadvantages must be considered as well. Among these are: the quality of some packages is not good enough; as R is open-source and non-proprietary, there is no single entity or “owner” to complain to if something doesn’t work; and R commands give little thought to active memory management (so R can consume all available memory).

Because of these, to overcome possible problems with students’ local R environments, we set up a stable, controlled R environment on the BU MET virtual laboratory.

The BU MET VLab infrastructure (Figure 8) – a kind of a private cloud - is available to students for the time of their training. The virtual machines (laboratories) are powered up on demand from custom templates, tailored to fulfil particular student’s needs, including all the required software and data for his courses.

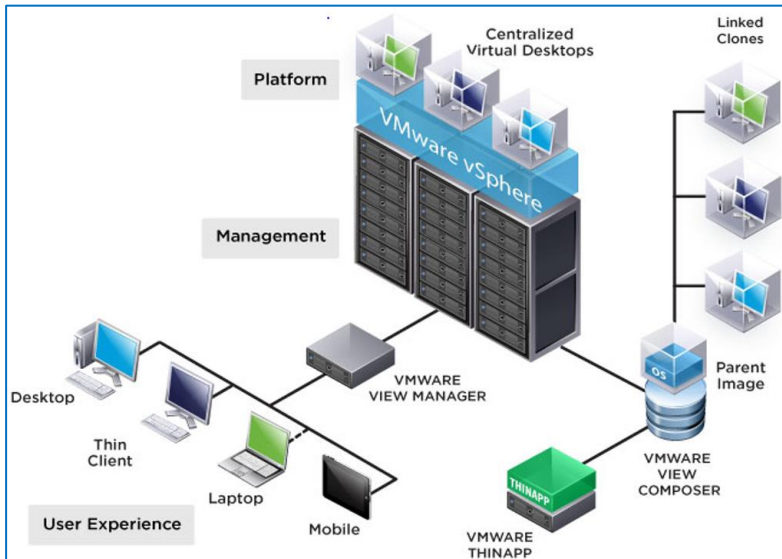


Figure 8. Controlling the environment – BU MET VLab

The swirl lesson file is stored in the VLab’s course folder as well. All datasets from the BU MET AD699 Data Mining for Business Analytics course curricula are also provided in the VLab.

The virtual laboratory eliminates the diversity of student local environments ensuring better support for students. It also eliminates the need for students to spend time and efforts setting up their local environments, which are non-trivial tasks for non-technical students, giving them better learning experience.

The “individual exercises” approach

The interactive hands-on tutorials are integrated into BU MET AD699 Data Mining for Business Analytics course lectures through individual exercises. Each individual exercise focuses on a particular subject from the lectures. It describes a scenario and sets expected results. The individual exercises give students the opportunity to practice on selected topics taking advantage of the interactive hands-on tutorials. They prepare the students for their individual assignments and for the team project.

Below is an example of individual exercise:

“Individual Exercise 6: Evaluating a Linear Regression Model

The file Advertising.csv contains 200 observations of 4 variables – “TV”, “radio”, “newspaper” (ads expenditures) and “sales”, and the observation ID “X”. Load this data into R and perform the following:

- 1. Answer the question whether there is a relationship between ads and sales.*
- 2. Determine whether any signs of linear correlation can be detected.*
- 3. Split the data into randomly generated training and validation partitions.*
- 4. Build a linear regression model for prediction of the sales, using the ads expenditures as a predictors (fitting it on the training set).*

5. *Run the model on the training set and on the validation set.*
6. *Evaluate the model's predictive performance on both sets (calculate prediction accuracy measures).*
7. *Compare the training and the validation performance.*
8. *Answer the questions whether the model fit is good, and whether the model's predictive performance is high.*
9. *Consider usage or removal as a predictor of the "newspaper" variable for the model performance improvement.*

The "Evaluating a Linear Regression Model" lesson from the AD699 Data Mining for Business Analytics swirl course will guide you through the above steps."

Current results and further developments

The interactive hands-on tutorials were included in the pilot runs of the BU MET AD699 Data Mining for Business Analytics graduate course delivered on-campus in Spring 2018 and on-line in Summer 1 2018. The approach yielded good results, was well accepted by students and was highly appreciated in the feedback.

The following swirl lessons were developed and are currently available in the AD699 Data Mining for Business Analytics swirl course:

1. Getting and cleaning data
2. Basic plots
3. Distribution plots and heatmaps
4. Multidimensional visualization and rescaling

5. Principal component analysis
6. Evaluating a linear regression model
7. Evaluating a classification model
8. Text mining – bag-of-words
9. Text mining – sentiment analysis
10. Visualizing and analyzing networks

As our approach was evaluated to be successful, further more swirl lessons / interactive hands-on tutorials on statistical models / machine learning algorithms will be developed and incorporated in BU MET AD699 Data Mining for Business Analytics course curricula in future.

Conclusions

The interactive hands-on tutorials proved to be an efficient method to teach data mining techniques to applied business analytics students. They leverage advanced technologies and deliver a complete educational product that resonates with students of different backgrounds, and with various preferred learning styles.

The selection of right technological platform for the particular needs of the education provider is crucial for successful implementation of the interactive content. The open source platforms give better flexibility to content developers and instructors, but are dependent on availability of skills and infrastructure at educational institution's side.

The integration of the interactive hands-on tutorials into the lecture materials is also important for achievement of learning efficiency. The individual exercises are instrumental for this integration.

The presented approach is generally suitable for teaching technical subjects to non-technical students.

References

Crawley, M. J. (2013). *The R Book (Second edition)*. Wiley.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, R. N., & Lichtendahl, Jr., K. C. (2018). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. Hoboken, NJ: Wiley.

¹ <http://www.bu.edu/academics/met/courses/met-ad-699/>

² <https://www.datacamp.com/>

³ <https://swirlstats.com/>

⁴ http://swirlstats.com/swirlify/introduction.html#what_is_swirl

⁵ https://github.com/swirldev/swirl_courses#swirl-courses

⁶ <https://swirlstats.com/instructors.html>

Author's Information



Penko Ivanov, MSc, PMP, CBAP, PMI-PBA, PhD Candidate,

New Bulgarian University, Department of Computer Science penko.ivanov@gmail.com

Major Fields of Scientific Research: IT project management, business analysis in IT projects, requirements engineering, software functional architecture design, Big Data analytics, IT product marketing, IT business development

CSECS 2018, pp. 271 - 297

Computer Science and Education in Computer Science
14th Annual International Conference
ISSN 2603-4794

June 29 – 30, 2018, Boston, USA

DATA ANALYTICS FOR DEVOPS EFFECTIVENESS

Alexandrina Ivanova (1), Penko Ivanov (2)

(1) Senior Software Developer, SAP Labs Bulgaria Ltd.

(2) New Bulgarian University, Department of Computer Science

***Abstract:** This paper discusses the opportunities and challenges associated with the data-driven approach to DevOps. The authors present analytical methods and techniques that can be applied to data collected from the DevOps process, as well as several ways in which that data can be used to improve the enterprises' development capabilities. The authors include specific recommendations regarding the data that should be collected over time, as well as common data storage best practices for enabling analysis and reporting.*

Metrics and DevOps effectiveness KPIs are described in the paper as well. As an example of KPI analytics the authors show particular application of machine learning algorithms for classification of new code change requests into cost categories, which facilitates deployment activities optimization and cost reduction.

The authors' proposed approach is explained in the context of use cases at existing internationally recognized leading companies, which run multiple large-scale software development projects

simultaneously. In addition, the paper explores the changing role of the DevOps engineer regarding data analytics, and highlights the requisite skills and knowledge for him to be successful in the big data era.

Keywords: *DevOps, Data Analytics, Metrics and KPIs, Process Automation, Continuous Integration, Continuous Delivery, Continuous Deployment*

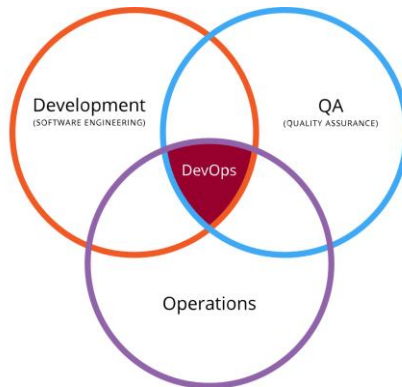
Introduction

Before going over to the essence of the problem addressed in this paper (namely how to achieve better DevOps effectiveness with the help of data analytics), we will make a short introduction to the underlying basic concepts.

DevOps – a leading practice

DevOps is the leading practice of enterprises to develop and deliver competitive software applications and solutions to the market. Its objective is to quickly deliver changes to meet shifting customer demands, while simultaneously improving code quality and reliability.

DevOps can be defined as the interaction of development (software engineering), operations and quality assurance (QA) (Figure 1).

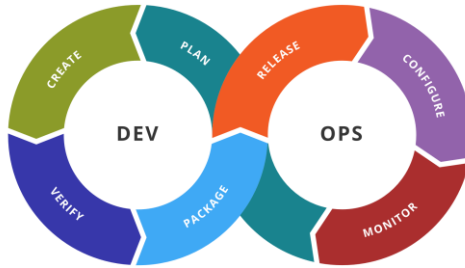


*Image by Rajiv.Pantderivative - work, CC BY 3.0,
<https://commons.wikimedia.org/w/index.php?curid=20202905>*

Figure 1. Venn diagram showing DevOps as the intersection of development (software engineering), operations and quality assurance (QA)

As DevOps is a cross-functional work culture, there is no single tool which is capable to cover all aspects and/or categories in the DevOps process. There are multiple tools forming the DevOps toolchain.

Figure 2 above illustrates the stages in a DevOps toolchain while expresses the continuous nature of the DevOps process.



*Image by Kharnagy - own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=51215412>*

Figure 2. Stages in a DevOps toolchain

DevOps effectiveness

DevOps' effectiveness is measured by evaluating the resolution of problems related to continuous integration, delivery, and deployment while ensuring timely delivery and high quality software. In the past, teams assessed progress with spreadsheets and periodic review meetings, but these methods are no longer effective now that code changes are delivered continuously. To maintain and increase code quality while ensuring customer satisfaction, robust analytics across the DevOps process is required; in fact, analytics are an essential component to the enterprise's DevOps implementation.

DevOps with analytics

In recent years experts say that DevOps with analytics is “the next big thing in software development”¹. This statement is based on expectations that DevOps with data analytics will give enterprises smarter and intelligent software delivery pipeline. The data-driven approach to DevOps has been established by the industry²,

and ensuring his efficiency has become a new challenge for the DevOps engineers.

Internet of Things in manufacturing

A parallel between the history of industrial manufacturing and the history of software manufacturing can be made. (Figure 3)



Figure 3. Parallel between the history of industrial manufacturing and the history of software manufacturing

Before the industrial revolution the manufacturing was carried out by hand. In most cases the manufacturers were separate single individuals or rural households. The work was performed without well-established processes and working procedures. Together with the automation, the industrial revolution brought new manufacturing processes and production methods.³ Now, as the manufacturing is automated, the winners are those who find a way to lower the cost of this automation.

Internet of Things (IoT) in industrial manufacturing

“The Internet of Things is a network of physical objects – vehicles, machines, home appliances, and more – that use sensors and APIs to connect and exchange data over the Internet”⁴ By using IoT technologies, the automated manufacturing becomes Smart Manufacturing. “Smart Manufacturing is being predicted as the next Industrial Revolution.”⁵ The main boosters of the Smart Manufacturing are the enabled by the latest technologies unprecedented access to very large amounts of data (Big Data), and the contextualization of this data.

IoT in software manufacturing

In software manufacturing (or software development), the story is pretty much similar.

Time ago, there were Developers, QAs, Tools, Infrastructure, Platform, Release, InfoSec, Operations etc. passing the baton much like what happens in a relay race. All the builds, deployments, validation, releases were executed separately, by the responsible persons and in most cases – by hand. And there were Program/Project Managers “managing” this release process and responsible persons through meetings and emails. All this was very hard, long lasting, error-prone and quite expensive work.

