ITHEA

International Journal INFORMATION MODELS & ANALYSES

2012 Volume 1 Number 3

International Journal
INFORMATION MODELS & ANALYSES
Volume 1 / 2012, Number 3

IJ IMA is official publisher of the scientific papers of the members of the ITHEA® International Scientific Society

IJ IMA rules for preparing the manuscripts are compulsory.
The **rules for the papers** for ITHEA International Journals as well as the **subscription fees** are given on www.ithea.org .

The camera-ready copy of the paper should be received by ITHEA® Submission system http://ij.ithea.org .
Responsibility for papers published in IJ IMA belongs to authors.

General Sponsor of IJ IMA is the **Consortium FOI Bulgaria** (www.foibg.com).

International Journal "INFORMATION MODELS AND ANALYSES" Vol.1, Number 3, 2012

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with
Institute of Mathematics and Informatics, BAS, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politechnika de Madrid, Spain,
Hasselt University, Belgium
Institute of Informatics Problems of the RAS, Russia,
St. Petersburg Institute of Informatics, RAS, Russia
Institute for Informatics and Automation Problems, NAS of the Republic of Armenia,
and Federation of the Scientific - Engineering Unions /FNTS/ (Bulgaria).

Publisher: **ITHEA®**
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: info@foibg.com
Technical editor: Ina Markova

Printed in Bulgaria

ISSN 1314-6416 (printed) **ISSN 1314-6424 (CD)** **ISSN 1314-6432 (Online)**

POLYNOMIAL APPROXIMATION USING PARTICLE SWARM OPTIMIZATION OF LINEAR ENHANCED NEURAL NETWORKS WITH NO HIDDEN LAYERS

Luis F. de Mingo, Miguel A. Muriel, Nuria Gómez Blas, Daniel Triviño G.

Abstract: *This paper presents some ideas about a new neural network architecture that can be compared to a Taylor analysis when dealing with patterns. Such architecture is based on lineal activation functions with an axo-axonic architecture. A biological axo-axonic connection between two neurons is defined as the weight in a connection in given by the output of another third neuron. This idea can be implemented in the so called Enhanced Neural Networks in which two Multilayer Perceptrons are used; the first one will output the weights that the second MLP uses to computed the desired output. This kind of neural network has universal approximation properties even with lineal activation functions. There exists a clear difference between cooperative and competitive strategies. The former ones are based on the swarm colonies, in which all individuals share its knowledge about the goal in order to pass such information to other individuals to get optimum solution. The latter ones are based on genetic models, that is, individuals can die and new individuals are created combining information of alive one; or are based on molecular/celular behaviour passing information from one structure to another. A swarm-based model is applied to obtain the Neural Network, training the net with a Particle Swarm algorithm.*

Keywords: *Neural Networks, Swarm Computing, Particle Swarm Optimization.*

ACM Classification Keywords*: F.1.1 Theory of Computation - Models of Computation, I.2.6 Artificial Intelligence - Learning, G.1.2 Numerical Analysis - Approximation.*

MSC*: 68Q32 Computational learning theory, 68T05 Learning and adaptive systems.*

Introduction

The only free parameters in the learning algorithm are the weights of one MLP since the weights of the other MLP are outputs computed by a neural network. This way the backpropagation algorithm must be modified in order to propagate the Mean Squared Error through both MLPs.

When all activation functions in an axo-axonic architecture are lineal ones ($f(x) = ax + b$) the output of the neural network is a polynomial expression in which the degree n of the polynomial depends on the number m of hidden layers ($n = m + 2$). This lineal architecture behaves like Taylor series approximation but with a global schema instead of the local approximation obtained by Taylor series. All boolean functions $f(x_1, \cdots, x_n)$ can be interpolated with a axo-axonic architecture with lineal activation functions with n hidden layers, where n is the number of variables involve in the boolean functions. Any pattern set can be approximated with a polynomial expression, degree $n + 2$, using an axo-axonic architecture with n hidden layers. The number of hidden neurons does not affects the polynomial degree but can be increased/decreased in order to obtained a lower MSE.

This lineal approach increases MLP capabilities but only polynomial approximations can be made. If non lineal activation functions are implemented in an axo-axonic network then different approximation schema can be obtained. That is, a net with sinusoidal functions outputs Fourier expressions, a net with ridge functions outputs ridge expressions, and so on. The main advantage of using a net is the a global approximation is achieved instead of a local approximation such as in the Fourier analysis.

A variety of general search techniques can be employed to locate a solution in a feasible solution space, in our case neural network weights. Most techniques fit into one of the three broad classes. The first major class

involves calculus-based techniques. These techniques tend to work quite efficiently on solution spaces with friendly landscapes. The second major class involves enumerative techniques, which search (implicitly or explicitly) every point in the solution space. Due to their computational intensity, their usefulness is limited when solving large problems. The third major class of search techniques is the guided random search. Guided searches are similar to enumerative techniques, but they employ heuristics to enhance the search.

Evolutionary algorithms(EAs) are one of the most interesting types of guided random search techniques. EAs are a mathematical modeling paradigm inspired by Darwin's theory of evolution. An EA adapts during the search process, using the information it discovers to break the curse of dimensionality that makes non-random and exhaustive search methods computationally intractable. In exchange for their efficiency, most EAs sacrifice the guarantee of locating the global optimum. Differential evolution (DE) and Particle Swarm Optimization, see figures 9 and 10, are both stochastic optimization techniques. They produce good results on both real life problems and optimization problems. A simple mixture between those two algorithms, called Differential Evolution - Particle Swarm Optimization (DE-PSO), is also considered. The explanation will no longer use the sine function, but the more frequently used sphere function. Also note that the explanation for this algorithm will not use a single value, but arrays (vectors) to represent particles and velocities. Therefore, it is compatible with more dimensions.

Enhanced Neural Networks

The most usual connection type in neural networks is the axo-dendritic connection. This connection is based on the fact that the axon of an afferent neuron is connected to another neuron via a synapse on a dendrite, and modelized in *ANN* model by a weighted activation transfer function. But, there exists many other connection types as: axo-somatic, axo-axonic and axo-synaptic [Delacour,1987]. This paper is focused on the second kind of connection type *axo-axonic*. Merely, the structure of the axo-axonic connection can be sketched by three neurons with a classical axo-dendritic connection and the synaptic axonal termination of N_3 connected to the synapse S_{12}. The principle consists on propagating the action of neuron N_3 as synapse S_{12}. In order to model previous connection type, two neural networks are required [Mingo,1998]. The first (assistant) one will compute the weight matrix of the second (principal) one. And, the second network will output a response, using the previously computed weight matrix, this architecture is named Enhanced Neural Networks *ENN* [Mingo,1999; Mingo,1999a; Mingo,1999b].

Taylor Approximation

Taylor approximation degree 2 of a function n-differentiable at a point $x = a$ can be obtained using the following expression as a power series:

$$\hat{f}(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + e(\xi) \tag{1}$$

, where ξ belongs to interval $[x, a)$.

If $f'''(x)$ is a continuos function in the closed interval $[a, x]$ then this derivate has a maximun M in such interval, and therefore, the error in the aproximation (equation 1) is measure by [Blum,1991]:

$$\max \left| f'''(x) \right| \leq M \tag{2}$$

$$|e(x)| \leq \frac{1}{6} M \left| x-a \right|^3 \tag{3}$$

In case an approximation degree n of function $f(x)$ must be obtained, previous equations can be generalized in order to get:

$$\hat{f}(x) = \sum_{i=0}^{n} \frac{f^{i)}(a)(x-a)^i}{i!} + \frac{f^{n+1)}(\xi)(x-a)^{(n+1)}}{(n+1)!} \tag{4}$$

provided following constraints are verified:

1. $f^{i)}(x)$ corresponds to the i-derivate of $f(x)$. Besides $f^{0)}(x) = f(x)$.
2. If $i = 0$ then $i! = 1$.
3. ξ is a point at interval $[x, a)$.

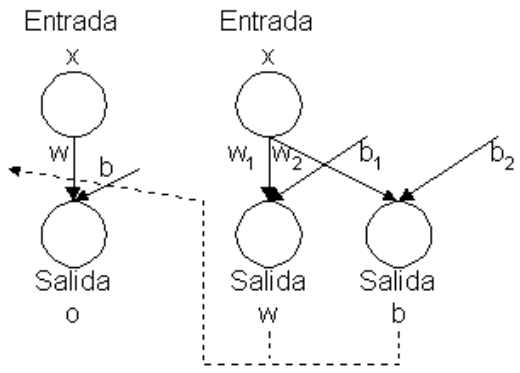
The approximation error, that is $f(x) - \hat{f}(x)$, can be measured if the $(n+1)$-derivate is a continuos function in interval $[a, x)$. Approximation error has a maximum defined by:

$$|e(x)| \leq \frac{1}{(n+1)!} M\,|x-a|^{(n+1)} \tag{5}$$

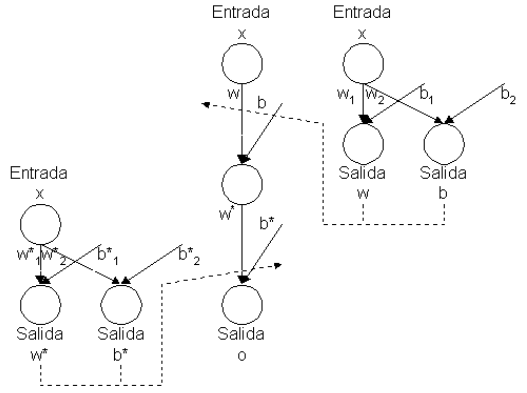
ENN as Taylor series approximators.

Above section has shown that a function can be approximated with a given error using a polynomial $P(x) = \hat{f}(x)$ with a degree n. The error $f(x) - P(x)$ is measure by equation (5) in such a way that in order to find a suitable approximation (error lower than a known threshold) it is only needed to compute sucessive derivates of function $f(x)$ until a certain degree n.

Enhanced Neural Networks behave as n-degree polynomial approximators depending on the number of hidden layer in the architecture. In order to obtain such behavior all activation functions of the net must be lineal function $f(x) = ax + b$.



$$\begin{aligned} o &= wx + b = \\ &= (w_1 x + b_1)x + w_2 x + b_2 = \\ &= w_1 x^2 + (b_1 + w_2)x + b_2 \end{aligned}$$



$$\begin{aligned} o &= w^*(wx + b) + b^* = \\ &= (w_1^* x + b_1^*)[(w_1 x + b_1)x + w_2 x + b_2] + = \\ &+ w_2^* x + b_2^* = \\ &= Ax^3 + Bx^2 + Cx + D \end{aligned}$$

Figure 1: ENN architectures and output expressions

As shown in figure 1 and output equations, the number of hidden layers can be increased in order to increase the degree of the output polynomial, that is, the number n of hidden layers control, in some sense, the degree $n + 2$ of output polynomial of the net.

Table 1 shows how the degree of the output polynomial increases according to the number of hidden layers in the net.

Table 1: Number hidden layers vs. degree of output polynomial

Hidden Layers	Degree $P(x)$	Output Polynomial
0	2	$o = a_2x^2 + a_1x + a_0$
1	3	$o = a_3x^3 + a_2x^2 + a_1x + a_0$
...	...	...
n	$n+2$	$o = \sum_{i=0}^{n+2} a_i x^i$

The only condition that the learning algorithm must verified is that weights must be adjusted to values related with the sucesive derivates of function $f(x)$ that pattern set represents. Usually such function is unkown therefore, if the network converges with a low mean squared error then all weights of the net have converged to the derivates of function $f(x)$ (the pattern set unkown function), and such weights will gather some information about the function and its derivates that the pattern set represents.

As an example, function $f(x) = sen(x)cos(x)$ can be approximated using equation (4), with a given point $a = 0$. Such equation can be reduced to $\tilde{f}(x) = x - \frac{4}{6}x^3$, using a polynomial $P(x)$ degree 3. This is a mathematical approach, but what happens if such function is the pattern set to an enhanced neural network mentioned before?.

A one hidden layer neural network must be used in order to obtain a 3-degree polynomial as the output expression. Figure 2 shows such architecture, after the training stage, the final configuration is shown. Output equation of the net is $o = x - \frac{4}{6}x^3$, equivalent equation with $\tilde{f}(x)$.

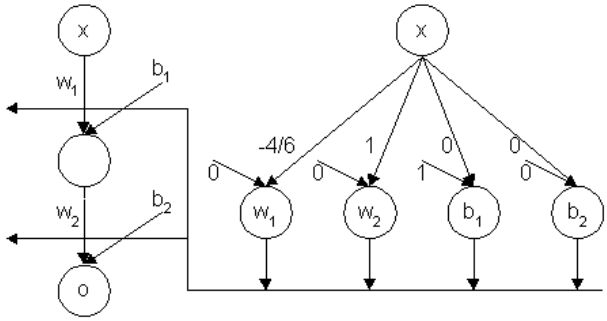


Figure 2: Approximation of $f(x) = sen(x)cos(x)$ with a one hidden layer

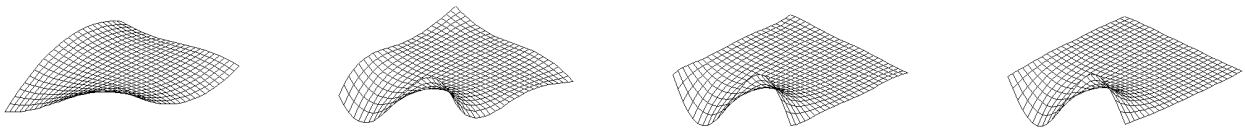
The approximation error using net in figure 2 can be computed using equation (5), and therefore $MSE \leq |e(x)|$. Such approximation is not the only one nor the best one, but it can be computed theoretically in order to provide the net some initial weights in order to speed up the learning process and to obtain a better approximation that the initial one with a lower error ratio. In sumary, Enhanced Neural Networks can be initialized to some weights computed using the Taylor Series of the function that the pattern set defines and after this initial stage the learning algorithm must be applied in order to achieved the best solution (the one that improves the Taylor Series error).

Figure 3 shows the surface computed by a net as the number of hidden layers is increased. The mean squared error is decreasing as the number of hidden layers goes up. This figure shows that this kind of neural net is very suitable when approximating functions, a given function or a function defined by the pattern set.

Non-Linear Activation

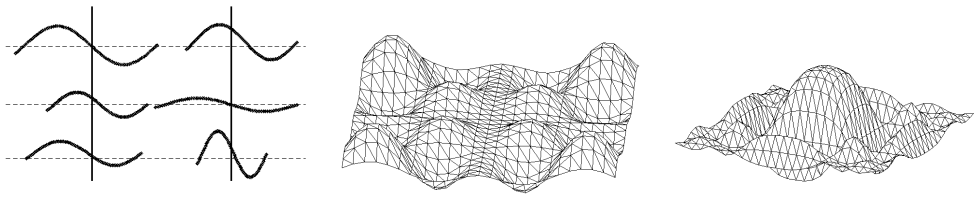
According to previous ideas, *linear ENNs* are better than linear *MLPs*, or at least, they are able to generate complex regions in order to divide the output space. When working with a *MLP*, only hyperplanes can be obtained. And moreover, the degree of the output equation increases according to the number of hidden layers.

Pattern set defined by $f(x, y) = sin(x)e^y$



2 hidden layers 4 hidden layers 6 hidden layers 8 hidden layers

Figure 3: Surface approximation depending on the number of hidden layers



a) Sinusoidal basis b) 3D Surfaces Approximation

Figure 4: Approximation with sinusoidal activation functions using basis of figure a).

In order to obtain a functional basis, one constraint must be made. It consists on implementing the network architecture with lineal *PE*s except the output neurons of assistant network. These neurons must have an activation function $g(x)$ which is used to computed the functional basis as the application of $g(x)$ to a non lineal combination of inputs. Figure 4 shows an example of a functional basis and the main network ourput.

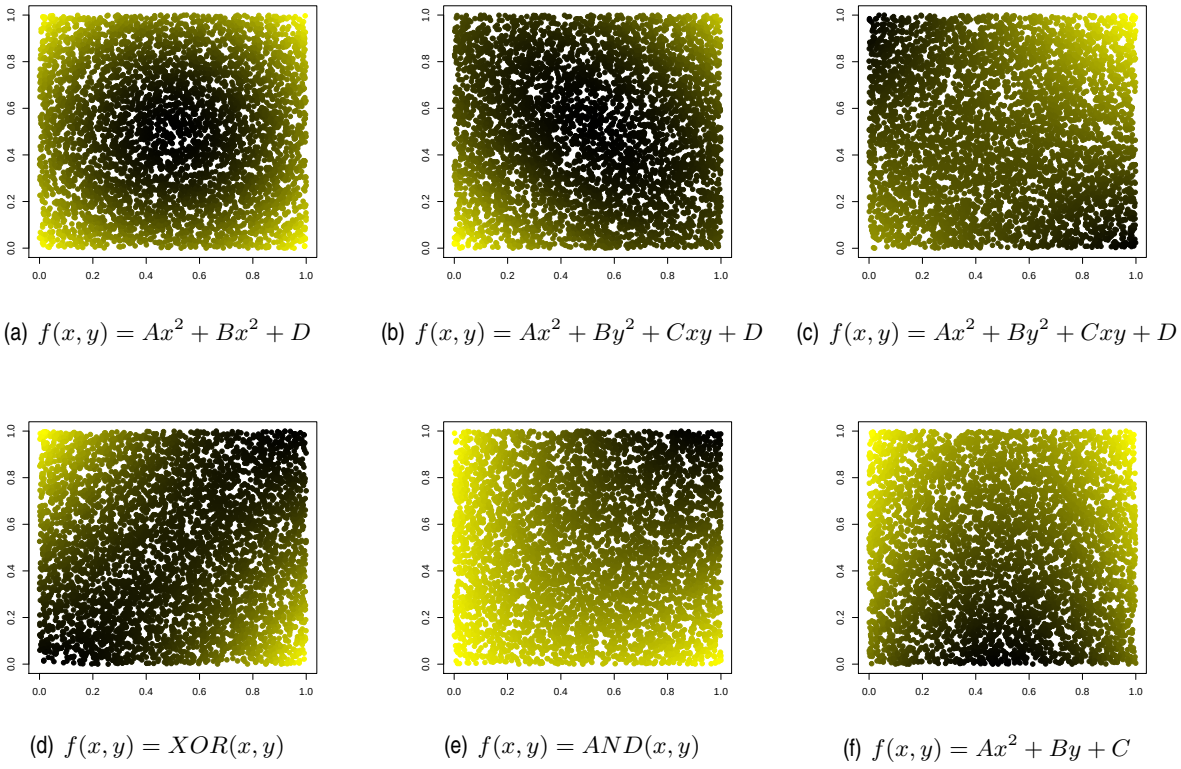
Depending on the activation function of output neurons belonging to assistant network, the main network will output an approximation function based on non lineal combination of elements belonging to the basis. That is if a sinusoidal activation function is implemented, then a cuasi-Fourier approximation is computed by the network; is a Ridge activation function is implemented, then a cuasi-Ridge approximation is computed and so on.

Main advantage of this new approximation method is that is absolutely easy to implement. And moreover, a global approximation to all the pattern set is perform. This way, if there are enough input patterns, then the generalization error will be minimized if there are enough learning iterations.

Enhanced Neural Networks as Universal Approximators

Along the paper [Mingo,1999a], this new architecture has shown that it is very suitable when dealing with any problem. Decision surfaces generated by the net are complex enough to represent any data set. The powerfull of these nets is in the number of hidden layers, that is, in the degree of the output polinomial associated to one output unit.

Funahashi Theorem can be directly apply to *Enhanced Neural Networks* in order to proof the universal approximation property of proposed networks, provided that activation function in hidden and output neurons belongs to a given class of functions stated by *Funahashi*. This way, *ENN* behave as universal approximators, that is, they are able to learn any pattern set.



(a) $f(x,y) = Ax^2 + Bx^2 + D$ (b) $f(x,y) = Ax^2 + By^2 + Cxy + D$ (c) $f(x,y) = Ax^2 + By^2 + Cxy + D$

(d) $f(x,y) = XOR(x,y)$ (e) $f(x,y) = AND(x,y)$ (f) $f(x,y) = Ax^2 + By + C$

Figure 5: Network output corresponding to 2-degree polynomial functions using an enhanced neural network with no hidden layers.

Results using Linear Networks with no hidden layers

The enhanced neural network architecture has the property that the output equation of an output unit is a n-degree polynomial if the activation function is a lineal one. This output can be compared with the Taylor approximation polynomials since the both methods are very similar. A set of data can be approximated by ENNs, computing a n-degree polynomial as the network output. But, the activation function can be a sinusoidal, instead being a lineal one. With this function, the approach of ENN is similar to Fourier series decomposition. This way, the activation function can be changed in order to get a better approximation than in the case of MLPs.

Figure 5 shows that the proposed network is able to learn different surfaces in a 2D space with a low MSE. This is mainly due to the special architecture of the net. The input to the net affects to the weights in the connection, and even changes them in order to optimize the error achieved by the net.

In a more complex pattern set, that is, a high dimension space, the proposed architecture is also stable, see figue 6 and note the correlation among all inputs and the correlation between the real output (OUT.1) and the desired response (OUT). Table 2 shows the final weights of the network.

Particle Swarm Optimization of Enhanced Neural Networks

Starting form general Particle Swarm Optimization algorithms formulas:

$$v_d^{(i)} = v_d^{(i)} + c_1\epsilon_1(p_d^{(i)} - x_d^{(i)}) + c_2\epsilon_2(g_d^{(i)} - x_d^{(i)}) \tag{6}$$

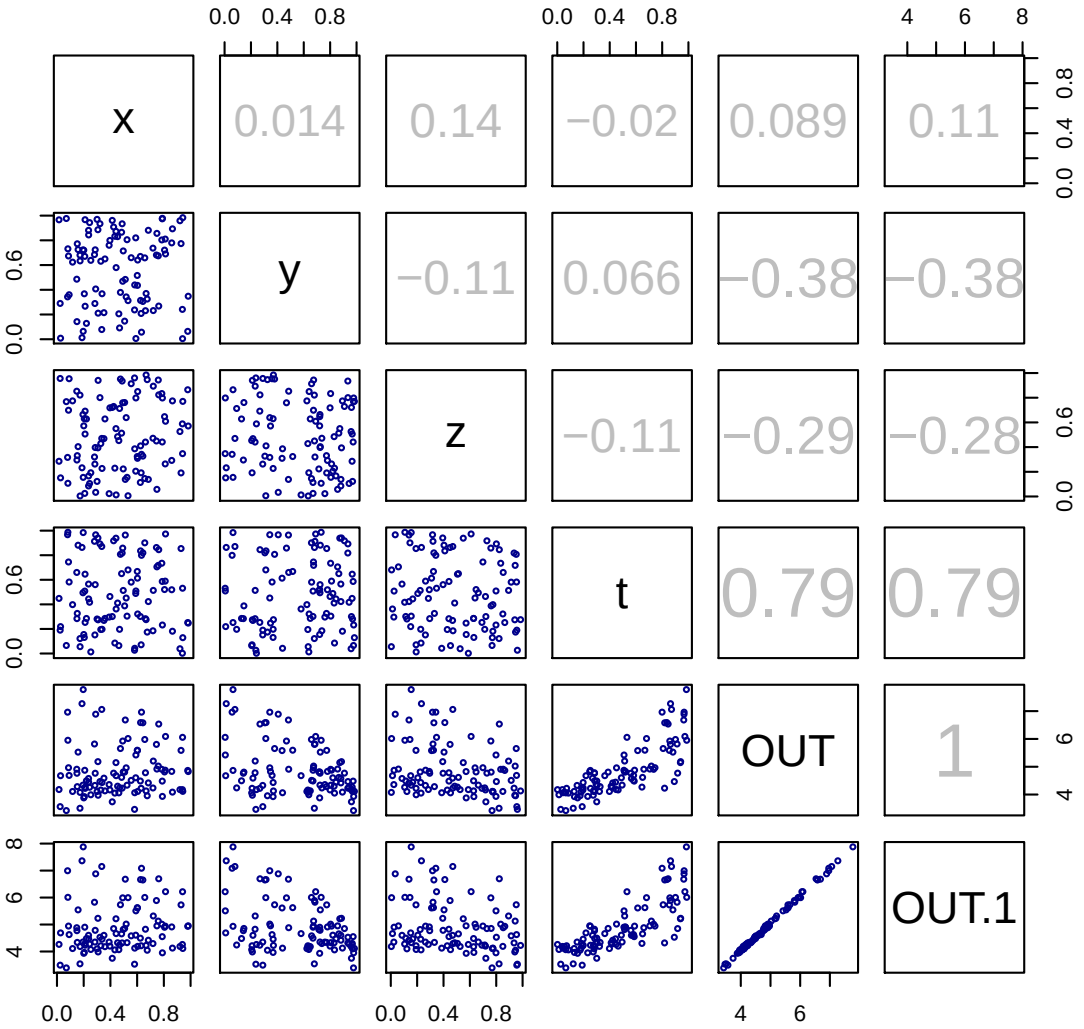


Figure 6: Correlation matrix of $f(x, y, z, t) = x^2z - (y + 1)^2t - z^2 + (t + 2)^2$, where $OUT = f(x, y, z, t)$ is the desired output and $OUT.1$ is the network output.

$$x_d^{(i)} = x_d^{(i)} + v_d^{(i)} \tag{7}$$

Regarding the PSO algorithm, different variants have been developed. Most of them aimed at speeding up the convergence of it. In addition to the unconstrained optimization problem in discrete or continuous variable, the multi target problem and the constrained problem have been addressed. We have also developed some hybrid optimization techniques. PSO technique has been tested with good results for training Artificial Neural Networks. When applying the method of Back Propagation, we are able to find appropriate weights that minimize an error function through a succession of iterations. Furthermore, by applying the PSO technique, the weights found are more efficient just by making small modifications to the algorithm. The new guidelines are aimed at avoiding PSO stagnation of the local optimal solutions.