DevOps is a “paradigm shift”⁶ to frequently released high quality software. Full automation is the key to achievement of this shift. And again - the winners are those who find a way to lower the cost of this automation.

The full automation and the latest technologies in place provide opportunity for gathering very large amounts of data (DevOps metrics) through a lifecycle of a software product, and for the

analysis of this data. The DevOps analytics is the Smart Manufacturing in software development.

Applying predictive and prescriptive analytics to DevOps metrics enables prediction of failures and business needs, as well as identification of areas for optimizations and actions for process improvements.

DevOps metrics

Here we are going to make a brief recall to our paper “Effective DevOps as a key to efficient large scale Agile software development - industry case study” - presented at CSECS 2017 and included in the conference proceedings – on how DevOps has been implemented in SAP - internationally recognized leading software company which is running multiple large scale agile development projects simultaneously.

Full automation

SAP developed a proprietary toolset that with a click of a button, automatically merges, builds, deploys, validates and releases a developer’s change (Figure 4). This enables to rapidly, reliably and repeatedly push out enhancements and bug fixes to customers at low risk and with minimal manual overhead.

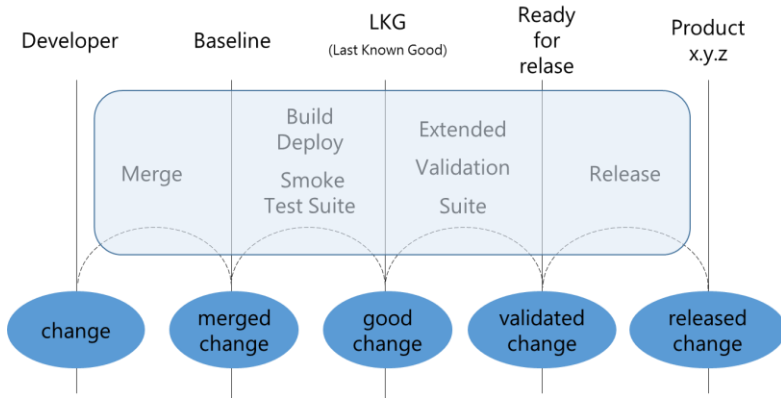


Figure 4. DevOps process automation in SAP – scope (within the blue frame)

Continuous monitoring

The DevOps automation toolset keeps all the history about requests, their changes, logs, authors, times, versions, etc. The statistical data being collected is available for further analysis. Figure 5 below shows an overview of the toolset components illustrating the continuous monitoring functionality.

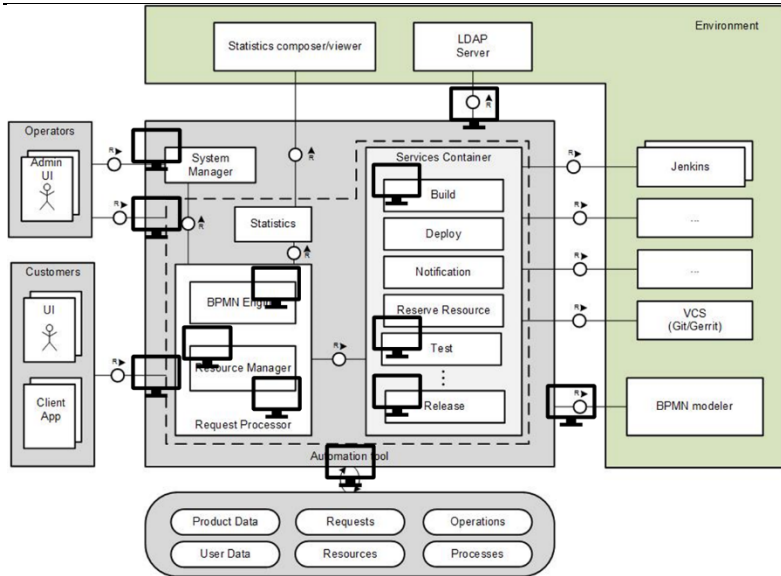


Figure 5. DevOps automation toolset in SAP – continuous monitoring

After the implementation of this complete DevOps solution providing full process automation and continuous monitoring, the next most logical step is to focus on how to define new metrics and identify the correct KPIs, which enable efficient application of predictive and prescriptive analytics, leveraging the data-driven approach to DevOps.

The data schema

We started with descriptive analytics on the data currently available. Figure 6 shows the current DevOps data schema.

The Change refers to the user stories in Jira (software development tool used by agile teams to plan, track, and release software). When the change in the software is ready, the developer (User) creates a Request to push his change in the productive branch of the desired Product. The change may affect several projects in the target product.

For each project change there is a RequestItem that represents the particular project and the particular change commit. The RequestItem is linked with the code repository from where data about the changed code can be gathered.

Each Request executes one Operation. The Operation consists of number of ordered OperationSteps. Each OperationStep executes one Job. Some of the Jobs use Resources that could be shared among the Requests and if the Resource is occupied by another Request, then the next Request should wait until the Resource is freed.

Of the available data, we extracted quantitative metrics about the code change requests. An important KPI among these metrics is the cost of the changes (Figure 7). The cost is being calculated for each change after the particular change is implemented, scoring the time for implementation, as well as the resources used for this implementation. The time for implementation includes not only the machine time needed to process the change, but also the time for retries and the time to wait for free resources.

Keeping the cost of the changes low means ensuring better effectiveness of the DevOps process. This was the reason for us to start our analysis from the cost of the changes.

| User_story | Author | Type | Changed_product | Changed_module | Changed_products | Changed_modules | Changed_files | Inserted_lines | Deleted_lines | Cost |
|------------|--------|------|-----------------|----------------|------------------|-----------------|---------------|----------------|---------------|------|
| 1 321 | 35 | 4 | 116 | 7 | 2 | 1 | 2 | 2 | 2 | 117 |
| 2 438 | 135 | 4 | 61 | 69 | 2 | 2 | 8 | 27 | 9 | 251 |
| 3 403 | 71 | 2 | 26 | 55 | 3 | 6 | 27 | 297 | 204 | 284 |
| 4 361 | 13 | 4 | 26 | 22 | 3 | 7 | 17 | 184 | 46 | 433 |
| 5 374 | 78 | 1 | 104 | 151 | 5 | 3 | 5 | 5 | 3 | 171 |
| 6 763 | 54 | 4 | 105 | 7 | 1 | 1 | 1 | 1 | 1 | 193 |
| 7 764 | 141 | 4 | 116 | 7 | 2 | 1 | 1 | 1 | 1 | 39 |
| 8 320 | 10 | 4 | 30 | 69 | 5 | 2 | 5 | 5 | 5 | 418 |
| 9 360 | 110 | 4 | 10 | 108 | 7 | 7 | 17 | 184 | 46 | 753 |
| 10 670 | 66 | 4 | 118 | 7 | 3 | 1 | 1 | 1 | 1 | 192 |
| 11 58 | 69 | 2 | 2 | 102 | 2 | 5 | 0 | 0 | 0 | 416 |
| 13 789 | 21 | 4 | 105 | 7 | 1 | 1 | 1 | 1 | 1 | 764 |
| 14 819 | 159 | 2 | 118 | 7 | 3 | 1 | 2 | 2 | 2 | 1644 |
| 15 394 | 50 | 2 | 119 | 7 | 4 | 1 | 3 | 3 | 3 | 33 |
| 16 528 | 93 | 4 | 105 | 7 | 1 | 1 | 1 | 1 | 1 | 81 |
| 17 359 | 156 | 4 | 116 | 7 | 2 | 1 | 4 | 4 | 4 | 24 |
| 18 732 | 127 | 2 | 61 | 69 | 2 | 2 | 3 | 3 | 3 | 164 |
| 19 705 | 55 | 2 | 116 | 7 | 2 | 1 | 1 | 1 | 1 | 19 |
| 21 861 | 130 | 4 | 105 | 7 | 1 | 1 | 1 | 1 | 1 | 26 |
| 22 820 | 54 | 4 | 118 | 7 | 3 | 1 | 1 | 1 | 1 | 1730 |
| 23 889 | 137 | 4 | 105 | 7 | 1 | 1 | 1 | 1 | 1 | 16 |
| 24 671 | 43 | 4 | 105 | 7 | 1 | 1 | 1 | 1 | 1 | 778 |

Figure 7. Change requests history

DevOps KPI analytics

Figure 8 below shows the statistical distribution of the single change costs in our historical data.

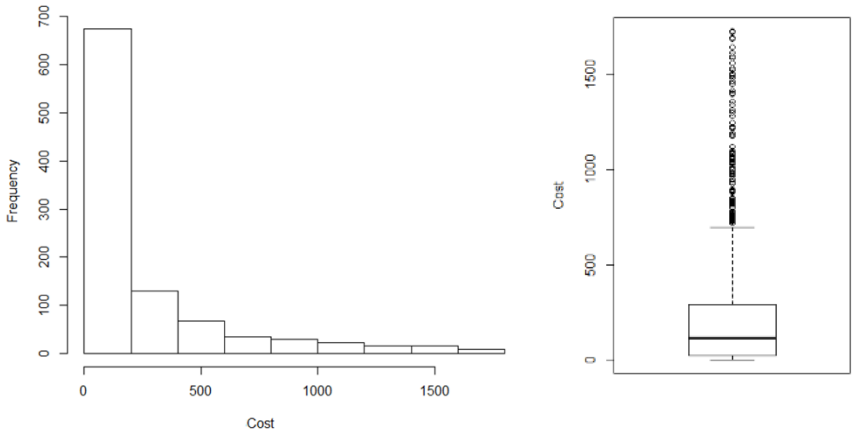


Figure 8. Distribution of the single change costs

We classified the change records in the historical data into low cost and high cost categories and labeled them accordingly. Changes with a cost of 120 which is the median of the costs distribution (see the boxplot on Figure 8 above) or less were classified as low cost changes, and respectively changes with cost of 121 or more were classified as high cost changes.

As it is clearly visible from the histogram on Figure 8, the majority of the changes are low cost changes. But in the same time the greater part of the total cost of changes is generated by high cost changes. A comparison between the low cost and high cost changes total costs is shown on Figure 9 below.

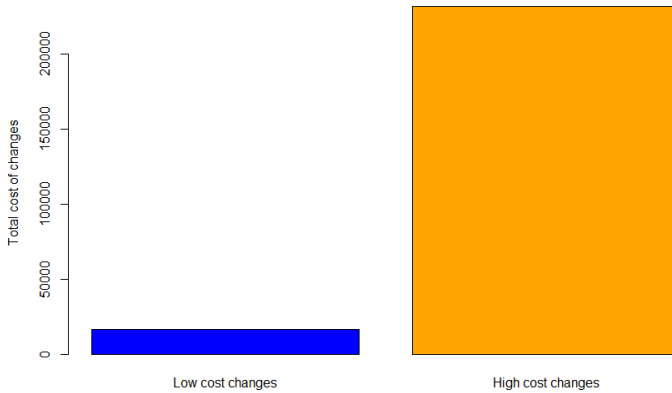


Figure 9. Low cost and high cost changes total costs

In order to be able to optimize the cost of the changes, first we need to be able to predict the potential cost of a new change request. In our case here, this task is not so trivial, and can't be solved by set up and application of simple business rules which are supported by the most of the DevOps tools as a standard functionality. This is because of the large scale of our products and projects, and because of the high number of contributors (software developers) to them.

Figure 10 shows the costs by changes' authors and types. As it is visible on the scatterplot, neither change author or change type is indicative for the change's cost. All of the more than 160 software developers which are authoring changes are submitting both low cost and high cost changes. Also changes from all of the four change types are being either low cost or high cost.