Shi and Eberhart [Shi,1998] proposed adjustments to the velocities of the particles by using a factor w called *inertial weight*. This factor utilizes the inertia of the particles in the process of friction when they are moving. This modification in the algorithm is done to control the search space. In order to do that it must change (8).The large inertia weight makes the global search easier; however small inertia weight does not improve local search. That is why was the initial value is greater than 1.0 to promote global exploration, and then gradually decreases to obtain more refined solutions. The algorithm decreases linearly at each iteration. Moreover, the use of inertial weight removes the restriction V_{max} on the velocity.

$$v_d^{(i)} = wv_d^{(i)} + c_1\epsilon_1(p_d^{(i)} - x_d^{(i)}) + c_2\epsilon_2(g_d^{(i)} - x_d^{(i)}) \tag{8}$$

In each iteration, inertia weight decrease linearly through the following expression:

$$w = w_{max} - (w_{max} - w_{min})\frac{g}{G} \tag{9}$$

g is the index of the generation, G is the maximum number of iterations previously determined, w_{max} is a value greater than 1, and w_{min} a value under 0.5. This variation of the method has proven to accelerate convergence.

Clerc and Kennedy [Clerck,2002] obtain another variation in the speed calculation. A constriction factor χ is introduced with that purpose, This factor depends on the constants that are used when calculating speed and it affects to the formula (6) The aim is to avoid the explosion of velocity:

$$v_d^{(i)} = \chi[v_d^{(i)} + c_1\epsilon_1(p_d^{(i)} - x_d^{(i)}) + c_2\epsilon_2(g_d^{(i)} - x_d^{(i)})] \tag{10}$$

χ is:

$$\chi = \frac{2}{|2 - \varphi - \sqrt{\varphi^2 - 4\varphi}|}, \quad \varphi = c_1 + c_2 = 4.1 \tag{11}$$

Table 2: Weight matrix corresponding to the auxiliary network when learning function $f(x, y, z, t) = x^2z - (y + 1)^2t - z^2 + (t + 2)^2$.

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.69392712	-0.03879246	0.636080632	0.1224748	0.3380812
[2,]	-0.07873902	-0.55159000	-0.009702458	-1.9004765	-0.4532731
[3,]	0.54488180	-0.02171687	-1.068343779	0.1584132	0.2621986
[4,]	0.06375846	-1.23761019	-0.131404413	1.0906522	-1.2437301
[5,]	0.36486673	-0.22649428	-0.091435548	-1.8485381	4.0222552

Function	Dim.	Search space.	Name
$f_1(x) = \sum_{i=1}^{d} x_i^2$	30	[-100,100]	Sphere/Parabola
$f_2(x) = \sum_{i=1}^{d-1} [100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2]$	30	[-30,30]	Rosenbrock Generaliz.
$f_3(x) = \sum_{i=1}^{d} (\sum_{j=1}^{i} x_j)^2$	30	[-100,100]	Schwefel 1.2
$f_4(x) = \sum_{i=1}^{d} cos(x_i)^2$	30	$(-\infty, \infty)$	
$f_5(x) = \sum_{i=1}^{d} (x_i^2 - 10cos(2\pi x_i) + 10)$	30	[-5.2, 5.2]	Rastrigin Generaliz.
$f_6(x) = -\sum_{i=1}^{d} x_i \sin(\sqrt{\lvert x_d \rvert})$	30	[-500,500]	Schwefel Generaliz. 2.6

Table 3: Functions tested with PSO algorithm

The results are: $\chi = 0.729$ and c_1=c_2= 2.05. These parameters were obtained by performing several tests.

χ factor is similar to the inertial weight. This means that controlling the velocity with $V_(max)$ is not required when χ is used. Bratton and Kennedy [Bratton,2007], analyzed the stability of this algorithm by using these values and by following a comparative study of both PSO algorithms (inertial weight and χ factor). Both of them are mathematically equivalent, in particular the algorithm with constriction factor is a special case of the inertial weight. Moreover, Parsopoulos al [Parsopoulos,2002], combined both for problems with constraints and they obtained equally good results in several tests.

We observe that the convergence always becomes slower when problem size increase, so when it comes to high-dimensional problems, a larger number of iterations occurs. Researchers Hatanaka et al [Hatanaka,2007] developed a PSO model, where velocity values are updated, by considering the rotation of the coordinate system. This model is aimed at problems of high dimensionality and it showed good results when applied to all functions of De Jong, (larger dimensions).

In order to test the standard PSO algorithm and two variants with incorporated and inertial weight factor χ, we have used some unrestricted functions which are commonly referred as *De Jong functions*'. The minimum of these functions is located in the search space. They were originally proposed by de Jong to measure the performance of genetic algorithms. However they have also been used to test the performance in PSO algorithms. Some of the other functions are unimodal and multimodal i.e Ring (*lbest*) and star (*gbest*) topologies. In the ring topology each particle is related to its two neighbours. In the star topology all particles are interconnected. A population of 20 particles was considered. Table 3 functions are tested with the standard PSO algorithm and two variations: inertia weight and constriction factor. The first three functions are unimodal and have the optimal solution $x^* = 0.0^d$ and the minimum value $f_i(x^*) = 0.0$. The following are multimodal functions, the function f_4 has the minimum value 0.0, the optimal solution is $\pm n\frac{\pi}{2}^d$, the function f_5 has optimal solution $x^* = 0.0^d$ and the minimum value of the function: $f(x^*) = 0.0$ and f_6 has the optimal solution $x^* = 420.968^d$ and the minimum value of the function : $f(x^*) = -12.569, 4866$.

Tables (4) and (5) show the results after 20 executions of the standard algorithm. Not only the inertial weight has been modified in this algorithm but also the constriction factor for each functions by using neighbourhood models *lbest* and *gbest*. The algorithm stops when two successive values of the best assessments of the swarm get close to each other. (A ϵ value is prefixed and so it is a maximum number of iterations). NPE (average number of assessments) shows the average number of evaluations for the function when applying PSO and its variants.

After PSO Algorithm and its variants are executed, results are collected; best results are found in approximately equal number of cases regardless the model is used (*lbest* or *gbest*), so we can not assure which topology is optimal. Moreover, we notice that when using the constriction factor, the convergence accelerates and results are better when compared to the exact solution. In some cases, we notice that the region in which the swarm of particles initially are, can affect the results. Regarding the number of particles of the swarm, when increased to more than 20, results did not improve.

Name	PSO Original		Weight Inertial		Factor Constriction	
	Best Solut.	NPE	Best Solut.	NPE	Best Solut.	NPE
f_1	**l**: 3,70889e-07	190.480	7,68359e-07	28.800	2,69078e-08	19.100
	g: 5,9803e-04	2×10^5	2,45656e-20	151.500	3,63055e-20	30.820
f_2	**l**: 5,60357e-06	2×10^5	2,2112e-05	831.580	4,5089e-06	1.257.520
	g: 4,51068e-02	4×10^6	9,24376e-12	4×10^6	2,59676e-09	5×10^5
f_3	**l**: 5,76231e-06	198.280	5,57901e-15	1.425.900	2,83076e-16	399.140
	g: 4,91261e-06	2×10^5	1,81786e-12	2.596.780	1,44177e-13	125.640

Table 4: Results obtained by applying the PSO algorithm to unimodal functions, l indicates the model *lbest*; g refers to model *gbest*.

Name	PSO Original		Weight Inertial		Factor Constriction	
	Best solut.	NPE	Best solut.	NPE	Best solut.	NPE
f_4	**l**: 2.18719e-05	2×10^5	4.36373e-19	114.560	6,7306e-19	4.174
	g: 3,9543e-12	180.820	1.3000e-12	15.120	1.7063e-15	14.160
f_5	**l**: 1,70688e-14	216.240	1,81227e-14	339.020	1,07181e-11	9.480
	g: 3,00262e-04	2×10^5	6,82288e-12	12.180	2,81599e-12	11.040
f_6	**l**: -12.568,2	2×10^5	-12.569,5	2×10^5	-12.569,5	2×10^5
	g: -12.569,5	6×10^5	-12.569,5	6×10^5	12.352,3	2×10^5

Table 5: Results obtained by applying the PSO algorithm to unimodal functions, l indicates the model *lbest*; g refers to model *gbest*.

Particle Swarm Optimization and Neural Networks

Particle swarm optimization can be applied to solve many problems. One of them could be the training of a neural network architecture: Given a neural architecture, the problem is to find weights that minimize the mean squared error of the net. Individuals code weights of the neural network, and the fitness function corresponds to the mean squared error. According to *Kolmogorov* a multilayer perceptron can approximate any function even when the number of hidden neurons is unknown.

Obviously, a neural network with i input neurons, h hidden neurons and o output neurons it has $(i+1)h+(h+1)o$ weights and therefore, individuals of the PSO have $(i+1)h+(h+1)o$ dimensions. By considering such number, any real application with neural networks has at least 20 weights. A classical particle swarm algorithm could be applied however individuals have a high dimension and then convergence depends on the random initialization.

Figure 7 shows the learning curve of the PSO algorithm applied to a XOR neural network. This network has a $2-2-1$ architecture. It can be seen that the random initialization of individuals affect the convergence process (columns of figure). And the number of iterations (100 or 1000, at each row) achieves a lower fitness (mean squared error). Anyway, this simple example is solved with 10 individuals in the population, with dimension 9.

Another example is a binary coding neural network. An exclusive 8-bit vector coded it in a 3-bit vector. This classical problem can be solved by using a multilayer perceptron with 3 hidden neurons. Table below shows the input/output patterns of the neural network and the final weights found applying the PSO algorithm. In this case the dimension of individuals is 39 with a population of 15 individuals.

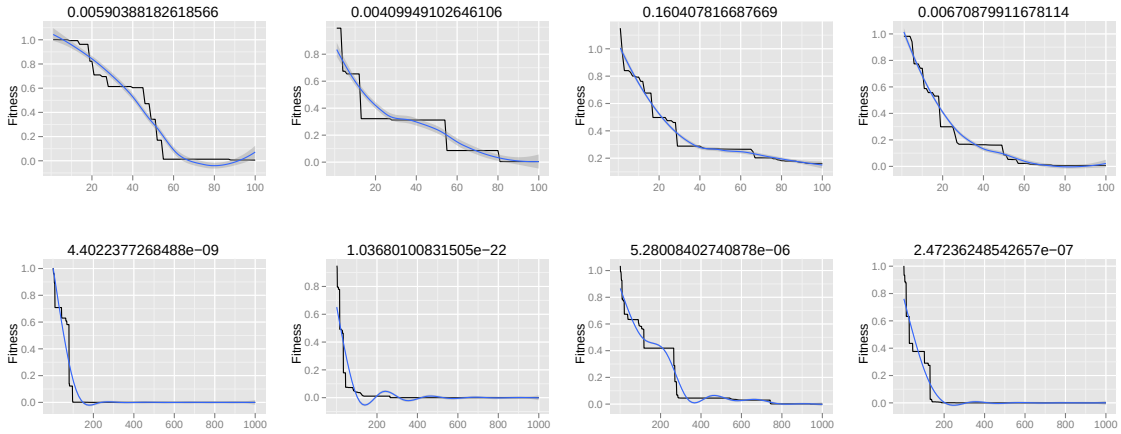


Figure 7: XOR multilayer perceptron with 2 hidden neurons and a particle swarm optimization learning using 10 individuals (individuals have 9 dimensions). Each column represents a different random initialization and each row a number of iterations (100 and 1000).

Input								Output		
1	1	1	1	1	1	1	-1	1	1	1
1	1	1	1	1	1	-1	1	1	1	-1
1	1	1	1	1	-1	1	1	1	-1	1
1	1	1	1	-1	1	1	1	1	-1	-1
1	1	1	-1	1	1	1	1	-1	1	1
1	1	-1	1	1	1	1	1	-1	1	-1
1	-1	1	1	1	1	1	1	-1	-1	1
-1	1	1	1	1	1	1	1	-1	-1	-1

Best fitness value: 5.067e-05

Best neural network weights with a $8-3-3$ architecture:

Input layer $\rightarrow$ Hidden layer

```
          0.1156528   1.097272  -0.946379977
         -0.3683220  25.945492  -1.703378035
          1.5325933  -9.765752  -0.636187430
          0.4830886  26.536611   0.002121948
         -1.5133460   0.790667  -0.131921926
         -1.7465932  -3.369892   0.987214704
          1.0519552  -4.920479   0.005404300
          0.3397246  -1.924014   3.273439670
Bias:     0.7092471  -5.304714  -0.331283300
```

Hidden layer $\rightarrow$ Output layer

```
          1.261584  -4.759320   0.9853185
        148.541038  -1.054461  -3.1161444
       -162.388447  -2.017318  -3.4691962
Bias:     0.090192   1.207254   1.9260548
```

These two neural examples have shown that the PSO can be successfully applied to the particle swarm algorithm in order to solve, in some way, the convergence of the algorithm when dealing with high dimension individuals.

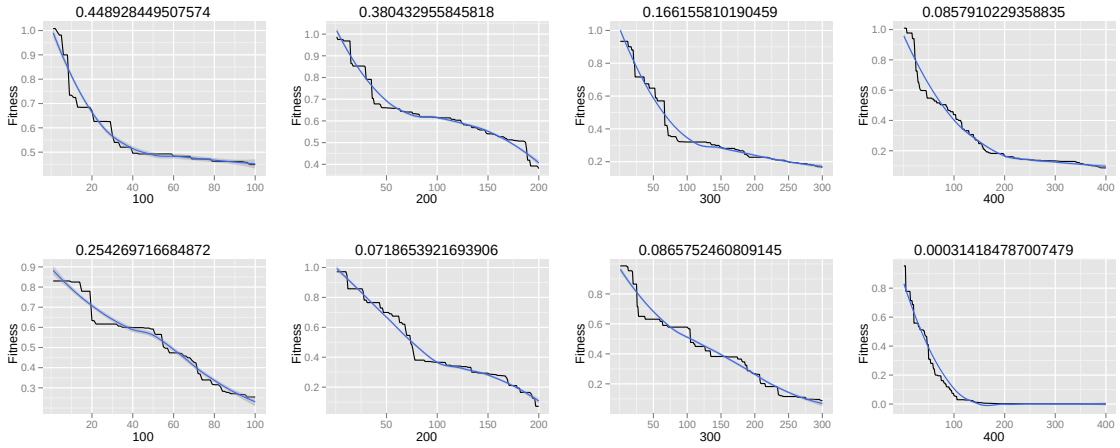


Figure 8: Binary coding neural network form 8 inputs to 3 outputs. First row is a neural network with 3 hidden neurons and second row a neural network with 5 hidden neurons. Mean squared error (fitness of the PSO) decreases as the number of iterations (100 to 400) increases.