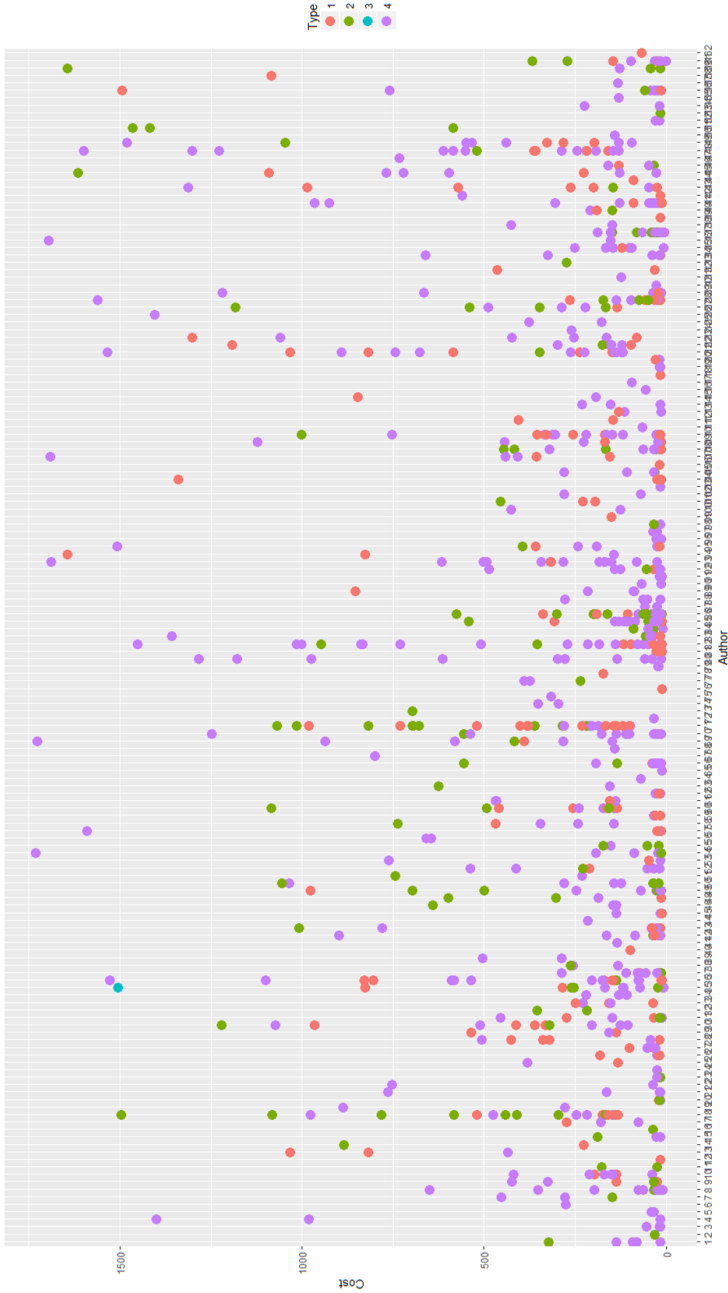


Figure 10. Costs of changes by authors and types

The rest of the changes' attributes such as the product and module being changed are also not indicative for the change's cost. There are both low cost and high cost changes in all products and in all of their modules. This in fact exhausted the possibilities for application of simple business rules for new code changes' cost prediction.

Another possible approach for prediction of a new change cost would be to use a statistical model.

We could build a multiple linear regression model in order to predict our dependent variable (the cost of a new change), using some of the change request's parameters (Changed_products, Changed_modules, Changed_files, Inserted_lines, Deleted_lines) as independent variables or predictors.

The problem with the linear regression model is that there is no strong correlation between the change cost and any of the change request's parameters. Figure 11 shows a correlation matrix of the change requests parameters. There is no any obvious linear relationship visible on the matrix. This will lead to low accuracy of the model.

Other disadvantage of the linear regression model in our case is that many of the change requests parameters are categorical variables. They won't be very efficient as a predictors in a linear regression model.

These were among the first indicators which directed us to the importance and the need of a well-established model of the metrics to be collected from each step of the DevOps process. The automated process and its continuous monitoring provides large amounts of data, but deeper data mining is necessary for more meaningful insights.

In that situation it was better to adopt classification techniques instead of just making predictions with high error rate. This allowed us to still utilize the data currently available, and to proof our concept for achieving better DevOps effectiveness with the help of data analytics.

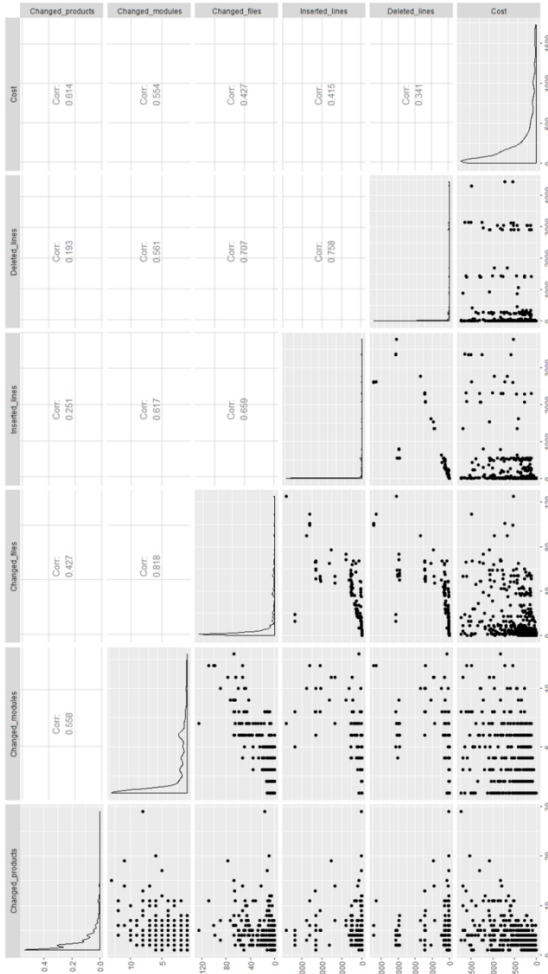


Figure 11 Change requests parameters correlation matrix

Because of the above described reasons, our selected approach was to apply machine learning classification algorithms in order to classify the new change requests into low cost and high cost categories. We use a standard classification process which looks as illustrated on Figure 12 below regardless of the particular technique.

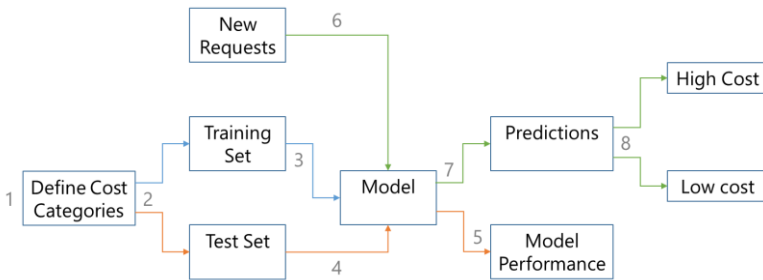


Figure 12. Applying machine learning classification algorithms

We have deployed different models such as k-nearest neighbors and logistic regression. The model selection is done dynamically based on model evaluation using the latest historical data. We have reached about 88% classification accuracy so far.

Being able to classify the new change requests in the queue into low cost and high cost categories is giving us different opportunities for optimization of the DevOps process.

We could estimate deployment times and prioritize the change requests accordingly. We could also provision additional resources on demand, dynamically, exactly when they are needed. The last has a direct financial impact in case of usage of cloud services.

Our next goal is to improve the classification model performance and even to become able to predict new change requests costs with reasonable accuracy by adding more meaningful data to our

models. Our strategic target is to explore furthermore metrics and to identify new currently unknown KPIs for DevOps effectiveness.

An important step toward our target would be the deployment of new DevOps OLAP database enabling deeper analysis and better reporting, including dashboards.

Model of the metrics to be collected

We work on research and development of comprehensive model of domain independent, industry commonly applicable DevOps metrics.

Metrics categories

We classified the metrics to be collected into 6 main categories: Development; Build | Validation; Deployment; Release; Production and DevOps Tools, covering the entire cycle of the DevOps process. Each category consists of sub-categories.

Table 1 below lists the DevOps metrics categories and sub-categories.

Table 1. DevOps metrics categories and sub-categories

| CATEGORY | SUB-CATEGORY |
|---------------------------|--------------|
| Development | Changes |
| | Tests |
| | Errors |
| | Developers |
| | Resources |
| | Times |
| Build Validation | Type |
| | Reliability |
| | Errors |
| | Resources |
| | Times |

Table 2 Lists the metrics in Development category grouped by their sub-categories.

| | |
|---------------------|----------------------------|
| Deployment | Frequency |
| | Changes |
| | Errors |
| | Resources |
| | Times |
| Release | Frequency |
| | BOM |
| | Resources |
| | Times |
| Production | Planned downtimes |
| | Outages |
| | Feature usage tracking |
| | User experience monitoring |
| | Direct feedback |
| | Tickets |
| | Compliance with SLAs |
| | Resources |
| DevOps Tools | Automation |
| | Feature usage tracking |
| | User experience monitoring |
| | Resilience |
| | Resources |

Table 2 – Development category DevOps metrics by subcategory

| SUB-CATEGORY | METRIC |
|--------------|--|
| Changes | Type: Feature request; Bug fix; etc. |
| | Changed projects |
| | Changed classes |
| | Lines of code |
| | Time for implementation (from task assigned to developer to change in VCS) |
| | Author (Developer) |
| | Reviewers (Developers) |
| | Test |
| Tests | Type: unit; integration; functional; performance |
| | Change |
| | Time for execution |
| | Author (Developer) |
| | Reviewers (Developers) |
| Errors | Type: functionality; test; infrastructure; performance |
| | Time for recovery |
| | Symptom |
| | Fix (Change) |
| | Phase of appearance (build, deployment, smoke validation, extended validation, performance validation) |
| | Effect (broken functionality, outages, performance) |
| | Ticket (customer ticket, internal ticket) |

| | |
|------------|-------------------|
| Developers | Position |
| | Education |
| | Trainings |
| | Skills |
| | Changes |
| | Reviews |
| | Tests |
| | Errors |
| Resources | Usage frequency |
| | Usage duration |
| | Cost |
| | Effectiveness |
| | Time for recovery |

Metrics from the other categories are currently under specification and refinement.

Benefits of well-established DevOps metrics

The benefits of well-established DevOps metrics and KPI analytics are not limited to the technical side. For instance, the Developers metrics category represents a software developer or operations engineer. He is the only human in the schema and is the author of all changes, tests, errors, reviews, etc. The Developer has the main role in the whole process. Thus, there is a good reason to pay a special attention to this particular category. Moreover there is a dedicated department (HR department) in each company that maintains very detailed database for each employee, including the software developers and operations engineers. It could be very useful for both sides – DevOps and HR, to incorporate the HR system into the DevOps analytics. If

DevOps infrastructure gets the full developer's HR profile, it could utilize pretty much everything out of it – the developer's position, education, trainings, skills, feedbacks, etc. DevOps could extend the developer's profile and combine it with more technical information: his changes, tests, errors caused by his changes, reviews. The technical part of the developer's profile will be continuously updated automatically. For the ease of analytics, the whole extended profile could be assembled and summarized into a single KPI – “rating”. The rating score will be calculated and being updated automatically. If a developer prepares many changes, with short times, covered with many tests, producing few errors, discovered at an early phases, then this developer will have a high rating.

The developer's rating could be efficiently used in the automated DevOps process. For example when a developer with a low rating creates a request to submit a new change, the automated procedure could require more code reviews, include extended validations, etc. If the developer's rating is used together with the characteristics of his change (changed projects, changed classes, changed lines of code, time for implementation) and the characteristics of the resources to be used in the automated procedure, this could bring quite a lot of value for analyzing and estimating/predicting the times (for the validation, for the used resources, for the whole procedure), the costs (for the used resources, for the engaged developers as reviewers), the bottlenecks, the risks of errors and their impact, etc.