The XOR example with dimension 9 and the binary-coding example with dimension 39 are a good starting point to combined classical neural networks with swarm intelligence.

Conclusion

The problem of nonlinear constrained optimization arises frequently in engineering. In general it does not have a deterministic solution. In the past, nonlinear optimization methods were developed and now it is a challenge to work with differentiable functions. Before gradient methods were used successfully for solving some problems. Evolutionary methods provide a new possibility for solving such problems. The PSO technique has been used successfully in optimizing real functions without restrictions, but it has been little used for problems with restrictions. This has happened mainly because there are no mechanism to incorporate restrictions on the *fitness* function. Evolutionary Computation has tried to solve the constrained optimization problem, either by bypassing nonfeasible solutions sequences, or by using a penalty function for nonfeasible sequences. Some researchers suggest to use two subfunctions of *fitness*. One helps to evaluate feasible elements and the other one evaluates the unfeasible one. In this regard, there are many criteria. Moreover, some special self adaptive functions have been designed to implement the penalty technique.

Hu and Eberhart [Hu,2002] presented a PSO algorithm. This algorithm bypasses nonfeasible sequences. it also creates a random initial population, in which nonfeasible sequences are bypassed until the entire population has only feasible particles. By upgrading the positions of the particles nonfeasible sequences are bypassed automatically. The cost of the technique that creates the initial populations is high; especially when it comes to problems with nonlinear constraints because then it must create an entire population of feasible individuals. In his work, Cagnina et al [Cagnina,2008] proposed the following strategies for implementing the PSO into problems with restrictions: **a)** If two particles are feasible, select the one with the best *fitness*. **b)** When a particle is feasible and the other is not, the feasible one is chosen. **c)** If two particles are nonfeasible, the one with the lowest degree of nonfeasibility is selected. These strategies are applied when the particles *gbest* and *lbest* are selected. The same authors also proposed an update in (6). This update considers three elements:

1) $p_d^{(i)}$ which is the best position reached by the particle i in its history. 2) $g_d^{(i)}$ which is the best position reached by the particles in its neigborhood and t_d which is the best position achieved by any particle in the whole swarm.

$$v_d^{(i)} = w(v_d^{(i)} + c_1\epsilon_1(p_d^{(i)} - x_d^{(i)}) + c_2\epsilon_2(g_d^{(i)} - x_d^{(i)}) + c_3\epsilon_3(t_d - x_d^{(i)})) \tag{12}$$

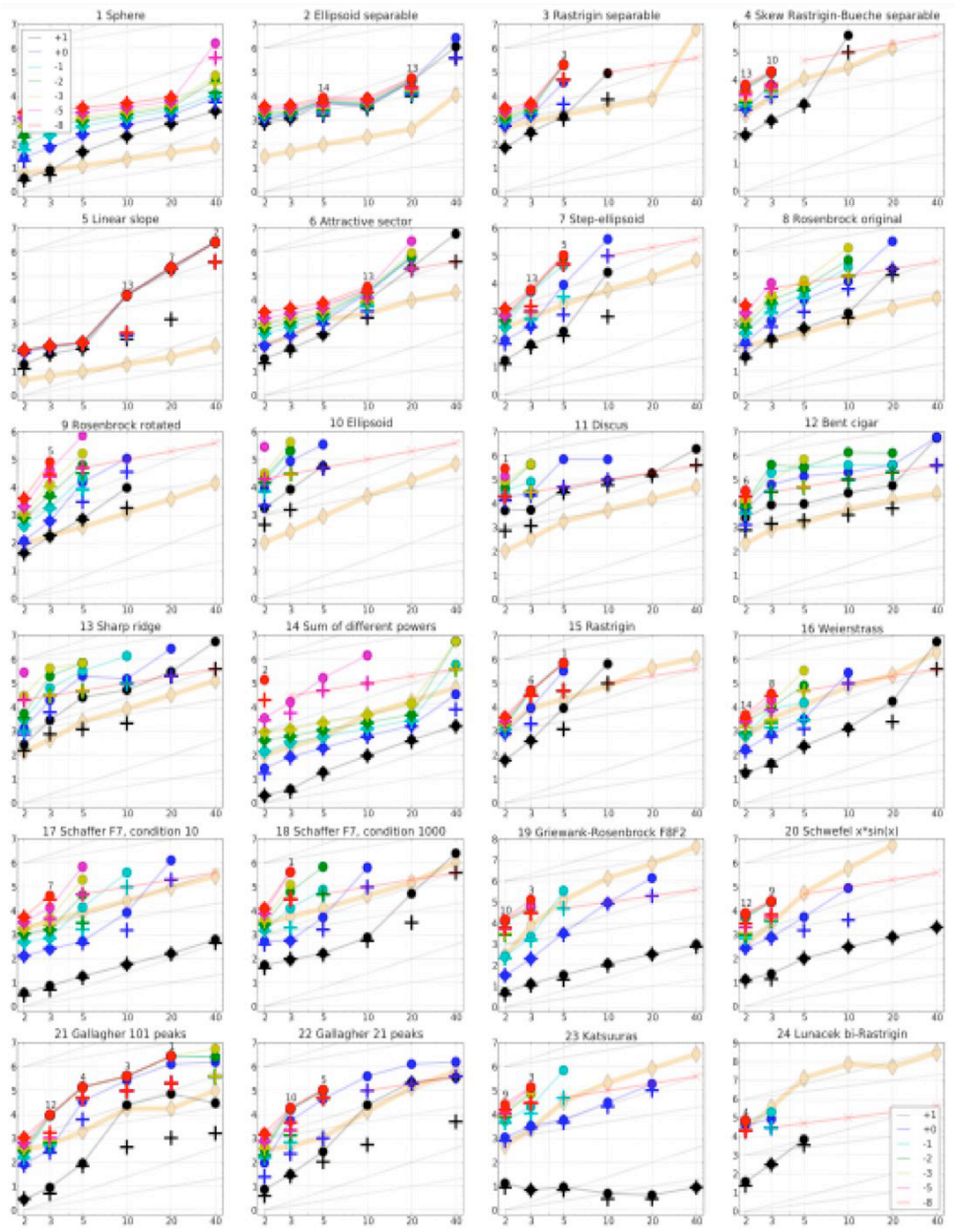


Figure 9: PSO Expected Running Time (ERT, ●) to reach $f_{\mathrm{opt}} + \Delta f$ and median number of f-evaluations from successful trials (+), for $\Delta f = 10^{\{+1,0,-1,-2,-3,-5,-8\}}$ (the exponent is given in the legend of f_1 and f_{24}) versus dimension in log-log presentation. For each function and dimension, $\mathrm{ERT}(\Delta f)$ equals to $\#\mathrm{FEs}(\Delta f)$ divided by the number of successful trials, where a trial is successful if $f_{\mathrm{opt}} + \Delta f$ was surpassed. The $\#\mathrm{FEs}(\Delta f)$ are the total number (sum) of f-evaluations while $f_{\mathrm{opt}} + \Delta f$ was not surpassed in the trial, from all (successful and unsuccessful) trials, and $f_{\mathrm{opt}}$is the optimal function value. Crosses ($\times$) indicate the total number of f-evaluations, $\#\mathrm{FEs}(-\infty)$, divided by the number of trials. Numbers above ERT-symbols indicate the number of successful trials. Y-axis annotations are decimal logarithms. The thick light line with diamonds shows the single best results from BBOB-2009 for $\Delta f = 10^{-8}$. Additional grid lines show linear and quadratic scaling.

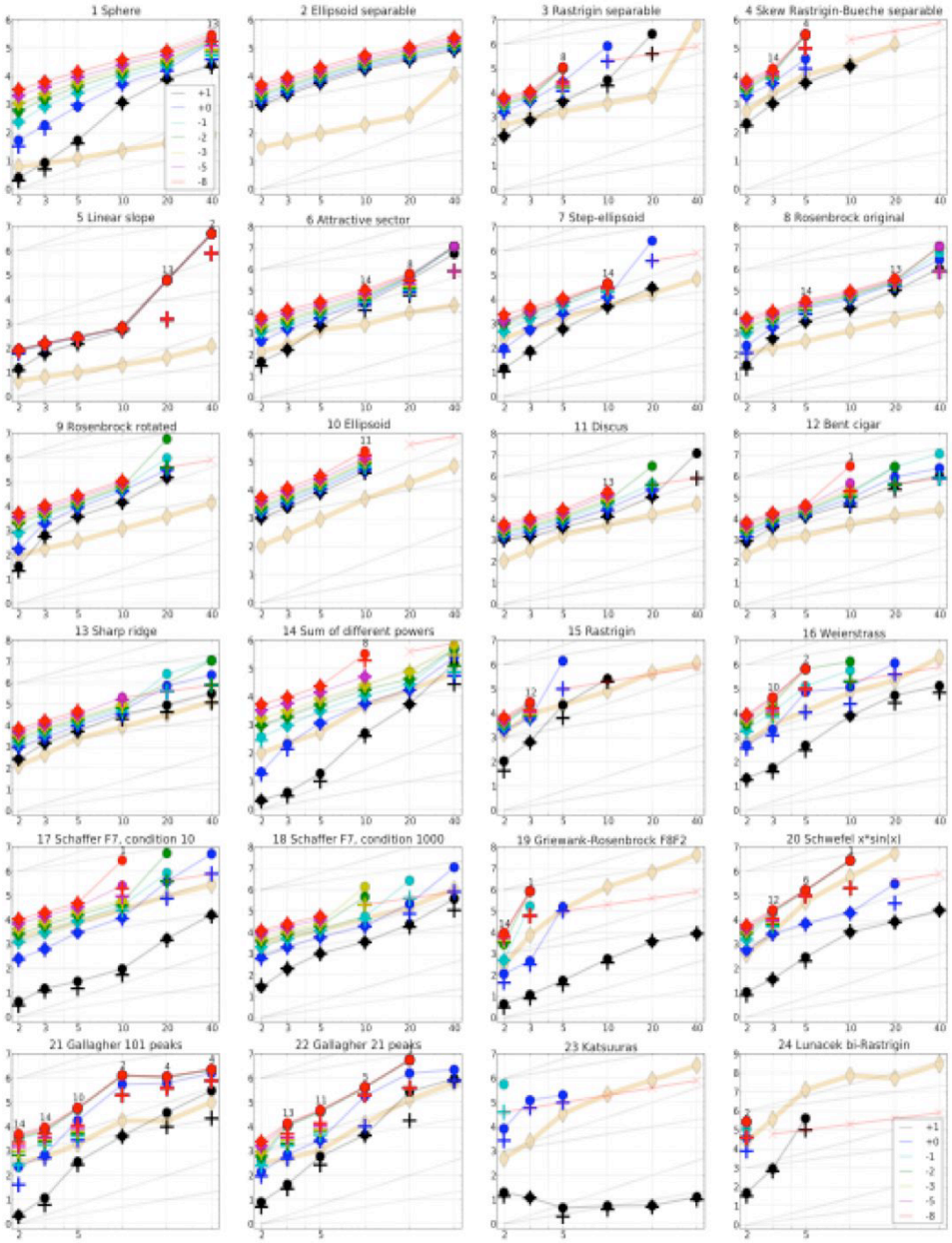


Figure 10: DE-PSO Expected Running Time (ERT, ●) to reach $f_{\mathrm{opt}} + \Delta f$ and median number of f-evaluations from successful trials (+), for $\Delta f = 10^{\{+1,0,-1,-2,-3,-5,-8\}}$ (the exponent is given in the legend of f_1 and f_{24}) versus dimension in log-log presentation. For each function and dimension, $\mathrm{ERT}(\Delta f)$ equals to $\#\mathrm{FEs}(\Delta f)$ divided by the number of successful trials, where a trial is successful if $f_{\mathrm{opt}} + \Delta f$ was surpassed. The $\#\mathrm{FEs}(\Delta f)$ are the total number (sum) of f-evaluations while $f_{\mathrm{opt}} + \Delta f$ was not surpassed in the trial, from all (successful and unsuccessful) trials, and f_{opt} is the optimal function value. Crosses ($\times$) indicate the total number of f-evaluations, $\#\mathrm{FEs}(-\infty)$, divided by the number of trials. Numbers above ERT-symbols indicate the number of successful trials. Y-axis annotations are decimal logarithms. The thick light line with diamonds shows the single best results from BBOB-2009 for $\Delta f = 10^{-8}$. Additional grid lines show linear and quadratic scaling.

c_1 is the personal learning factor and c_2 and c_3 are the social learning factors. According to Michalewicz et al [Michalewicz,1998] and [Michalewicz,1996] constrained optimization methods are classified as:

1. Methods based on preserving feasibility of solutions.
2. Methods based on penalty functions
3. Methods that make a clear distinction between feasible solutions and infeasible sequences.
4. Methods based on decoders
5. Hybrid methods

We propose to analyze the penalty methods under E.A. perspective (Evolutionary Algorithms). The penalty methods use functions (penalty functions) that degrade the quality of the nonfeasible solution. In this way the constrained problem becomes a problem without constraints by using a modified evaluation function:

$$eval(x) = \begin{cases} \text{f(x)} & x \in \mathcal{F} \\ \text{f(x) + } \textit{penalty}\text{(x)} & eoc \end{cases} \tag{13}$$

$\mathcal{F}$ is the set created by the intersection of all sets that are the restrictions of the problem (Feasible region). The penalty is zero if no violation occurs and it is positive otherwise. The penalty function is based on the distance between a nonfeasible sequence and the feasible region $\mathcal{F}$, It also works for repairing solutions outside of the feasible region $\mathcal{F}$.

There are many penalty methods. The main difference between the methods is the way the penalty function is designed and applied to the nonfeasible sequences. Some methods associate a penalty function f_j, $(j = 1, \ldots, m)$ with a constraint, which measures the violation of the restriction j as follows:

$$f_j(x) = \begin{cases} max\{0, g_j(x)\}, & si \;\; 1 \leq j \leq p \\ |h_j(x)| & si \;\; p+1 \leq j \leq m \end{cases} \tag{14}$$

Acknowledgements

The research was supported by the Spanish Research Agency projects TRA2010-15645 and TEC2010-21303-C04-02.

Bibliography

[Shi,1998] Shi, Y., Eberhart, R., A Modified Particle Swarm Optimizer, Proc. IEEE World Congr. Comput. Intell., 1998, pp. 69

[Clerck,2002] Clerc, M., Kennedy, J.,*The particle Swarm: Explosion, Stability, and Convergence in a Multidimensional Complex Space*,IEEE Trans. Evol. Comput., vol. 6, no. 1, pp. 5873, Feb. 2002.

[Bratton,2007] Bratton, D., Kennedy, J.,*Defining a Standard for Particle Swarm Optimization*.Proceedings of the 2007 IEEE Swarm Intelligence Symposium (SIS 2007).

[Parsopoulos,2002] Parsopoulos, K.E., Vrahatis, M. N.,Particle Swarm Optimization Method for Constrained Optimization Problems, 2002.

[Hatanaka,2007] Hatanaka, T., Korenaga, T., Kondo, N., Uosaki, K.*Search Performance Improvement for PSO in High Dimensional Space*,2007.

[Hu,2002] Hu, X., Eberhart, R., *Solving Constrained Nonlinear Optimization Problems wirh Particle Swarm Optimization*. Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informaticsics, SCI2002 , Vol. 5, IIIS, July 2002.

[Cagnina,2008] Cagnina, L.C., Esquivel,S. C., Coello, C.,*Solving Engineering Optimization Problems with the Simple Constrained Particle Swarm Optimizer*, 2008.

[Michalewicz,1998] Michalewicz, Z. , Fogel, D., *How to solve it. Modern Heuristics*. Springer,1998.

[Michalewicz,1996] Michalewicz, Z. , Shoenauer, M., *Evolutionary Algorithms for Constrained Parameter Optimization Problems*. Evolutionary Computation, 1996, Vol. 4, pp. 132.

[Delacour,1987] Delacour, J.; Apprentissage et Memoire: Une Approache Neurobiologique. Masson(Ed.) September. (1987).

[Mingo,1999] Mingo L.F., Arroyo F., Luengo C., Castellanos J.; *Learning HyperSurfaces with Neural Networks*. 11th Scandinavian Conference on Image Analysis. SCIA'99. June 7-11. Kangerlussuaq, Greenland. Pp: 731-737. 1999.

[Mingo,1999a] Mingo L.F., Arroyo F., Luengo C., Castellanos J.; Enhanced Neural Networks and Medical Imaging. 8th International Conference on Computer Analysis of Images and Patterns. CAIP'99. September 1-3. Ljubljana, Slovenia. 1999.

[Mingo,1999b] Mingo L.F., Giménez V., Castellanos J.; Interpolation of Boolean Functions with Enhanced Neural Networks. Second Conference on Computer Science and Information Technologies. CSIT'99. August 17-22. Yerevan, Armenia. 1999.

[Mingo,1998] Mingo L.F., Castellanos J., Giménez V.; A New Kind of Neural Networks and Its Learning Algorithm. Information Processing and Management of Uncertainty in Knowledge Based Systems. IPMU'98. Paris, France. July 6-10. Pp: 1913-1914. 1998.

[Blum,1991] Blum, E. K. & Leong, L.: Approximation Theory and Feedforward Networks. Neural Networks, 4. Pp. 511-515. 1991.

Authors' Information

Luis Fernando de Mingo López *- Dept. Organización y Estructura de la Información, Escuela Univesitaria de Informática, Universidad Politécnica de Madrid, Crta. de Valencia km. 7, 28031 Madrid, Spain; e-mail: lfmingo@eui.upm.es*
Major Fields of Scientific Research: Artificial Intelligence, Social Intelligence

Nuria Gómez Blas *- Dept. Organización y Estructura de la Información, Escuela Univesitaria de Informática, Universidad Politécnica de Madrid, Crta. de Valencia km. 7, 28031 Madrid, Spain; e-mail: ngomez@eui.upm.es*
Major Fields of Scientific Research: Bio-inspired Algorithms, Natural Computing

Miguel A. Muriel *- Dept. Tecnología Fotónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, Avenida Complutense 30, Ciudad Universitaria, 28040 Madrid, Spain; e-mail: m.muriel@upm.es*
Major Fields of Scientific Research: Theoretical Computer Science, Microwave Photonics

Daniel Triviño García *- Natural Computing Group, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s.n., 28660, Boadilla del Monte, Madrid, Spain; e-mail: d.trivino@fi.upm.es*
Major Fields of Scientific Research: Artificial Intelligence, Computer Science

Appendix A - Linear Enhanced Neural Networks (with no hidden layers) Implementation in R

```
train <- function (iter, alpha, patrones_in, patrones_out, verbose) {
    entradas <- ncol(patrones_in)
    salidas <- length(patrones_out[1,])
    num_patrones <- nrow(patrones_in)
    num_pesos <- (entradas+1)*salidas
    matriz_pesos_auxiliar <- matrix(runif((entradas+1)*num_pesos),nrow=(entradas+1),ncol=num_pesos)
    matriz_pesos_principal <- matrix(runif((entradas+1)*salidas),nrow=(entradas+1),ncol=salidas)
    for (i in 1:iter) {
        mse <- 0.0
        for (id_patron in 1:num_patrones) {
            patron_in <- c(as.matrix(patrones_in[id_patron,]),-1)
            patron_out <- c(as.matrix(patrones_out[id_patron,]))
            salida_red_auxiliar <- patron_in %*% matriz_pesos_auxiliar
            matriz_pesos_principal <- (matrix(salida_red_auxiliar,nrow=(entradas+1),ncol=salidas))
            salida_red_principal <- patron_in %*% matriz_pesos_principal
            error <- (salida_red_principal - patron_out)
            mse <- mse + sum(error*error*0.5)
            variacion_pesos <- -alpha * (error)
            matriz_pesos_principal <- matriz_pesos_principal + t(matrix(variacion_pesos,
                    nrow=(salidas),ncol=(entradas+1))) * (matrix(patron_in,entradas+1,salidas))
            vector_salida_red_auxiliar <- c(as.matrix(matriz_pesos_principal))
            error_auxiliar <- (salida_red_auxiliar - vector_salida_red_auxiliar)
            variacion_pesos_auxiliar <- -alpha * error_auxiliar
            matriz_pesos_auxiliar <- matriz_pesos_auxiliar + t(matrix(variacion_pesos_auxiliar,
                    nrow=(num_pesos),ncol=(entradas+1))) * (matrix(patron_in,entradas+1,num_pesos))
        }
        if (((i%%10)==0)&& verbose) {
            cat("Iteration ", i, " --> MSE",(mse/num_patrones)/salidas,"\n")
        }
    }
    train <- matriz_pesos_auxiliar
}

test <- function  (matriz_pesos_auxiliar,patrones_in, patrones_out) {
    entradas <- ncol(patrones_in)
    salidas <- length(patrones_out[1,])
    num_patrones <- nrow(patrones_in)
    num_pesos <- (entradas+1)*salidas
    salida_red <- patrones_out
    for (id_patron in 1:num_patrones) {
        patron_in <- c(as.matrix(patrones_in[id_patron,]),-1)
        salida_red_auxiliar <- patron_in %*% matriz_pesos_auxiliar
        matriz_pesos_principal <- (matrix(salida_red_auxiliar,nrow=(entradas+1),ncol=salidas))
        salida_red_principal <- patron_in %*% matriz_pesos_principal
        salida_red[id_patron,] <- (salida_red_principal)
    }
    test <- salida_red
}

panel.cor <- function(x, y, ...) {
    par(usr = c(0, 1, 0, 1))
    txt <- as.character(format(cor(x, y), digits=2))
    text(0.5, 0.5, txt, cex = 1.4 * ( abs(cor(x, y))) + 2, col="grey" )
}

plot_correlation <- function(patrones_in, patrones_out, network_output) {
    pairs(data.frame(patrones_in, patrones_out, network_output),upper.panel=panel.cor,
        main="", col="darkblue", cex=0.5)
}

plot2d_interval <- function(matriz_pesos_auxiliar, patrones_out) {
    patrones_in <- data.frame(runif(5000), runif(5000))
    sal <- test(matriz_pesos_auxiliar, patrones_in, patrones_out)
    color <- (sal[,1]-min(sal[,1]))/max(sal[,1]-min(sal[,1]))
    plot(patrones_in,col=rgb(color,color,0),xlab="",ylab="",pch=19,cex=1)

}
```

SOLVING DIOPHANTINE EQUATIONS WITH A PARALLEL MEMBRANE COMPUTING MODEL

Alberto Arteta, Nuria Gomez, Rafael Gonzalo

***Abstract**: Membrane computing is a recent area that belongs to natural computing.. P-systems are the structures which have been defined, developed and implemented to simulate the behavior and the evolution of membrane systems which we find in nature. Diophantine equations are those equations that have integer solutions. Currently, the extended Euclidean algorithm works to find integer solutions. .This paper shows a step by step procedure that solves a Diophantine equation by processing the extended Euclidean Algorithm*

***Keywords**: Extended Euclidean Algorithm, Membrane systems .*

Introduction

Natural computing is a new field within computer science which develops new computational models. These computational models can be divided into three major areas:

1. Neural networks.
2. Genetic Algorithms
3. Biomolecular computation.

Membrane computing is included in biomolecular computation. Within the field of membrane computing a new logical computational device appears: The P-system. These P-systems are able to simulate the behavior of the membranes on living cells. This behavior refers to the way membranes process information. (Absorbing nutrients, chemical reactions, dissolving, etc)

In this paper, we design a MEIA system just by explaining the process of encrypting the information that membrane systems process.

In order to do this we will take the following steps:

- Introduction to P-systems theory.
- Introduction to encryption algorithms
- Integration of the encryption with membrane systems
- Description of MEIA
- Applications of MEIA

Introduction to P-systems theory

I. A P-system is a computational model inspired by the way the living cells interact with each other through their membranes. The elements of the membranes are called objects. A region within a membrane can contain objects or other membranes. A p-system has an external membrane (also called skin membrane) and it also contains a hierarchical relation defined by the composition of the membranes. A multiset of objects is defined within a region (enclosed by a membrane). These multisets of objects show the number of objects

existing within a region. Any object 'x' will be associated to a multiplicity which tells the number of times that 'x' is repeated in a region.