The benefits from HR's point of view could also be significant. With the help of the DevOps analytics areas for improvement could be identified for each developer, and trainings, a mentoring, a courses or a specific technical literature for self-education can be recommended. After completing the particular

recommendation (internal training, for example), the DevOps infrastructure will update the developer's profile, recalculate the developer's rating score. Last but not least, the developer's rating could be a factor in the developer's salary or benefits. In case a developer leaves the company, the HR will have well refined requirements for his preplacement.

We could go even further and let the DevOps infrastructure track the effect from the completed training with the help of analytics. That way HR could get automated maintenance of the training profile or the trainer's rating. The DevOps analytics could also detect the need for specific course or training.

Conclusions

The DevOps effectiveness is crucial for development and delivery of high quality, competitive software products to the market in today's dynamic business environment. The automated DevOps processes and their continuous monitoring collecting large amounts of data making them available for further analysis. The data analytics are the key to optimization of the DevOps processes in the contemporary enterprises.

Collecting the right metrics, their efficient storage and access to them is decisive for the performance of the deployed analytical models.

Well-established DevOps metrics and KPI analytics provide number of benefits – from time and cost optimizations to quality improvements. The successful DevOps engineer of the 21st century analyzes data as part of his everyday job, applying descriptive, predictive and prescriptive analytics to deliver business value to his organization.

References

- Abbott, D. (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*. Indianapolis: Wiley.
- Baesens, B. (2014). *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*. Hoboken: Wiley.
- Bas, L., Weber, I., & Zhu, L. (2016). *DevOps, a Software Architect's Perspective*. Old Tappan, NJ: Addison-Wesley.
- Davis, J., & Daniels, K. (2016). *Effective DevOps*. Sebastopol: O'REILLY.
- Kim, G., Humble, J., Debois, P., & Willis, J. (2016). *The DevOps Handbook*. Portland: IT Revolution Press.
- Rozanski, N., & Woods, E. (2012). *Software Systems Architecture*. Upper Saddle River, NJ: Addison-Wesley.

¹ <https://blog.aspiresys.com/infrastructure-managed-services/devops-with-analytics-is-the-next-big-thing-in-software-development/>

²

https://www.ibm.com/cloud/garage/content/learn/practice_devops_through_analytics/

³ <https://en.wikipedia.org/wiki/Manufacturing>

⁴ <https://www.sap.com/trends/internet-of-things.html>

5 <https://www.manufacturingtomorrow.com/article/2017/02/what-is-smart-manufacturing--the-smart-factory/9166>

6 <https://www.cygnet-infotech.com/blog/devops-bringing-paradigm-shift-to-software-development-lifecycle-management>

Author's Information



Alexandrina Ivanova, MSc
Senior Software Developer, SAP Labs Bulgaria Ltd. alexandrina.ivanova@sap.com

Major Fields of Scientific Research: software development, software architecture, DevOps



Penko Ivanov, MSc, PMP, CBAP, PMI-PBA, PhD Candidate,

New Bulgarian University, Department of Computer Science penko.ivanov@gmail.com

Major Fields of Scientific Research: IT project management, business analysis in IT projects, requirements engineering, software functional architecture design, Big Data analytics, IT product marketing, IT business development

CSECS 2018, pp. 299 - 317

Computer Science and Education in Computer Science
14th Annual International Conference
ISSN 2603-4794

June 29 – 30, 2018, Boston, USA

END-USER APPLICATION FOR EARLY FOREST FIRE DETECTION AND PREVENTION

**Peter PEINL¹, Micha HEIDERICH¹, Ivan CHISTOV¹,
Jugoslav ACHKOSKI², Nikola KLETNIKOV², Igorche
KARAFILOVSKI², Nikola MANEV², Rossy GOLEVA³,
Alexander SAVOV³, Ivelin ADREEV⁴**

***Abstract:** In this paper, we describe a Web application that has been designed and implemented by Fulda University of Applied Sciences in the context of the ASPires project. The application extends the functionality available to Crisis Management Centers (CMC). Actual readings from sensors installed in the test areas, for example national parks, are made available to CMC personnel, as well as pictures from cameras that are either mounted on stationary observation towers or taken by Unmanned Aerial Vehicles (UAVs) in the area of an actual or supposed forest fire. Data are transmitted to the AspIRES cloud and*

¹ Fulda University of Applied Sciences, Fulda (DE)

² Military Academy Skopje (MK)

³ Comicon Ltd., Sofia, (BG)

⁴ ICB - InterConsult Bulgaria, Sofia (BG)

delivered swiftly to the Web application via an open interface. Furthermore, fire alarms raised by novel detection algorithms are forwarded automatically to the application. This clearly improves the potential for the early detection of forest fires in rural areas.

Keywords: *forest fire, fire detection systems, end-user, Web application, crisis management center*

ACM Classification Keywords: *A.0 General Literature – Conference proceedings; C.2.4 Distributed Systems – Distributed applications.*

Introduction

The ASPires project primarily aims at improving the capabilities of Crisis Management Centers (CMCs) to deal effectively with forest fires. The costs of fighting an actual fire and the damage resulting from such a fire depend heavily on the time to detect it. Therefore, ASPires (Advanced Systems for Prevention and Early Detection of Forest Fires) aims at developing and at assessing novel techniques and technologies (Wireless Sensor Networks (WSNs), specialized optical and thermal cameras and drones) to reduce the time of delivery of relevant data and information to Crisis Management Information Systems (CMIS). This will reduce reaction times and improve analyses in the early stages of a forest fire, or to prevent those. Additional services will be provided to the CMISs, like automatic alert generation based on the new and existing data.

The intended beneficiaries, called end-users herein, are primarily interested in the new and/or extended functionality of the technology. To corroborate its claims, the ASPires project has designed and implemented two graphical end-user applications that illustrate the novel features, services and benefits to the end-users. Those applications either might be stand-alone or integrated into existing CMISs. Both applications include all aspects and invoke all the components of the ASPires platform, i.e. they are end-to-end.

The focus of this paper is on the University of Applied Sciences Fulda (UFulda) application. Military Academy Skopje is the author of the other Web application.

ASPires System Architecture

Only those parts and aspects of the ASPires system architecture that are necessary to explain the role of Web applications in and their relationship to the elements of the overall architecture will be outlined. The figures in this Section are from Deliverable D.C.3 [1] of the project.

The major part of Figure 1 depicts devices and equipment of the novel technologies employed by ASPires to enhance the forest fire prevention and early detection capabilities of CMISs. There are sensors on the ground, cameras mounted on watch towers or carried by drones. Field gateways on the ground, carried by drones or mounted on watch towers, transfer information gathered in the field into the ASPires cloud, which is shown at the upper right corner of Figure 1.

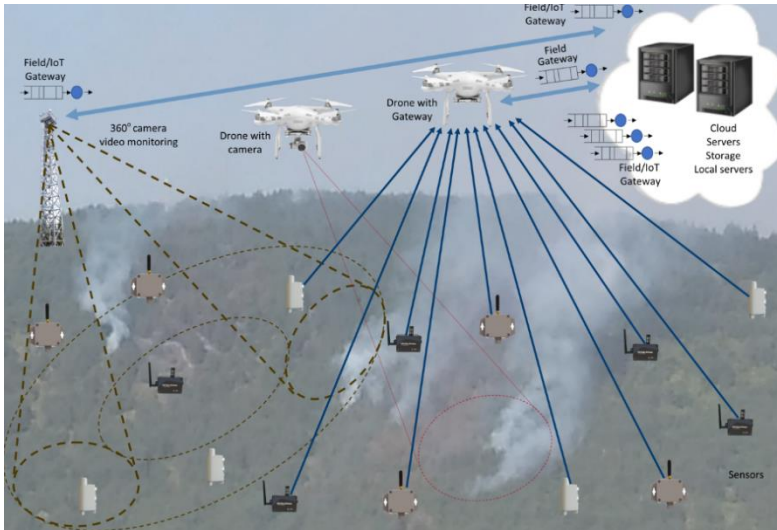


Figure 1. ASPires components and equipment

Figure 2 abstracts from the data gathering part of ASPires and puts the cloud in the center of the architecture. Conceptually, the cloud acts as the central storage area for the data arriving from the data capturing equipment. Raw data may be aggregated and analyzed. The cloud's storage capabilities enable the system to archive huge amounts of data on which new, additional services can be built.

The green line on the left side of the cloud represents all the interfaces (south-bound interface of the cloud) that ingest the the information captured by the equipment in the field into the cloud.

End-users, in general, are positioned to the right of the blue line (north-bound cloud interface) in Figure 2. Among them, CMCs, EFFIS and MKFFIS (European and Macedonian Forest Fire Information System) are of special interest to the ASPires project. The blue line comprises all interfaces that allow the end-users to

the right of the cloud to properly request and receive data from the cloud. Development of Web applications has to be based on those interfaces, as was done for the software described in this article.

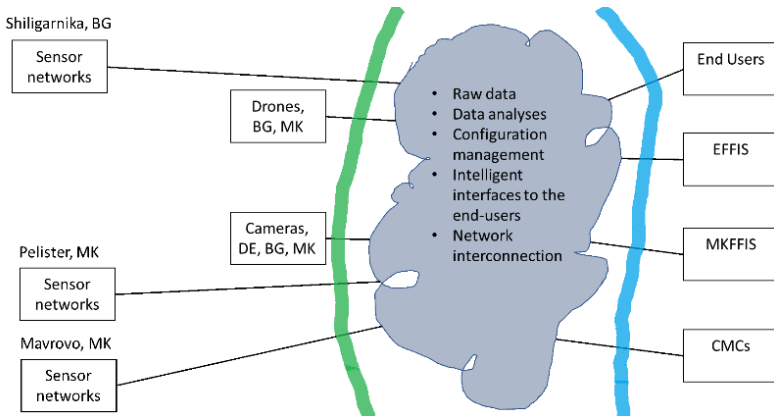


Figure 2. ASPires virtualized platform general view

End-users

A very general definition of the term end-user includes various groups of persons, organizations and institutions that might profit from the added benefits of an early forest fire warning and prevention system like ASPires. This is discussed in greater detail in Deliverable D.D.1 [2]. From a purely academic perspective, those groups might be identified, their needs analyzed and described and then specific applications be sketched. An incomplete list of different types of end-users in the wider sense and their potential needs and benefits of Web applications includes the following groups.

Citizens will profit from information generated by ASPires in the case of an actual forest fire, similar to KATWARN [3], the predominant catastrophe warning and information system in Germany. KATWARN informs about fires, heavy storms, thunderstorms, and other unexpected disasters. The content, timing and scope are decided by authorized institutions and security organizations.

Citizens might also send observations of an incipient fire to the proper authorities by a dedicated Web application or might communicate among themselves about the fire or get directions where to move and how to assist the authorities, i.e. what do to and, in particular, what not to do to impede the efforts of the fire brigades.

Fire-fighting and rescue teams and operations might be supported by different kinds of Web applications. Information has to be restricted to that particular group of users because of confidentiality. Information might come from the CMC or higher authorities. A Web application also might aid the fire fighters on the ground by offering closed user groups by secured (and monitored) communications.

CMCs are the authorized source of warnings to the general public and the proper management and control of fire-fighting on the ground. Therefore, they are the natural first recipients of novel information that is provided by ASPires.

Applications for ASPires End-users

The term application [4] needs to be defined first, because it often is not enough distinguished from the umbrella term software, which may encompass everything from computer games to Microsoft Excel.

A Desktop application is a native application that executes on a user's local machine. This application may or may not have a network component. An example of a desktop app would be MS Word, Adobe PhotoShop, or a Web browser.

A Mobile application is an application built to run natively on a mobile device. The most common devices at present either are iOS or Android based, but there are more alternatives. Mobile applications, similar to desktop apps, may or may not have a network component. An example of a mobile app would be Harvest for iOS or the SMS app.

A Web application is an application that runs a 100% within a browser. These applications are URL driven and work over the internet. There are a group of apps on both the desktop and on mobile that are just Web apps, running within an app-specific browser.