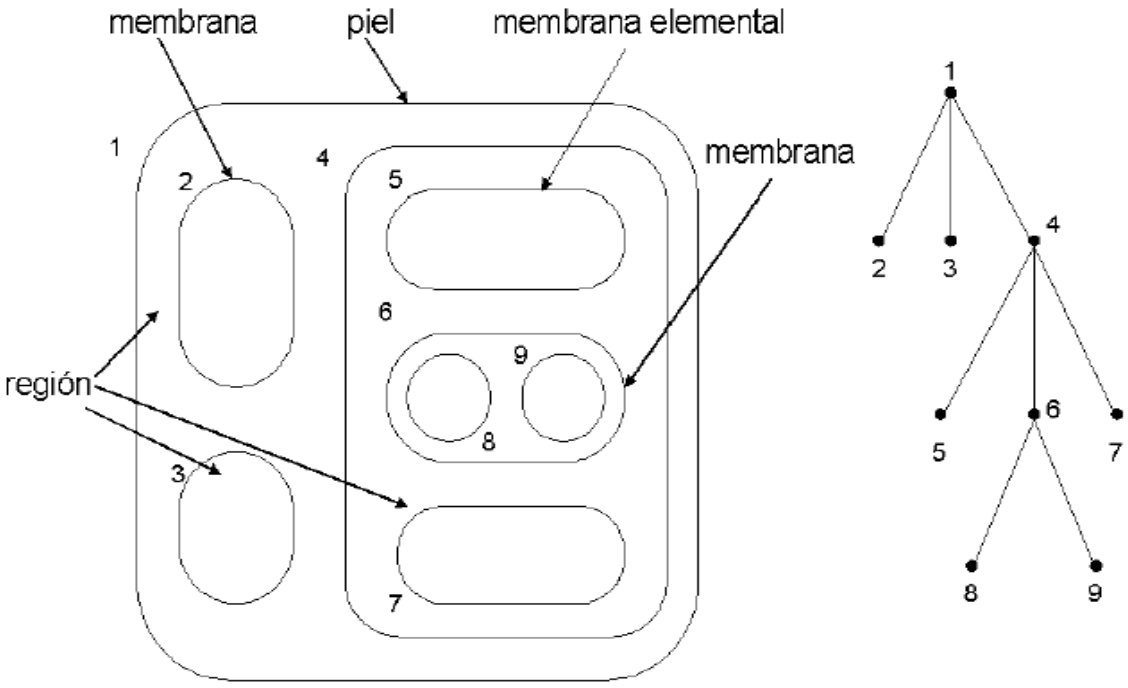


Fig. 1. The membrane's structure (left) represented in tree shape (right)

According to Păun 's definition, a transition P System of degree n, n > 1 is a construct: [Păun 1998]

$$\prod = \left(V, \mu, \omega_1, .., \omega_n, (R_1, \rho_1), .. (R_n, \rho_n), i_0\right)$$

Where:

1. *V* is an alphabet; its elements are called objects;
2. μ is a membrane structure of degree n, with the membranes and the regions labeled in a one-to-one manner with elements in a given set ; in this section we always use the labels 1,2,..,n;
3. $\omega_i \; 1 \le i \le n$, are strings from V^* representing multisets over V associated with the regions 1,2,..,n of μ
4. $R_i \; 1 \le i \le n$, are finite set of evolution rules over V associated with the regions 1,2,..,n of μ; ρ_i is a partial order over $R_i \; 1 \le i \le n$, specifying a priority relation among rules of R_i . An evolution rule is a pair (u,v) which we will usually write in the form $u \to v$ where u is a string over V and v=v' or v=v' δ where v' is a string over $\left(V \times \{here, out\}\right) \cup \left(V \times \{in_j \; 1 \le j \le n\}\right)$, and δ is a special symbol not in. The length of u is called the radius of the rule $u \to v$

5. i_o is a number between 1 and n which specifies the output membrane of Π

Let *U* be a finite and not an empty set of objects and N the set of natural numbers. A *multiset of objects* is defined as a mapping:

$$M : V \to \mathrm{N}$$
$$a_i \to u_1$$

Where a_i is an object and u_i its multiplicity.

As it is well known, there are several representations for multisets of objects.

$$M = \{(a_1, u_1), (a_2, u_2), (a_3, u_3)...\} = a_1^{u_1} \cdot a_2^{u_2} \cdot a_n^{u_n} .....$$

Evolution rule with objects in *U* and targets in *T* is defined by $r = (m, c, \delta)$ where
$m \in M(V), c \in M(VxT)\ and\ \delta \in \{to\ dissolve, not\ to\ dissolve\}$

From now on *'c'* will be referred to as the consequent of the evolution rule *'r'*

The *set of evolution rules* with *objects* in *V* and targets in *T* is represented by R *(U, T).*

We represent a rule as:

$x \to y \quad or \quad x \to y\delta$ where x is a multiset of objects in M((*V*)xTar) where Tar ={here, in, out} and y is the consequent of the rule. When δ is equal to "dissolve", then the membrane will be dissolved. This means that objects from a region will be placed within the region which contains the dissolved region. Also, the set of evolution rules included on the dissolved region will disappear.

P-systems evolve, which makes it change upon time; therefore it is a dynamic system. Every time that there is a change on the p-system we will say that the P-system is in a new transition. The step from one transition to another one will be referred to as an evolutionary step, and the set of all evolutionary steps will be named computation. Processes within the p-system will be acting in a massively parallel and non-deterministic manner. (Similar to the way the living cells process and combine information).

We will say that the computation has been successful if:

1. The halt status is reached.
2. No more evolution rules can be applied.
3. Skin membrane still exists after the computation finishes.

Extended Euclidean Algorithm

The algorithm Is used to find single solutions of Diophantine equations and great common divisor, given 2 numbers. This is the algorithm:

r_0 : = *a*, x_0 : = 1, y_0 : = 0. *Initial values*.

r_1 : = *b*, x_1 : = 0, y_1 : = 1.

i : = 1.

IF r_i = 0 RETURN r_{i-1}.. **we are done**

IF r_i > 0, DO

Dividing r_{i-1} by r_i obtaining k_i y r_{i+1} .

$r_{i-1} = k_i r_i + r_{i+1}$ $\quad 0 \le r_{i+1} < r_i$.

$x_{i+1} := x_{i-1} - k_i x_i$ e

$y_{i+1} := y_{i-1} - k_i y_i$.

i : = i + 1 and GOTO step 4.

If r_n is the last non-zero remain we are done.

$$\text{mcd}(a, b) = r_n = a x_n + b y_n$$

In fact $r_i = a x_i + b y_i, \ \forall\ i = 0, ..., n$.

By following this algorithm we can build this table.

$r_0 = a$	$r_1 = b$	r_2	...	r_{n-1}	r_n	$r_{n+1} = 0$
	k_1 =	k_2		k_{n-1}	k_n	
$X_0 = 1$	$x_1 = 0$	x_2		x_{n-1}	x_n	
$Y_0 = 0$	$y_1 = 1$	y_2		y_{n-1}	y_n	

Exercise: Find the gcd(282,84) and a linear combination that relates gcd(282,4) to 282 and 84..

i	**0**	**1**	**2**	**3**	**4**	**5**
r_i	282	84	30	24	**6**	0
k_i		3	2	1	4	
x_i	1	0	1	-2	***3***	
y_i	0	1	-3	7	***-10***	

Following the algorithm we obtain:

The last non-zero remain is $r_4 = 6$. Therefore gcd(282, 84) = 6.

Furthermore we obtain the following linear combination:

$$6 = 282 \cdot x_4 + 84 \cdot y_4$$

$$6 = 282\ (3)+84\ (-10)$$

The use of the extended Euclidean algorithm is useful for solving Diophantine equations.

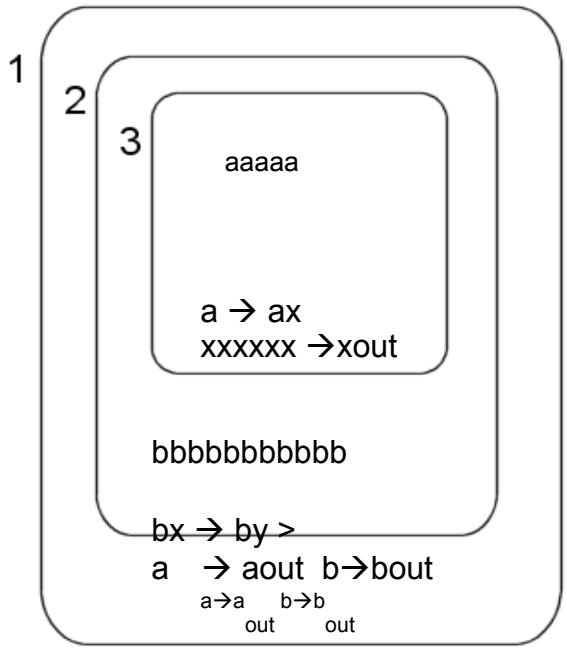
Transition P-systems solving the equation

Let us be the equation ax+by=c

The transition P-systems consist of 3 membranes:

Membrane 1 contains Membrane 2. and membrane 2 contains membrane 3.

The idea is that membrane 1 does the calculation



i

The 2 numbers involved are the number of a's and b's. The number of a's are decreasing ans passed by to membrane 2, which will do the intermediate operation to obtain y, So in the end the number of a's and b's out

The membrane 3 synchronized itself with membrane 2 . The idea is that membrane 3 is applying the euclidean algorithm step by step rising the number of 'x' until obtained the halt condition. And returns the number of a's and' b's solution (x,y) of the initial equation.

In the end:

x=aaaaaaa..aaa

y=bbbbbbb..bbbb

When this operation occurs is important to point out that it occurs in a parallel and non deterministic manner, which implies that the performance of this method is optimal in comparison with the other processing models such as the Turing machine.

Conclusions

The idea of implementing this biological model is taking advantages of the parallelism to do these operations. This theoretical model proposed here shows that it'd be possible to obtain solutions to Diophantine equations or, maybe calculate the great common divisor in a minimum number of operations. With these kind of models and optimal performance is guaranteed.

Bibliography

[Păun 1998] "Computing with Membranes", Journal of Computer and System Sciences, 61(2000), and Turku Center of Computer Science-TUCS Report nº 208, 1998.

[A. Arteta, 2008] I "Algorithm for Application of Evolution Rules based on linear diofantic equations" Synasc 2008, Timisoara Romania September 2008[1] A. Syropoulos, E.G. Mamatas, P.C. Allilones, K.T. Sotiriades "A

[Arroyo, 2001] "Structures and Bio-language to Simulate Transition P Systems on Digital Computers," Multiset Processing (. [Arroyo, 2003] "A Software Simulation of Transition P Systems in Haskell, Membrane Computing,"

[Ciobanu, Pérez-Jiménez, Ciobanu, Păun 2006] M. Pérez-Jiménez, G. Ciobanu, Gh. Păun. Applications of Membrane Computing, Springer Verlag. Natural Computing Series, Berlin, October, 2006.

[Fernández, Castellanos, Arroyo, tejedor, Garcia 2006] L. Fernández, J.Castellanos, F. Arroyo, J. tejedor, I.Garcia. New algorithm for application of evolution rules. Proceedings of the 2006 International Conferenceon Bioinformatics and Computational Biology, BIOCOMP'06, Las Vegas, Nevada, USA, 2006.

[Fernández, Martínez, Arroyo, Mingo 2005] L. Fernández, V.J. Martínez, F. Arroyo, L.F. Mingo. A Hardware Circuit for Selecting Active Rules in Transition P Systems. Proceedings of International Workshop on Theory and Applications of P Systems.Timisoara (Romania), September, 2005.

[Pan, Martin 2005] L. Pan, C. Martin-Vi de. Solving multidimensional 0-1 knapsack problem by P systems with input and active membranesl. Journal of Parallel and Distributed Computing Volume 65 , Issue 12 (December 2005)

[Păun 2000] Gh. Păun. Computing with Membranes. Journal ofComputer and System Sciences, 61(2000), and Turku Center of Computer Science-TUCS Report nº 208, 1998.

[Păun 2005] Gh. Păun. Membrane computing. Basic ideas, results, applications. Pre-Proceedings of First International Workshop on Theory and Application of P Systems,Timisoara (Romania), pp. 1-8, September , 2005.

[Qi, Li, Fu, Shi, You 2006] Zhengwei Qi, Minglu Li, Cheng Fu, Dongyu Shi, Jinyuan You. Membrane calculus: A formal method for grid transactions. Concurrency and Computation: Practice and Experience Volume18,Issue14 , Pages1799-1809. Copyright © 2006 John Wiley & Sons, Ltd.

Authors' Information

Alberto Arteta Albert *– Associate professor U.P.M Crtra Valencia km 7, Madrid-28031, Spain; e-mail: aarteta@eui.upm.es*
Research: Membrane computing, Education on Applied Mathematics and Informatics

Nuria Gomez*– Associate professor U.P.M, Crtra Valencia km 7, Madrid-28031, Spain; e-mail: ngomez@eui.upm.es*
Research: Membrane computing, Education on Informatics
Rafael Gonzalo *- Professor, faculty of informatics. Campus de Montegancedo. Boadilla e-mail: rgonzalo@fi.upm.es*
PhD on Artificial Intelligence, Education on Mathematics and Informatics

AUTOMATED SYSTEM FOR QUANTIFYING THE LEVEL OF PREPARATION IN COLONOSCOPY

Leticia Angulo-Rodríguez, Xuexin Gao, Dobromir Filip, Christopher N. Andrews and Martin P. Mintchev

***Abstract**: Colonoscopy is the gold standard method for the diagnosis of colorectal cancer (CRC). It detects the first clinical manifestation of CRC, known as polyps.*

One night prior to a colonoscopy procedure, patients are instructed to take laxative agents in order to completely cleanse the colon. This process is called bowel preparation. Contemporary sensitivity of colonoscopy for detecting polyps of a size larger than 10 mm is 98% with the limitation in detection mainly due to poor visualization related to inadequate bowel preparation.

Unfortunately, there is not yet a metric (formally recommended by means of guidelines) for the quantification of bowel preparation. Scales used nowadays are not objective, because generally colonoscopists estimate the level of cleanliness after the conclusion of the colonoscopic test.

This limitation leads to the formalization of the present study, which focuses on the development of a novel cleansing evaluation system for bowel preparation and the assessment of its clinical efficacy. The proposed system consists of a computer-based tool that can automatically measure the quantity of stool and waste matter existing within the patient during a colonoscopy procedure. As these metrics can be obtained automatically, the proposed method can lead to future quality control in daily medical practice. Furthermore, it can be used to create best practice standards for colonoscopy training or as part of medical skill evaluation.

***Keywords**: Colonoscopy; Colon preparation; Efficacy; Quality measurement metrics; Video segmentation*

***ACM Classification Keywords**: A.0 General Literature - Conference proceedings; J.3. Life and Medical Sciences*

Introduction

A. Colonoscopy and colon cancer

Colonoscopy is a procedure that allows real-time visualization of the colon. Additionally, it enables polypectomy (removal of polyps), tissue biopsy and the usage of instruments that can be inserted within the colonoscope [1].

Colonoscopy can assist doctors to diagnose some diseases with a variety of symptoms, including colonic bleeding, abdominal pain, and weight loss [2]. Most importantly, colonoscopy is now considered the “gold standard” for the detection of colorectal cancer (CRC) [3]. Based on the fact that CRC is the second most common cancer in the world causing more than 600 000 deaths annually, colonoscopy has become one of the most commonly performed medical procedures in the United States [4].

Early evaluation by means of colonoscopy is related to the decrease of CRC incidences [5]. This was demonstrated by the National Polyp Study, the objective of which was the long term monitoring of patients with polypectomy. Published results from this study demonstrated a reduction in the incidence of CRC ranging

from 76 to 90% [6]. After this publication supported the use of colonoscopy as the major screening method for CRC, the demand for this procedure increased significantly. In 2004 alone, 22 million colonoscopies were performed, resulting in an increase of 8 million compared to the previous year [7].

B. Colon preparation for colonoscopy

Bowel preparation is a very important stage in the colonoscopy procedure, since it cleanses the colon walls from fecal matter for optimal visualization [8]. Poor bowel preparation inhibits the detection of small and large polyps. In such situations clinicians often clean unclear sections of the colon using the water source that is embedded in the colonoscopy system. Such an approach is time consuming and leads to the prolongation of intubation, withdrawal times and sedation, causing an increased risk for the patient. In addition, there is an immediate impact on the associated costs, especially when the examination has to be repeated [9]. Thus, poor preparation is one of the biggest obstacles to colonoscopy effectiveness. Figure 1 demonstrates a clean colon (left) in comparison to colonic segments with poor bowel preparation (middle and right). In the poorly prepared colon, small or even moderately sized polyps can be obscured by stool and turbid fluid.

The American Society for Gastrointestinal Endoscopy (ASGE) and the American College of Gastroenterology (ACG) have recommended that quality of bowel preparation should be registered in the final report after each individual screening [1]. Unfortunately, there is not yet a metric formally recommended by means of guidelines. This is because all three presently introduced scales for bowel preparation measurements are considered to be subjective. They depend purely on the colonoscopist's appreciation of cleanliness [10].These 3 scales use subjective terms for grading, such as "excellent," "good, "fair" and "poor" [1]. Although these scales have been a welcome tool for assessing quality of preparation, their subjective nature illustrates the need for a standard quality measurement. These scales are: the Aronchick, the Ottawa and the Boston scales [10].

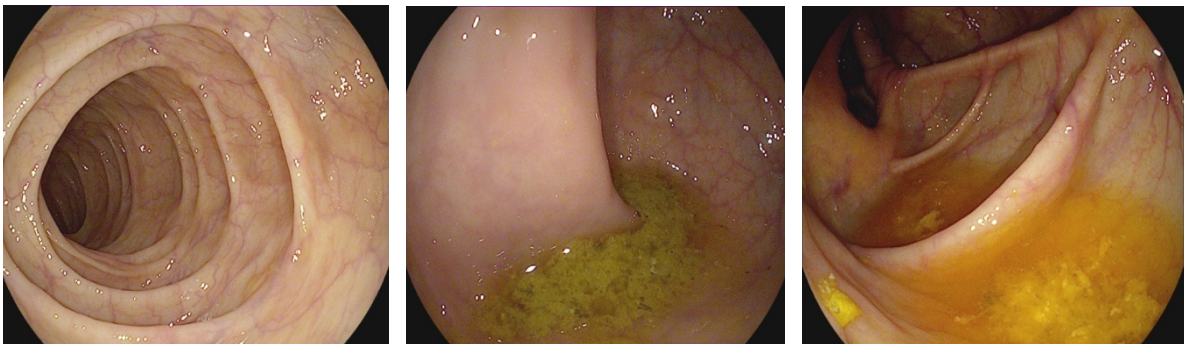


Figure 1. Left: Bowel segment with good preparation. Middle and right: Poorly prepared bowel segments.

C. Quantification scales for bowel preparation

Aronchick scale

The Aronchick Scale evaluates the quality of the colonoscopy according to the percentage of colon surface that is clearly visible [11]. The scale rates bowel preparation as "excellent" when more than 95% of the surface is seen and there is only a small amount of clear liquid. The preparation is "good" if large amounts of clear liquid cover 5% to 25% of the surface allowing the visualization of 90% of the surface. It is "fair" if there is some semisolid stool that could be washed away but still more than 90% of the surface is perceptible. Finally, the bowel preparation is "poor" when semisolid stool cannot be suctioned or washed away and it is possible to see less than 90% of the surface only [11].

Ottawa scale

The Ottawa Bowel Preparation Scale (OBPS) evaluates each segment of the colon (right colon, mid colon and rectosigmoid) regarding the presence of fecal matter. The sum of all segment scores provides a total score for

bowel the preparation [10]. This scale calculates cleanliness and fluid volume independently. Cleanliness is graded from 0 to 4 in each of the three segments of the colon. The obtained 3 numbers are added to indicate total cleanliness. On the other hand, fluid quantity is an overall value for the entire colon ranging from 0 to 2 [12]. OBPS assigns a total score of 0 to an excellent preparation where no fluid is on the colon surface. Conversely, it assigns 14 overall points (12 for poor cleanliness and 2 for large fluid quantity) to a poor preparation containing large amount of fluid in each of the three segments [11].

The criteria for cleanliness of each segment in the Ottawa scale are as follows:

♣ Excellent (0): the mucosa is completely visible, there is almost no stool, although there can be fluid, but it should be clear.

♣ Good (1): there is stool and non-transparent fluid, but colon wall is still clearly visible and it does not require washing and suctioning.

♣ Fair (2): there is turbid fluid and stool hindering the visualization of the mucosa. However, mucosal detail becomes visible by suctioning but not washing.

♣ Poor (3): there is stool on the mucosa. By suctioning and washing a fair image of the mucosa can be retrieved.

♣ Inadequate (4): It is not possible to see the colon properly even after washing and suctioning.

Weighing for the presence of fluid is: 0 for small; 1 for moderate; and finally, 2 for large amount of fluid [11].

Boston Scale

The Boston Bowel Preparation Scale (BBPS) is a 10-point scale that assesses bowel preparation with the indication that it has to be done during colonoscope withdrawal, which occurs strictly after the completion of all cleansing maneuvers [8]. In this scale, similarly to the scales discussed above, subjective terms, such as “excellent”, “good”, “fair”, “poor” and “unsatisfactory” are inferred. 4-point (from 0 to 3) scoring system is applied to each of the 3 different segments of the colon. The maximum BBPS score for a perfectly clean colon without any residual liquid is 9, and the minimum BBPS score for an unprepared colon is 0. Conversely, if the procedure has to be aborted due to insufficient preparation, all the segments are assigned a score of 0 [8].

D. Aim of the paper

Despite the fact that colonoscopy is widely practiced nowadays, an established standard for measuring the quality of colonoscopies is still missing. Such system could allow for the continuous improvement of colonoscopy practices. The objective of this study is to propose a novel colon cleansing evaluation system for bowel preparation and to assess its clinical efficacy.

Methods

The core factor to be detected in the process of developing a colon cleansing evaluation system is the level of cleanliness in the colon. In the present study the evaluation of the level of cleanliness was based on the Ottawa scale. The developed algorithm to perform stool detection utilizes color recognition, which is a major approach in feature segmentation [13].

Matlab software (The MathWorks Inc., Natick, Massachusetts) was selected for the implementation of the algorithm because of its friendly interface and the advantage of having an image processing toolbox, in which image recognition by color and shape is facilitated. This algorithm was tested in a set of 13 videos, which were acquired in the McPhail Colon Cancer Screening Center, Foothills Hospital, Calgary, Alberta, Canada. The colonoscopy videos were anonymized videos without any patient and endoscopist information.

Finally, a correlation was performed between the Ottawa 5 point-scoring system and the percentage of stool in the colon retrieved by the proposed software.

A. Color detection

The elements for characterizing an image are color, shape and texture. Of the above mentioned factors, color is the most important feature to segment images [13].

Hue-Saturation-Value color model (HSV) is a method to define color according to its three basic features:

1) Hue
2) Saturation; and
3) Lightness [14].

Hue represents a specific tone of color. Saturation is the estimation of the purity of hue and is related to the intensity of the latter. When a color is completely saturated it excludes any gray from its content. Conversely, when the saturation is low, a color turns entirely into gray. The lightness component determines if the color turns lighter or darker [15].

HSV color space is shown on Figure 2. The values for the hue component range from 0° to 360°, where red is set at 0° and black at 360°. For saturation and lightness, the values range from 0 to 1 [16]. From these 3 components, hue expresses the main characteristic of a color [16]. For instance, in the present study, hue represents the highest contrast between the colon wall and the stool in comparison with the other 2 components. Once the ranges in the HSV color space have been set for a specific target object, only the pixels in the image that are within these limits are extracted.

The spectral characteristics of a camera sensor and its lighting conditions determine the level of color in an image [17]. The use of the HSV color model is suitable for the present study because it is more consistent and efficient than the Red-Green-Blue color space (RGB) while working on color detection, since the hue component remains immune to lighting behavior conditions. This means that the hue histogram, which provides the values for each one of the HSV color space components in an image, remains about the same regardless the change of the illumination level [17].

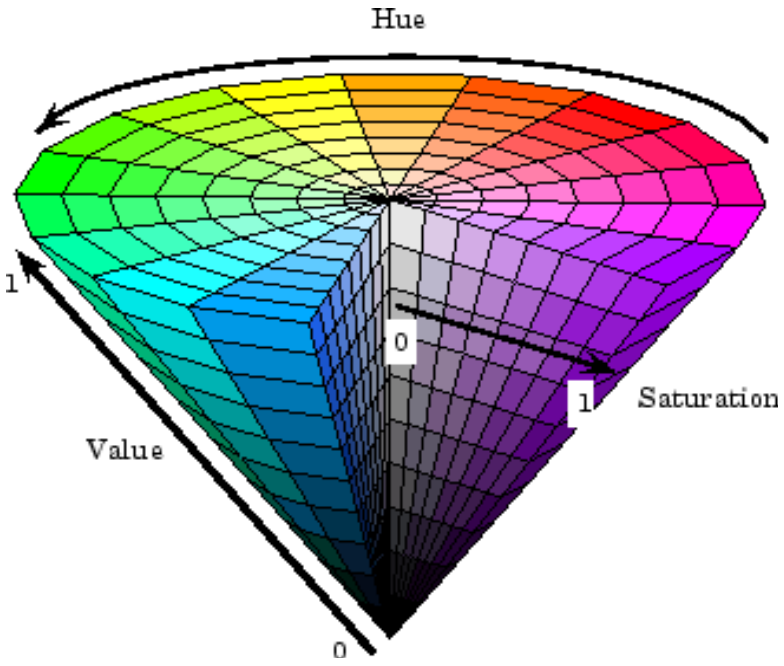


Figure 2. HSV Color Space [17].

B. Post-processing and quantification

The post-processing and quantification algorithm includes the following steps:

1. Increasing the contrast of the RGB image in order to discriminate better between the colon wall and the stool matter on it.

2. Separating the three color components of the HSV color space. As a result of that partition, hue, saturation and lightness components are separately quantified. The target color is defined by limiting the range of these component values as follows: Hmin< H <Hmax, Smin< S <Smax and Lmax< L <Lmin, where min and max are the maximum and minimum values, respectively [18]. The examined color is detected in the image if the pixel color lays within the boundaries of the HSV zone. These values should be previously identified from the HSV histograms. One histogram for each one of the components in the HSV color space was computed.

3. Every small object that has less than a certain number of pixels is removed from the image in order to avoid saturating the screen and allowing only the segmentation of areas that considerably contribute to the percentage that is displayed.

4. A morphological closing is performed on the grayscale image. The ratio of the black pixels, (which represent the pixels of stool) over the total number of pixels in the matrix is calculated. With this method the percentage output of pixels of stool is computed.

5. The perimeter of the detected area is overlaid on the original image to examine the correctness of the detection.

Figure 3 shows the output of the proposed software. At the top, the percentage of stool matter in the image is indicated. The contour around the area with poor bowel preparation is also displayed. The complete procedure is explained as a flow chart in Figure 4.

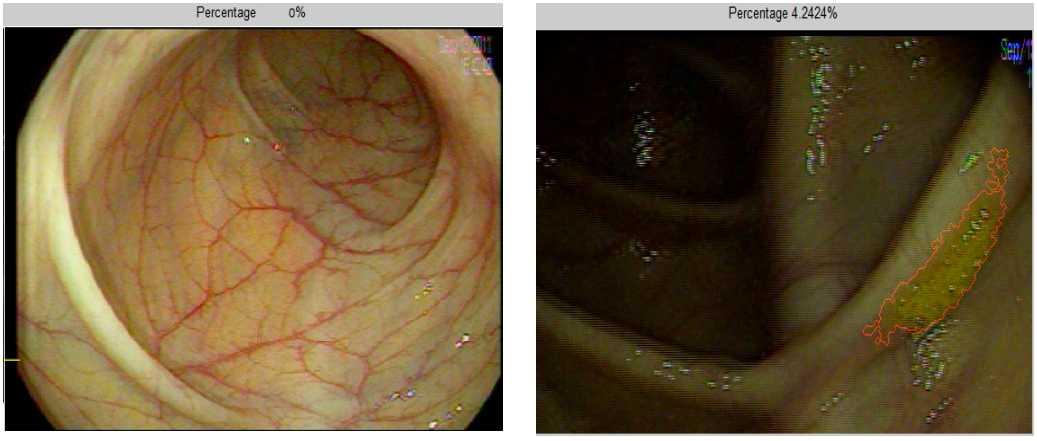


Figure 3. Output of the software. Left: A clean colon where no percentage of stool was recognized. Right: Colon with stool where the target area is highlighted.