The decision to develop **Web applications** in the context of **ASPIres** was based on the following user needs and priorities [5]. First, the app needs to guarantee a high degree of reliability. Second, the app should be accessible over a wide range of devices and not depend on their OS or their hardware capabilities. Third, the app should be easy to build, feature an interactive, user friendly interface, and allow for regular updates and maintenance. All these characteristics imply that a Web application is the right and only choice for ASPIres.

Existing CMIS and Potential Improvements

The ASPIres project closely cooperates with the Macedonian CMC, which has a modern up-to-date CMIS (MKFFIS) at its disposal and considerable operational and technical know-how in

that domain. Some of the many benefits of MKFFIS are as following.

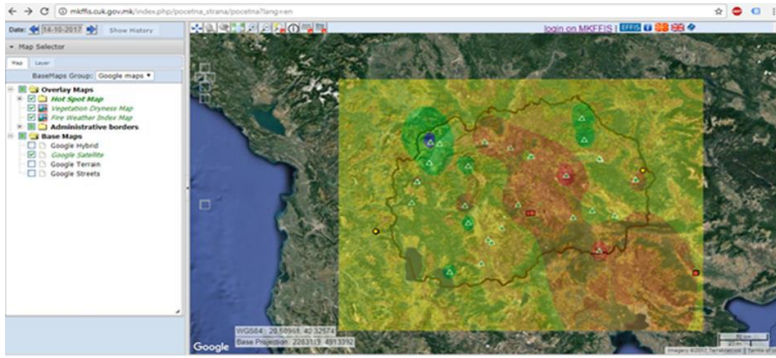


Figure 3. MKFFIS GUI of a CMIS in an existing CMC

The risk of forest fires and the danger they represent has been reduced. The prevention of and successful response to forest fires has been improved. The working conditions in the institutions that are a part of the wider crisis management system are much better than before.

However, MKFFIS is based only on satellite data which sometimes are too coarse and not as rapidly available as necessary. Detection of forest fires is often not reliable enough and fails to pinpoint the location of a wildfire. Figure 3 shows a typical screenshot taken from MKFFIS in one of the ASPires regions. No actual information about sensors and their readings is available. ASPires will close that information gap by adding novel technologies and thereby extend a state of the art CMIS and make it even better.

Overview of End-user Functionality

UFulda does not have sensors except for an IR camera at its disposal. So the Web application is entirely based on the ASPires cloud and the data from the ASPires data model retrieved via the northbound cloud interface.

The following screenshots are from the operational application. The layout is final with some improvements and refinements to be made in the course of the validation and test phase of the project. The cloud will then be populated with more data. They will become accessible to the UFULDA Web application without major software changes.

The functionality and information available to the end-users is structured into the following domains:

- Administration (including user and profile management)
- ASPires region page (map based information)
- Sensor information (including statistics)
- Camera information
- Alarms

User and profile management, as well as security and reliability are standard features of this type of application. Therefore, they will only be briefly summarized here. Besides the CMC personnel (role end-user), there is also an administrator of the system (role admin), who among others checks the credentials of new user to be registered user. A secure login process for different ASPires regions will be supported as well as the individual customization of the GUI. End-users may set multiple profile parameters like

language and display styles. The GUI will be described in the following by a series of commented screenshots.

ASPIres region page (map based information)

After successful login, the end-user is shown a map of the ASPIres region (Figure 4). All sensors and cameras installed in that region are represented by intuitive, self-explanatory icons. Their position in Figure 4 is based on geo-coordinates. Maps are fetched using the Google Maps API [6].

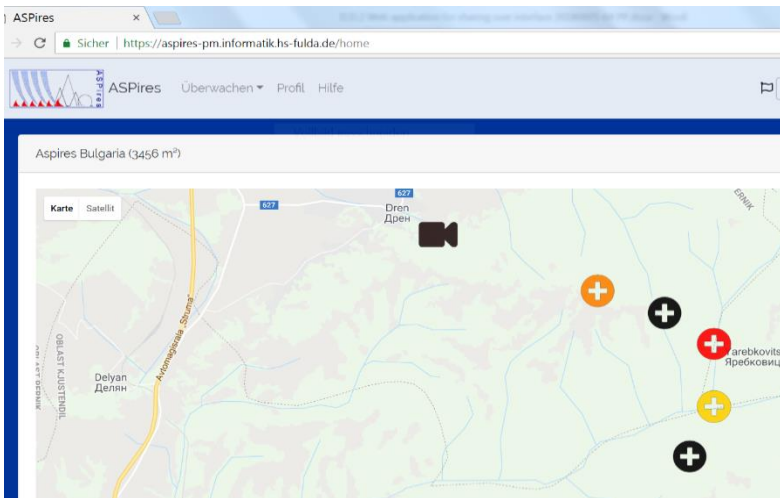


Figure 4. Full screen view of the ASPIres area Web page

All Information available to the end-user can easily be reached by clicking on icons or menus on this pivotal Web page. The map itself may be enlarged, scaled down or re-centered thanks to the built-in functionality of Google maps [6].

Each circle shaped icon (pin) points to a sensor site. By default, the colors in **Error! Reference source not found.** indicate the temperature. Black color signals that a sensor currently does not

deliver any temperature data. Colors, in general, are associated with a range of values. The color scheme is configurable.



Figure 5.List of all sensors in an ASPires region

A region might be very large or densely populated with sensors and cameras. Therefore, the Web application may display the complete list (Figure 5) of all available devices.

Detailed sensor data

Actual data from individual sensors may be obtained by a click on the sensor icon (pin) on the map (Figure 6). The map can be shown in several view style (satellite, terrain, etc.) supported by the Google API. The most recent reading of “Sensor Lisets” is automatically displayed above its icon. Figure 6 also tells the end-user that two other sensors (the black pins) do not deliver temperature readings. This could either mean that currently no temperature data is available or that this sensor does not capture this type of parameter at all. Sensor types and the physical or chemical parameters they capture will be discussed in the following paragraphs.

Sensors have been treated rather summarily in this section. Yet, the term sensor has a more general meaning in ASPires. It denotes a module potentially housing several devices, each

capturing one or more chemical or physical parameters, such as CO, CO₂, temperature, pressure, etc. In addition, there are parameters that report on the operational state of sensors. The plot in Figure 7 indicates a constant battery voltage of the sensor. Hence, it ought to be operational. **Error! Reference source not found.** also demonstrates that the Web application can plot time-series data.

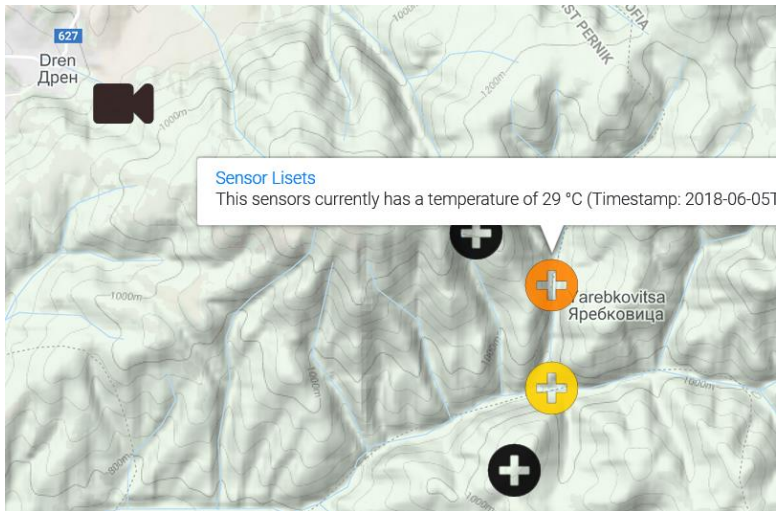


Figure 6. Most actual reading of selected sensor (terrain view)

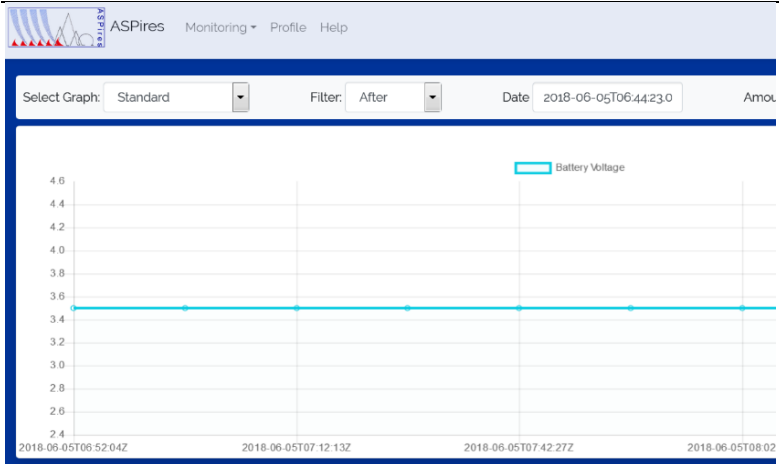


Figure 7. Displaying smoothed time series of captured sensor data (1)

The graphics package [7] used in the implementation automatically smooths the curves (**Error! Reference source not found.**). More display options are available through the main menu (before, after, between, amount of values).

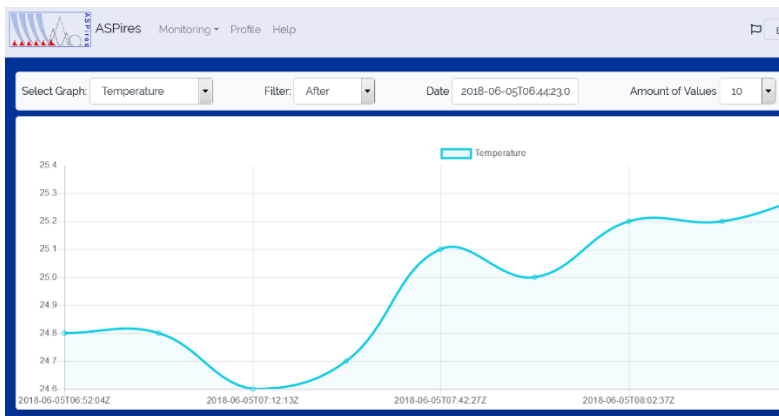


Figure 8. Displaying smoothed time series of captured sensor data (2)

Camera

Images taken by cameras constitute a very powerful type of “sensor” data made available by the ASPires project. Especially, cameras carried by drones add a flexible, mobile and low-latency component to the data capturing side of the project. Images, when available, are extremely helpful in verifying automatically generated alarms. Human experts may check the validity of alarms by analyzing images of the location of a supposed fire or of unusual sensor readings.

Figure 9 shows an image taken by an optical camera made available to the end-users through the Web application. Geo-coordinates, exact time of capture and an approximation of the period elapsed since the image was taken, are added at the bottom of the page. A gallery of most recent pictures is also available.

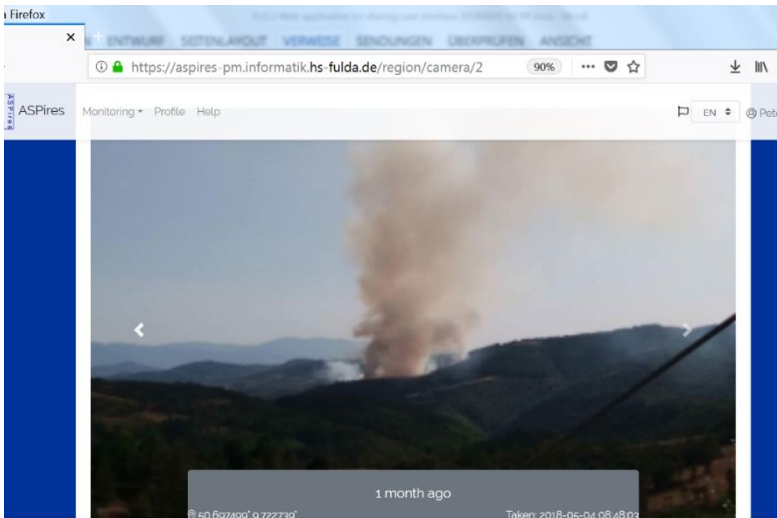


Figure 9. Displaying camera images

Alarm

The last, but extremely important, new feature that ASPires will give the CMCs, is automatic generation of alarms based on the information gathered by the sensor infrastructure. Though this feature will only become fully available in the course of the activities within the test and validation phase, the functionality to

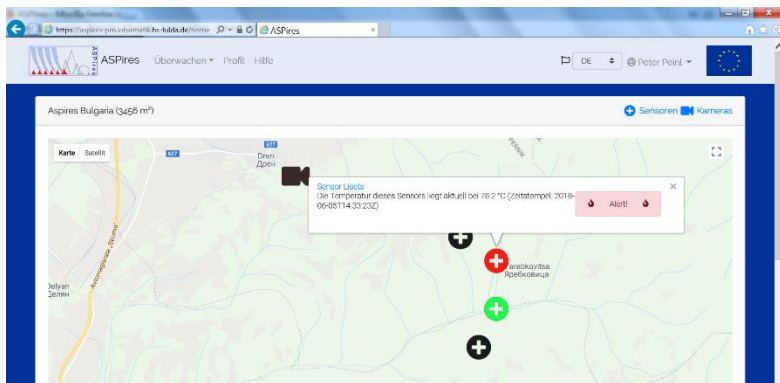


Figure 10. Alarm raised

display such an event has already been foreseen in the Ufulda Web application (Figure 10).

Conclusion

A Web application that enhances the functionality of existing CMIS in their capacity to detect and prevent forest fires in an early stage and that can easily be integrated has been designed and implemented in the context of the ASPires project has been described in great detail.

The viability of the approach and the benefits for the end-users are widely documented by means of an almost complete set of screenshots of the GUI of the Web applications.

The Web application is ready for test and validation in the test and evaluation phase of the project, where refinements and improvements will be applied in close cooperation with the end-users in the test areas.

New services and data not yet available in that form or with that speed in existing CMIS, like sensor data, including time-series, pictures from watch towers and/or drones and automatically generated fire alerts have been included. Those will be available to the CMCs.

The application also validates the claim made by ASPIRES to be able to develop end-to-end applications from the data capturing side, through the cloud to the end-users, for example at existing CMCs thereby improving their work and creating European added value.

Bibliography

- [1] D.C.3., "Advanced systems for prevention and early detection of forest fires design," *Deliverable D.C.3, ASPIRES project, ARES(2016)6878547-09/12/2016, European Commission*, 8 2017.
- [2] D.D.1., "Proper standards communication protocol between end-users experimental development and implementation," *Deliverable D.D.1, ASPIRES project, ARES(2016)6878547-09/12/2016, European Commission*, 9 2018.
- [3] KATWARN, "KATWARN German catastrophe information

- system," Fraunhofer FOKUS institute, 15 05 2018. [Online]. Available: <https://www.katwarn.de/warnsystem.php>.
- [4 D. Bychkov, "Desktop vs. Web Applications: A Deeper Look and Comparison," 03 06 2013. [Online]. Available: . Segue Technologies. June 7th, 2013. <https://www.seguetech.com/desktop-vs-web-applications/>. [Accessed 15 06 2018].
- [5 J. Summerfield, "Mobile Website vs. Mobile App: Which is Best for Your Organization?," [Online]. Available: <https://www.hswsolutions.com/services/mobile-web-development/mobile-website-vs-apps/>. [Accessed 10 06 2018].
- [6 Google Maps, "Google Maps Javascript API," Google Corporation, [Online]. Available: <https://developers.google.com/maps/documentation/javascript/?hl=en>. [Accessed 05 06 2018].
- [7 ChartJS, "ChartJS Diagrams," 02 05 2018. [Online]. Available: <http://www.chartjs.org/docs/latest/>. [Accessed 05 06 2018].

Authors' Information



Peter Peinl, Ph.D., MSc Computer Science, MBA, Professor of Computer Science at Fulda University of Applied Sciences; peter.peinl@informatik.hs-fulda.de.

Major Fields of Scientific Research: Database and Information Systems



Ivan Chistov., Dipl. Ing., Fulda University of Applied Sciences; ivan.chistov@informatik.hs-fulda.de

Major Fields of Scientific Research: Robotic Operating Systems, Embedded Systems, Computer Vision and Image Processing



Micha Heiderich., BSc Computer Science, micha.heiderich@informatik.hs-fulda.de

Major Fields of Scientific Research: Software Architecture, Web Application Design and Development



Jugoslav Achkoski, PhD Assistant Professor, Military Academy „General Mihailo Apostolski “Skopje, jugoslav.ackoski@ugd.edu.mk.

Major Fields of Scientific Research: Information Technology, Computer Science



Nikola Kletnikov, MsC, Military Academy „General Mihailo Apostolski “Skopje, nikola.kletnikov@ugd.edu.mk.

Major Fields of Scientific Research: Crisis management, Operation Planning



Igorce Karafilovski, MsC, Crisis Management Center “Skopje, igorce.karafilovski@cuk.gov.mk.

Major Fields of Scientific Research: Information Technology, Computer Science



Nikola Manev, BSc in Defence Resource Management, Military Academy “General Mihailo Apostolski” Skopje

Major Fields of Scientific Research:

Development, modeling, analysis and application of mechatronic systems, process automation and wireless sensor and actuator networks



Rossitza Goleva, Ph. D., Assistant-Professor, Department of Communication Networks Technical University of Sofia

Major Fields of Scientific Research: *Distributed Networks and Cloud Computing Communication Networks*



Aleksandar Savov, MSc in Computer Science, General Manager of Comicon Ltd., Bulgaria

Major Fields of Scientific Research: *Sensor Technologies, Distributed Networks, Automation*



Ivelin Andreev, MSc in Informatics is with Interconsult Bulgaria last 15 years.

Major Fields of Scientific Research: *Cloud computing, data mining, application design*

CSECS 2018, pp. 319 - 320

Computer Science and Education in Computer Science
14th Annual International Conference
ISSN 2603-4794

June 29 – 30, 2018, Boston, USA

BLOCKCHAIN SOLUTION FOR ANNUAL EVALUATION IN BULGARIAN SCHOOLS

Delyan Keremedchiev, Juliana Peneva

***Abstract:** The educational system in Bulgaria covers twelve grades. The quality of the education is measured at the end of the academic year. Students take written exams in the fourth, seventh and twelfth grade. The evaluation of these exams is concerned with a lot of paper work and documents exchange. The results of the evaluation should be provided to the authorities in a secure and reliable way. With the modern blockchain a school-based distributed computing environment can be created and many of the current difficulties can be overcome. By design, a blockchain is resistant to data modification and provides security and data integrity. In order to create such a blockchain each school has to support a blockchain based peer in its IT infrastructure. The Ministry of Education will be able to access all relevant data in a convenient and secure way. The blockchain technology will prevent manipulation of test results. Reliable information can be provided for interested institutions such as universities and potential employers. In this paper possibilities for blockchain*

implementation in schools which leads to improvement of Bulgarian school educational system are discussed..

Keywords: *Distributed database, blockchain, personal data, privacy.*

CSECS 2018, pp. 321 - 322

Computer Science and Education in Computer Science
14th Annual International Conference
ISSN 2603-4794

June 29 – 30, 2018, Boston, USA

INTRODUCTION OF BELL-LANCASTER METHOD AND LEARNING BY DOING INTO THE PRACTICAL CURRICULUM AT UNDERGRADUATE AND GRADUATE LEVELS

Valentina Ivanova, Mariyana Raykova
New Bulgarian University, Department of Informatics

***Abstract:** Micro IT projects are considered an educational innovation in the context of undergraduate education at department Computer Science at New Bulgarian University. The management of micro IT project in the educational context is challenging from organizational and academic point of view. This paper presents the introduction of Bell-Lancaster method into the practical curriculum at under-graduate and graduate levels from educational and from management perspective. The discussion covers the micro IT projects practicum set-up, the learning outcomes, the success rate of the projects and the students' feedback. The pilot started in 2013/2014 with two*

teams, now it is scaled up to than 23 teams with more than 80 participants.

Authors' Information



Valentina IVANOVA, PhD, Chief Assistant Professor, New Bulgarian University, 21, Montevideo Str., 1618 Sofia, Bulgaria; v.ivanova@nbu.bg

Major Fields of Scientific Research: Project Management, Software Engineering



Mariyana RAYKOVA, Ph.D. Chief Assistant Professor, New Bulgarian University, Informatics Department, 21 Montevideo Str., 1618 Sofia, Bulgaria, mariana_sokolova@yahoo.com.

Major Fields of Scientific Research: e-Learning, automatic test generation, programming

REPRESENTATIVE SAMPLE AS A LP PROBLEM

Dimitar Atanasov

New Bulgarian University
Department of informatics

Abstract: *Empirical analyses usually estimate the effect of some variables, considered as independent, on one (or more), considered as dependent. It is rare situation to have the whole population available for the analysis. Then, usually a subsample is chosen and the conclusions are generalized to the whole population. This approach is based on the concept that the sample represents, in some meaning the population. If a sample is large enough and the selection process is random, usually the sample is considered to be representative one. However in some cases the sample should satisfy some restrictions on selected characteristics, which are previously known for the population. Here we propose an approach to this problem as a solution of a task of integer linear programming.*

DON'T BE AFRAID TO COMMIT EXPERIENCES USING GITHUB CLASSROOM FOR TEACHING CS

Andrew Walfe

***Abstract.** Many of us teaching Computer Science are familiar with version control systems, whether from academic experience or from industry. In professional environments, such systems are often viewed as vehicles for delivering a product. However, in a very real sense these systems are also an individual vehicle for delivering their individual work for integration into a product. In this way, the conveyance of the CIS student's work for grading bears a lot of resemblance to version control. And, of course, students going into a professional environment are likely to be required to use such tools on a daily basis.*

At the same time there are many drawbacks to teaching computer science or CIS using Blackboard or other learning management systems ("LMS"). The most serious problem is that realistic assignments involve files that depend on other files and on runtime context, like particular folder or initialization files. Such context is lost in the flattened submission format of the LMS. In addition, LMS assignment structures also obstruct the provision of 'sample' or 'starter' files, and of the directory/folder structures often needed to make toolsets work effectively.

*In order to meet these needs, I decided to use a service called GitHub Classroom in three courses from Fall 2017 to the present. GitHub is well known as a colossal version-control repository server using the **git** distributed version-control system. (In 2017 GitHub had 24 million users with 67 million version-control repositories.) GitHub Classroom is a service that adapts the usual professional Git workflow to distribute individually to students; it is free to those with certifiable faculty positions. The first course on which I used GitHub Classroom was CS 689, the newly-reworked course "Designing and Implementing a Data Warehouse." However, in the middle of the Spring semester, I also deployed GitHub Classroom on the Term Project for CS 669, "Database Design and Implementation for Business."*

I found GitHub Classroom to be extremely useful and, with adaptations, intend to use it for future courses. This paper will present my experience as a case study on the mechanics, advantages, and drawbacks of using this service for teaching courses relating to computer science and information systems.

“Take the first step in faith. You don't have to see the whole staircase, just take the first step.”
-- Dr. Martin Luther King, Jr.

SEEING THE STAIRCASE: REFLECTIONS ON A FIRST SEMESTER TEACHING DATA MINING TO BUSINESS STUDENTS

Greg Page, Lecturer, Boston University - Metropolitan College

Abstract: In this paper, the author presents several challenges associated with teaching a technical subject that blends statistics, computer science, and specific subject-matter knowledge to an academically-accomplished but mostly non-technical audience. He also talks about some of the difficulties that he faced as an instructor dealing with three levels of newness: the course was brand new, so there was no prior blueprint upon which to fall back; the concepts presented in the course were completely new to the students; and some of the material was relatively new to him as well. After describing these challenges, he shares several lessons learned regarding instructional material and approaches that succeeded with the audience, as well as some candid reflections about areas in which he can improve future iterations of the course.

Among the areas explored in depth here are:

- The use of distributed technology, such as instructional coding videos and in-console coding lessons, to convey key ideas regarding programming and data mining;
- The need to design assessments that not only gauge student learning but help reinforce the areas that deserve the most emphasis;
- Ensuring that course material is challenging enough to ultimately prove rewarding to students, but without becoming so difficult as to overwhelm or discourage;
- Finding the most effective ways to take advantage of face-to-face meeting time during weekly class sessions of 165 minutes;
- Calibrating student expectations for the course.

Finally, the author shares his thoughts about the most important takeaways from this inaugural semester of *AD699: Data Mining for Business Analytics* and his vision for the next several iterations of the course.

USING GOOGLE BIGQUERY FOR DATA ANALYTICS IN RESEARCH AND EDUCATION

**Dimitar TRAJANOV¹, Ivana TRAJANOVSKA¹,
Lubomir CHITKUSHEV^{1,2}, Irena VODENSKA^{1,2}**

¹⁾ ss “Cyril and Methodius” University, Faculty of Computer
Science and Engineering, Skopje, Macedonia

²⁾ Boston University, Metropolitan College, Boston, USA

***Abstract:** Big data has been generating increased attention among researchers, scholars, businesses, media and even consumers. The trend of "big data growth" presents enormous challenges, but it also provides immense research and business opportunities. Large number of companies are entering this area and are developing and offering a wide spectrum of solutions and platforms. Google's BigQuery platform does not require infrastructural knowledge, and is one of the data analytics tools with shortest learning curve, which makes it very suitable for education in Data Science and introduction to Big Data. The performance of the platform and the flexible pricing model makes it also very suitable for Data analytics research, although the tool itself currently does not cover of all the phases of the Data science process.*

***Keywords:** big data, data science, open research, education, BigQuery.*

***ACM Classification Keywords:** Information systems → Information systems applications → Data mining.*

Introduction

Big Data [1] is data that exceeds the processing capacity of conventional database systems - the data is too big, moves too fast, or doesn't fit the structures of traditional database architectures. To gain value from this data, one must choose an alternative way to process it. Big Data technologies have a huge variety of sources, a large volume of information, and much shorter time required to process information thanks to parallel processing and clustering infrastructure. Data that has a continuing and massive impact on our business or personal lives is truly Big Data. Big Data provides unique insights into the behavior of complex systems in the real world. Data Science is an interdisciplinary field which involves using automated methods to analyze massive amounts of data and to extract knowledge from them. With implementation of automated methods for analysis of readily-available volumes of information, data science is creating today new research fields which are influencing many areas of social science and the humanity itself. Data from mobile sensors, complex instruments, and the Web is rapidly growing, and most of this available data can be processed with already existing automated methods or with the new ones that are not difficult to establish. Data Science provides a cross section of three major disciplines - computer science, applications, and mathematical statistics. With this combination of skills and science every single piece of data can be grouped, examined, researched, linked, etc. Using data science, the vast amounts of data can be turned into new insights and ultimately - knowledge [13].

A large number open source projects and commercial companies are developing tools and platforms for Big Data and Data Science [2] [14] [23]. Google as one of the global cloud providers is

developing its own platform named The Google Cloud Platform (GCP) [4]. GCP consists of a dozens of products and services, divided into several groups: Compute, Storage and Databases, Networking, Big Data, Machine Learning, Management Tools, Developer Tools, and Identity & Security. Google BigQuery which is a part of Google Big Data products and services is intended to be used as a Data Analytics tool. The BigQuery, was already proposed to be used as a tool in Introductory Technology Course [16]. Having in mind that the tool is offering a different approach than other similar products, and is enabling very short learning curve, we propose the usage of the BigQuery as a tool in Data Analytics education and also we describe its potential to be used as a research tool for Data Science. The rest of the paper is organized as follow: first we are giving a short overview of the Google Big Data Products, then we describe the Google BigQuery, and finally, through real example we are proposing methodology on how the tool can be used for educational and research purposes. We are presenting also a comparison of the proposed methodology and the performance of Google BigQuery and a Python based approach.

Google Big Data Products

The Google Big Data services currently consist of six products, that intent to cover different aspects of Big Data processing [5]: BigQuery, is a fast, economical and fully managed data warehouse for large-scale data analytics; Cloud Dataflow is a real-time data processing service for batch and stream data processing; Cloud Dataproc is a managed Spark and Hadoop service that is fast, easy to use, and low cost; Cloud Datalab is a powerful interactive tool that is used to explore, analyze and visualize data; Cloud Pub/Sub is a fully-managed real-time messaging service that allows sending and receiving messages

between independent applications; and Google Genomics helps the life science community organize the world's genomic information and make it accessible to general public.

The Google Cloud Big Data Reference Architecture is shown in Figure 1, and in addition to the services that are part of Google Big Data product line, two data storage services Cloud Storage and Cloud Bigtable are included. The Google BigQuery product take the role of analytical engine and provides output to Cloud Datalab service and several external data exploration products.

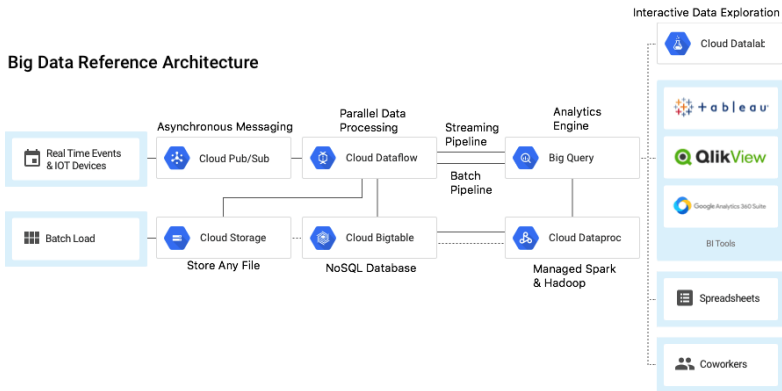


Figure 1. Google Cloud Big Data Reference Architecture [5]

Although it can work as part of this reference architecture, the BigQuery service is designed to be used as a standalone tool for data analytics. Google BigQuery is the one of most suitable cloud services when it comes to the Big Data analysis via Web technologies [7]. In contrast to other Big Data services, BigQuery provides an easy to use an interactive Web interface, the REST API and helper libraries for Java, .NET, Go, JavaScript, Node.js, Objective-C, PHP, Python, and Rubi. BigQuery is serverless, there is no infrastructure to manage, and there is no need for a database administrator, so the focus can truly be on analyzing

data to find meaningful insights, use familiar SQL, and take advantage of the pay-as-you-go model. There are also a variety of third-party tools that can be used to interact with BigQuery for visualization or for data upload and management.

BigQuery stores data using a hierarchy of containers called projects, datasets, and tables. The associated actions of uploading data, running queries, and exporting information are called BigQuery jobs [8]. Projects are top-level containers in Google Cloud Platform, and they are used for holding computing resources and user data. Also, projects store information such as billing data and authorized users. Each Google Cloud Platform project is distinguished by three identifiers: Project number, Project ID and Project name. Datasets are one level lower in the hierarchy than projects, and are used to cluster tables into groups. A dataset is assigned to a single project and the control of table access is done by using access control lists (ACLs). Tables contain the data, along with a corresponding table schema that describes field names, types, and whether specific fields are mandatory or optional. Tables are required in various data import and export data actions. BigQuery also supports views, which are virtual tables defined by an SQL query. Jobs are actions that are constructed by developers and executed by BigQuery on their behalf. Jobs include actions to load data, query data, export data, and copy data. Because BigQuery is typically used with large datasets, jobs may take a long time to execute, so there is an API call that is used to poll query status. Job history is saved for all jobs associated with a project.

The process of using and interacting with BigQuery starts with loading data into BigQuery Storage. To get the data back out of BigQuery, data can be exported. A table can be set up as a federated data source which allows using a query to transform

data as it is loaded. The next step is querying and viewing data. The web UI tool can be used to query the data interactively and then to export the results in CSV file or save the data as a table. The Web UI as shown in Figure 2, where the user can Save Query for future use, explore the results or see the explanation of the query execution. Querying data can be realized by calling the `bigquery.jobs.query()` or `bigquery.jobs.insert()` method with a query configuration. Viewing data is done by calling the `bigquery.tabledata.list()` or `bigquery.jobs.getQueryResults()` method. In addition to querying and viewing data, users can easily manage their data in BigQuery storage.

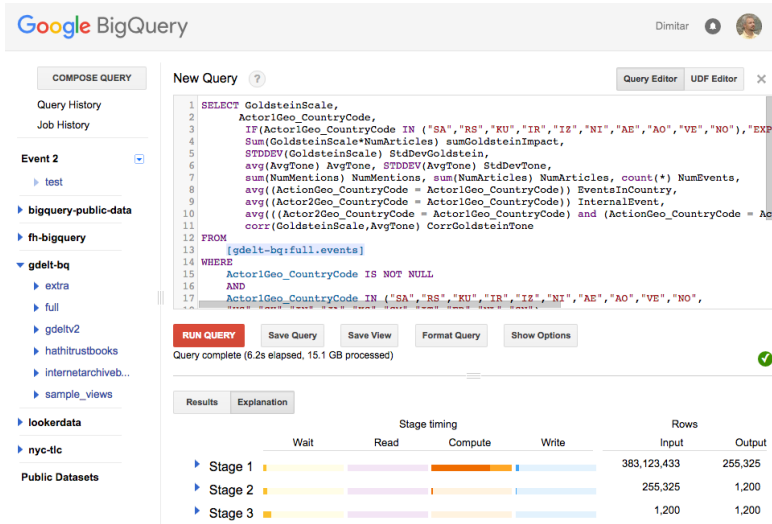


Figure 2. Google BigQuery Web User Interface

Google BigQuery Public Datasets

In order to boost the usage of its platform and as an example for the power of BigQuery, Google is hosting dozens of public datasets. A public dataset is any dataset that is stored in BigQuery which is made available to the general public. The collection

includes datasets like USA Names (provided by the U.S. Social Security Administration); NYC TLC Trips (taxi trips); Hacker News (developer-oriented social media posts); USA Disease Surveillance (weekly nationally notifiable disease reports); GDELT Book Corpus (digitized books from the Internet Archive); and NOAA GSOD (climate data from the National Oceanic and Atmospheric Administration), GitHub Archive (the dataset is automatically updated every hour), The Global Database of Events, Language, and Tone (GDELT) events dataset (all events that happen all over the world and are updated every 15 minutes) [6]. Along with the datasets mentioned above and listed by Google, much more are available, ranging from soccer data to cancer genomics [9] [10]. Developers can also share their own datasets simply by changing their permissions.

BigQueey Pricing

BigQuery offers scalable, flexible pricing options. BigQuery charges for data storage, streaming inserts, and for querying data, but loading and exporting data are free [15]. Storage of data currently costs \$0.02 per GB per month and \$0.01 per GB per month for a long time storage (if a table is not edited for 90 consecutive days). The queries are charged \$5 per TB, and the first 1 TB of data processed per month is free of charge (per billing account). Loading, copying and exporting data is free of charge. Query pricing refers to the cost of running your SQL commands and user-defined functions. BigQuery charges for queries based on the number of bytes processed. When you run a query, you're charged according to the total data processed in the columns you select, even if you set an explicit LIMIT on the results. The total number of bytes per column is calculated based on the types of data in the column. Google is offering 60 days

free trial account with \$300 free credit, so the users can use it to test all the features of the platform.

Using Google BigQuery in Education

The Google BigQuery is very easy to setup and use, making it quite convenient for introductory classes in Data Analytics and Data Science. In order to start using the tool, the Google account is needed, and the user needs to setup a project in Google Cloud Console, to create a service account and to enable billing for the project. Querying only public data sets doesn't require billing so that the users can test the tool for free. If you need additional data storage, or you are processing more than 1 TB of data, then the users can use the 60 days free trial.

The BigQuery doesn't require neither knowledge of any infrastructure setup, nor familiarity with database import or export tools. Only the basic knowledge of SQL is needed, which makes it an excellent option for using it in classes to teaching data analytics.

The existence of many continuous almost-real-time updated public datasets covering different areas, provides a unique opportunity for students to have hands-on experience with a powerful data analytics sets and tools, and to create applied real data driven application.

Using Google BigQuery in Research

The straightforward and easy to use infrastructure is especially beneficiary for data scientists, as they can focus on data analytics and use the power of BigQuery to handle the data storage and management. The platform is of superior performance and gives

an opportunity to manage data in many ways, and to test and explore different scenarios.

However, for a more comprehensive research and data analysis, some additional tools are needed, such as more sophisticated statistical or machine learning algorithms. One solution would be preprocessing of data in BigQuery and then using R, Python or RapidMiner to make the additional fine-grained analysis.

There is a number of research studies that have been done by using Google BigQuery. For example, in [17] the authors present the concept and implementation of a Big Data automation service based on BigQuery, which is used to store real-time process data from systems/machines in a cloud, from where it can be analyzed with user-specific algorithms. In [18] the analysis of public opinion about energy policy of the Spanish Government using the GDELT data set and BigQuery is presented. In the [19] the BigQuery is used for collection and mining of GitHub data, trying to understand GitHub user behavior and project success factors. The Predictive Analytical Decision Support System presented in [20] is based on BigQuery, and is used to connect healthcare organizations in order to share electronic health records and statistical reports. The next example of BigQuery usage is in [21] where it is used as a platform in a System for customer analytics on social media using for improving target advertising and improved business decision making. In [22] the authors are using the HTTPArchive data set that is publicly available on Google BigQuery to identify critical factors affecting software performance and to derive the performance metrics.

Case Study: Processing and Querying GDELT Data

In order to test the performance of BigQuery, we used the GDLET dataset. The GDELT dataset consists of over a 380 million event records in more than 300 categories covering the entire world from 1979 to present and connecting them in a large network of connected persons, organizations, locations, and themes [11] [12]. This large amount of almost real-time data is making it an ideal dataset for testing the BigQuery. To compare the ease of use and performance of the BigQuery we have set up the same analysis in Python.

Processing with Python

When working with Python, there is an option to download and extract only a subset of the data or to download the reduced format of the database and to perform an advanced data investigation. The reduced format of the database is a better option for making more general analysis and for processing the data on a local machine, although it is slower than writing queries on the endpoint. After downloading the data in its reduced format and writing the scripts in Python for processing and querying it, it has been executed on a local machine via multithreading using all eight threads, in order to get the results faster. Depending on the complexity of the queries and the requirements set in the scripts, the execution time can vary from 5 to 45 minutes. Thanks to Python's library Pandas there are limitless possibilities to process and query the data. The data have been read and transformed into Pandas DataFrame. The usage of SQL for querying the DataFrames is a good method to get and process the results in an adequate format. After this step, we apply different statistical or machine learning algorithms on the result dataset.

Processing with Google BigQuery

When making more precise analysis and processing the records in their original format, Google BigQuery is a better choice for that purpose. Querying the data is very fast, even when complex queries are implemented. The execution time varies depending on the query complexity but is always measured in seconds, which is one of the most significant advantages of using the endpoint.

Generally, the extracted data will contain the subset of fields that are required for analysis, and some aggregate function applied to them. The results provided a starting point for further analysis. With this method of data processing, we can save a lot of time and took advantage of complex data queries to get results. After that, we need to download the data in the CSV format that is used as an input the next step – visualization and advanced analysis. The final processing and visualization were done use Python, RapidMiner and Microsoft Excel.

One of the main disadvantages of BigQuery is the variety of aggregate functions supported by SQL. In our case study instead of Pearson product-moment correlation coefficient that is implemented by SQL CORR() Function we have a need to use Spearman's rank correlation coefficient. This was more complicated to implement in SQL than in Python where this function and many others are freely available and easy to use.

Comparison

Since GDELT is a really large data set, after approaching the two methods for processing the data a decision was made about which one is better. With Python, the downloading takes a long time and the datasets are too large to read entirely into Python unless the person is really familiar with the database and knows what is looking for. Also, there is a redundancy in the variables and the

dataset can be reduced to a smaller size. On the other hand, working with the reduced format of the dataset saves a lot of time, but it also suffers from missing information which was a big drawback for our research. In order to process the whole dataset on a local machine, and to do it fast and effective, it is required to have the right hardware and infrastructure, although that option is very expensive. For that reason Google provides running SQL queries using the processing power of Google's infrastructure. That is a big advantage and is very helpful in running the queries efficiently. Most important of all, Google BigQuery enables near-real-time querying over the entire GDELT dataset. After executing any complex or simple query, there is an option to download the results in a desired format. Google BigQuery is definitely the better choice for querying the GDELT dataset since Google is handling all the hard work. It is always fast and effective no matter how complexity the query is.

Conclusion

In this paper, we described the Google BigQuery platform, provided analysis and presented a set of recommendations on how the platform can be used in data analytics education and research. The main advantage of the BigQuery platform is its almost-zero-time setup effort, which makes it very useful for introductory courses in data science and data analytics. In addition, the BigQuery is hosting dozens of free datasets that can be analyzed almost for free (the first 1TB processed data every month is free), so the students can immediately try and test processing and analytics on big data sets.

In this paper, we presented a short survey of research work that was based on usage of BigQuery. We found out that the platform

has been used in different research fields such as event processing, energy, software projects analytics, social media, software performance, and healthcare. This further demonstrated that the tool is very easy to use and can be easily adopted in very different research projects.

In order to compare the BigQuery approach with a more classical approach based on data analytics, we presented a comparison of analysis of the GDEL dataset in Python and BigQuery. Our main conclusion is that BigQuery is showing superior performance and has very simple setup steps. However, to perform a more sophisticated analysis and to apply machine learning algorithms, Python or alternatives like R or RapidMiner are needed.

Bibliography

- [1] Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.
- [2] Xu, Xiang, Fumin Zou, Quan Zhu, and Xianghai Ge. "Comparison and Test for Several Typical Cloud Computing Platforms." In *Intelligent Data Analysis and Applications*, pp. 427-435. Springer International Publishing, 2015.
- [3] Feinleib, David. *Big data bootcamp: What managers need to know to profit from the big data revolution*. Apress, 2014.
- [4] Google Cloud Platform, <https://cloud.google.com/>, (accessed May 25, 2016)
- [5] Google Cloud Big Data Products, <https://cloud.google.com/products/big-data/>, (accessed May 25, 2016)
- [6] Google Public Data Sets, <https://cloud.google.com/bigquery/public-data/>, (accessed May 25, 2016)

- [7] Langmann, R. "Google cloud and analysis of realtime process data." In Remote Engineering and Virtual Instrumentation (REV), 2015 12th International Conference on, pp. 81-85. IEEE, 2015.
- [8] Gonzalez, Jose Ugia, and S. P. T. Krishnan. Building Your Next Big Thing with Google Cloud Platform: A Guide for Developers and Enterprise Architects. Apress, 2015.
- [9] David Ramel, Google Hosts Public Datasets for BigQuery Analytics, Application Development Trends Magazine, 03/30/2016, <https://adtmag.com/articles/2016/03/30/google-public-datasets.aspx>, (accessed May 25, 2016)
- [10] Reddit: Datasets publicly available on Google BigQuery, <https://www.reddit.com/r/bigquery/wiki/datasets>, (accessed May 25, 2016)
- [11] Kalev Leetaru, Philip A. Schrodt, GDEL: Global Data on Events, Location and Tone, 1979-2012, International Studies Association meetings, San Francisco, April 2013.
- [12] The GDEL Project, <http://gdeltproject.org/about.html>
- [13] McAfee, Andrew, Erik Brynjolfsson, Thomas H. Davenport, D. J. Patil, and Dominic Barton. "Big data." The management revolution. Harvard Bus Rev90, no. 10 (2012): 61-67.
- [14] Von Laszewski, Gregor, Javier Diaz, Fugang Wang, and Geoffrey C. Fox. "Comparison of multiple cloud frameworks." In Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on, pp. 734-741. IEEE, 2012.
- [15] Pricing - BigQuery — Google Cloud Platform, <https://cloud.google.com/bigquery/pricing> (accessed May 25, 2016).
- [16] Frydenberg, Mark. "Introducing Big Data Concepts in an Introductory Technology Course." Information Systems Education Journal 13 (2015): 5.
- [17] Langmann, R. "Google cloud and analysis of realtime process data." In Remote Engineering and Virtual Instrumentation (REV), 2015 12th International Conference on, pp. 81-85. IEEE, 2015.

- [18] Bodas-Sagi, Diego J., and José M. Labeaga. "Using GDELT Data to Evaluate the Confidence on the Spanish Government Energy Policy." *International Journal of Interactive Multimedia & Artificial Intelligence* 3, no. 6 (2016).
- [19] Chatziasimidis, Fragkiskos, and Ioannis Stamelos. "Data collection and analysis of GitHub repositories and users." In *Information, Intelligence, Systems and Applications (IISA)*, 2015 6th International Conference on, pp. 1-6. IEEE, 2015.
- [20] Neto, Silvino, and Felipe Ferraz. "Disease Surveillance Big Data Platform for Large Scale Event Processing.", ANCS'16, Santa Clara, CA, USA, March 17–18, 2016
- [21] Patel, Aditya, Hardik Gheewala, and Lalit Nagla. "Using social big media for customer analytics." In *IT in Business, Industry and Government (CSIBIG)*, 2014 Conference on, pp. 1-6. IEEE, 2014.
- [22] Patel, Charmy, and Ravi Gulati. "Identifying ideal values of parameters for software performance testing." In *Computing, Communication and Security (ICCCS)*, 2015 International Conference on, pp. 1-5. IEEE, 2015.
- [23] Chandrasekhar, Udaigiri, Arun Reddy, and Roi Rath. "A comparative study of enterprise and open source big data analytical tools." In *Information & Communication Technologies (ICT)*, 2013 IEEE Conference on, pp. 372-377. IEEE, 2013.

Authors' Information



Dimitar Trajanov, PhD, Professor, ss "Cyril and Methodius" University, Faculty of Computer Science and Engineering, Skopje, Macedonia, Ruger Boskovik 16 Skopje – Macedonia, dimitar.trajanov@finki.ukim.mk.

Major Fields of Scientific Research: Data Science, Semantic Web, Big Data, Computer Networks, Parallel processing



Ivana Trajanovska, Student, ss “Cyril and Methodius” University, Faculty of Computer Science and Engineering, Skopje, Macedonia, Ruger Boskovik 16 Skopje – Macedonia, trajanovska.ivana@students.finki.ukim.mk.

Major Fields of Scientific Research: Data Science, Semantic Web, Big Data



Lubomir Chitkushev, PhD, Associate Professor, Boston University, Metropolitan College, Boston, USA , LTC@bu.edu.

Major Fields of Scientific Research: Computer Networks, Health Informatics, Information Security, Complex Systems and Data Analytics



Irena Vodenska, PhD, Associate Professor, Boston University, Metropolitan College, Boston, USA , vodenska@bu.edu

Major Fields of Scientific Research: quantitative finance, complex networks, big data analytics

Computer Science and Education on Computer Science
(CSECS 2018)

14-th Annual International Conference

June 29 - 30, 2018 Boston, USA

Editors: Petya Assenova, Guanglan Zhang, Peter Peinl

Copyright © 2018

New Bulgarian University

21, Montevideo Str.,

1618 Sofia, Bulgaria

University of Applied Sciences

123 Leipziger Str.,

36037 Fulda, Germany

Boston University, MET

808, Commonwealth Avenue

02215 Boston, USA

ISSN 2603-4794