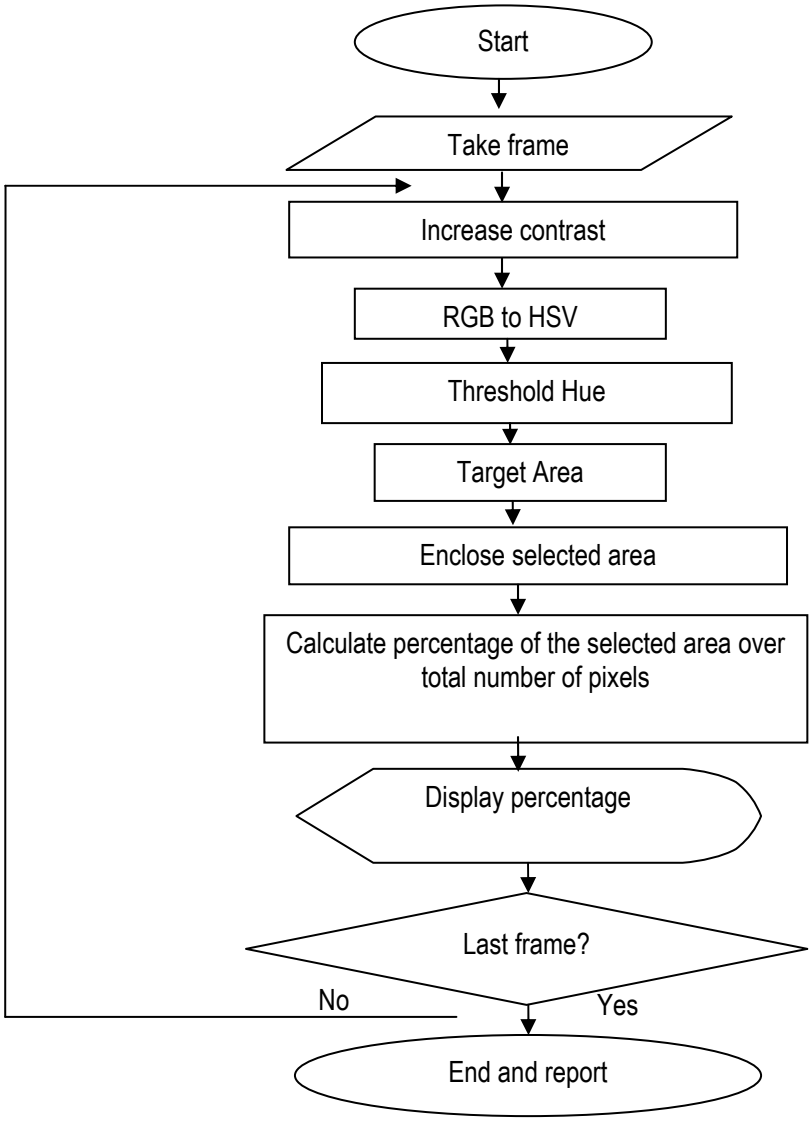


Figure 4. Algorithm for stool detection

C. Ottawa scale integration

Thirteen complete colonoscopy videos have been graded by a professionally licensed academic endoscopist, and each of the 3 segments in the colon (right colon, mid colon and rectosigmoid) was marked according to the Ottawa scale. For the preliminary analysis the value for the quantity of fluid was not taken into consideration, since the fluid recognition stage has not been yet implemented in the proposed software.

All videos were analyzed in Matlab, at 30 frames per second. All of them had different time durations, ranging from 4 to 8 minutes. The percentage of stool in each of the frames was retrieved from each of the colon segments. From the collected data, the mean of the percentages of stool and the standard deviation for each of the three colonic segments in the 13 colonoscopic videos were calculated.

Results

Table 1 lists the results utilized to perform the correlation between the Ottawa scores given for each one of the segments in the colon and the mean value of the percentages of pixels identified as stool over the total pixels in the image for each segment in the 13 videos. The values are plotted in Figure 5, which demonstrates the relationship between the Ottawa scale scores and the percentages retrieved from the proposed system. After performing Pearson correlation analysis [19], the obtained correlation coefficient r was equal to 0.61 which confirmed that there was a statistically significant correlation between the Ottawa scale score given by the endoscopist and the percentage output of the software ($p<0.01$). Alternatively, the coefficient of determination (r^2) was obtained by squaring Pearson correlation coefficient. It calculates the linear relationship strength between two variables [20]. In this study, the coefficient of determination estimated that 37% of the variances of either variable (Ottawa scale scores and percentages of stool as determined by the proposed algorithm) are shared between one another.

Table 1. Comparison between the proposed system results and the corresponding Ottawa scores. The percentages represent the per-video average of the ratios of pixels identified as stool over the total number of pixels per frame. The standard deviation of these percentages per video is also listed.

Segment	Right			Mid			Rectosigmoid		
Video #	Ottawa score	% of stool	Standard Deviation	Ottawa Score	% of stool	Standard Deviation	Ottawa score	% of stool	Standard Deviation
1	1	2.879	3.406	2	22.913	34.456	2	13.743	20.200
2	2	3.004	7.279	3	12.434	21.211	3	37.656	36.917
3	3	27.226	32.404	2	22.271	28.398	2	30.891	38.104
4	3	36.028	29.084	3	36.110	32.423	2	37.289	30.765
5	3	39.683	30.648	1	18.530	26.308	2	20.496	33.054
6	1	2.609	8.174	1	4.209	12.638	1	5.449	15.121
7	1	20.761	17.325	2	5.169	11.220	1	1.420	3.878
8	2	38.260	34.029	0	11.344	20.069	1	3.522	13.671
9	1	18.812	30.531	2	17.530	27.703	1	5.267	18.404
10	2	40.112	26.426	2	20.196	27.842	1	2.231	9.393
11	2	35.074	36.736	1	25.571	32.370	2	33.520	32.678
12	1	12.957	12.521	1	24.967	30.424	1	12.057	20.579
13	0	10.950	11.503	1	19.738	29.246	1	10.450	25.965

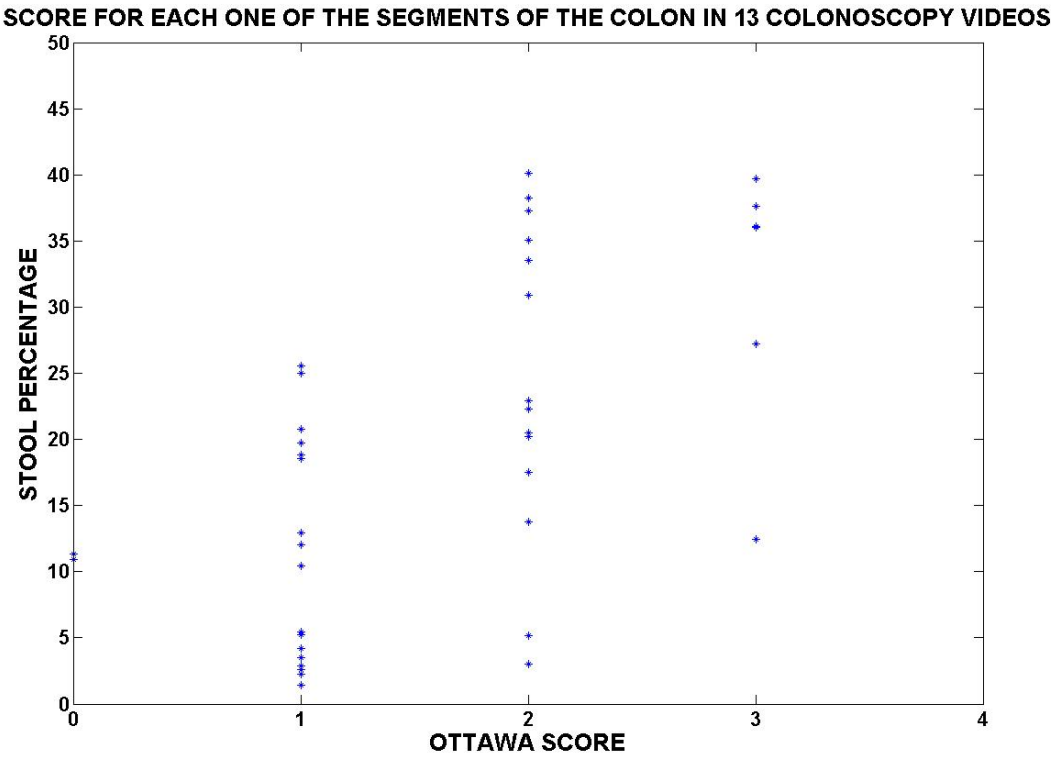


Figure 5. Graphical representation of the statistically significant correlation ($p<0.01$) between Ottawa scale scores and the mean of the stool percentages in each segment of the colon retrieved from the proposed system in 13 videos.

Discussion

The present study describes the development of a system that measures the amount of stool matter in the colon as an approach to quantify the quality of colonoscopy. The main part of the algorithm is color-based detection by hue color space.

In order to set the percentage range that would determine whether the quality of bowel preparation is “Excellent”, “Good”, “Fair”, “Poor” or “Inadequate”, the software should be improved in order to get more consistent results and have less variability in the group of points that correspond to each value in the Ottawa scale. Therefore, the reliability of the software cannot be presently evaluated, and it will be necessary to include some improvements to the code in order to get a greater determination coefficient in order to confidently conclude that the results are close to those in the Ottawa scale.

There are some potential factors that have an impact on the number of target pixels detected. Among them is the type of consumed laxative, since this might determine the color of the stool, ranging from yellow shade color to a green shade color. Furthermore, in order to agree with the Ottawa scale, the amount of fluid within the colon should be detected separately and automatically displayed by the system as a percentage output. Finally, an additional factor that affects the measurement is the quantity of blurry frames that can be identified in the video. These frames hamper the correct detection of stool matter in the colon. Future work should consist of automatically removing unclear image frames, especially when cleaning maneuvers are being conducted. So far, it has been thought that edge analysis could be an appropriate method to discard blurry frames, based on the assumption that the latter would not contain edges.

In summary, our approach for evaluating the quality of bowel preparation accounts for 3 factors:

1. Color of stool
2. Amount of fluid
3. Quantity of blurry images

Since these factors are uncontrollable, it is important to implement a calibration system that can distinguish between different color tonalities of fluid, stool matter and colon walls and can automatically adjust the HSV threshold values accordingly.

Finally, in order to have a general evaluation of colonoscopy procedures, in a future stage the proposed system has to be incorporated with other useful tools that measure additional parameters of importance in colonoscopy, such as the colonoscope withdrawal time and the polyp detection rate. Thus, the proposed software could provide complete quantitative results about colonoscopy quality and would be able to document them in a database. Overall future clinical testing of the developed system should include comparing polyp detection rate in colonoscopy utilizing colon cleanliness evaluation.

The proposed system can lead to future quality control in daily medical practice and can be used to create best practice standards, since colonoscopy quality metrics would be automatically obtained. In addition, the software could also be part of training programs for colonoscopists that would allow for an easier continuing professional development and competence maintenance, as well as for becoming a tool for medical skills evaluation. Finally, it could be pivotal for extracting information from already documented colonoscopy studies to conduct requirement analyses for future improvements of the procedure.

Conclusions

Quality is a very important issue in medical practice. Having guidelines that can help in the establishment of quality standards can significantly improve health care delivery service.

Colonoscopy has become the gold standard for the diagnosis, monitoring and therapeutic treatment of colorectal cancer, yet its continuous improvement is hampered by the lack of quantitative standards for best practices.

In this paper, the development of an objective quantification of bowel preparation stage during colonoscopic procedures is proposed, since the perception of bowel cleanliness may vary from one clinician to another. A detailed revision of the 3 scales currently available was performed for the purpose of establishing quantitative threshold levels for bowel cleanliness. Stool detection algorithm using thresholding in the hue component of HSV space and additional imaging processing procedures was proposed.

Based on the performance of the proposed evaluation system in processing 13 colonoscopic videos, correlation was established between the system and the Ottawa Bowel Preparation Scale.

Bibliography

[1] Y. Hazewinkel and E. Dekker, "Colonoscopy: basic principles and novel techniques", *Nat. Rev. Gastroenterol. Hepatol.,* vol. 8, pp. 554-564, Sep 6, 2011.

[2] National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), "**Colonoscopy**". January, 2010. Retrieved on March 29, 2012. Available: http://digestive.niddk.nhi.gov/ddiseases/pubs/colonoscopy/

[3] American Society of Colonoscopy, "*Colorectal cancer facts & figures 2011-2013*", American Cancer Society, Atlanta, GA, 2011. Available: http://www.cancer.org/Research/CancerFactsFigures/ColorectalCancerFactsFigures/colorectal-cancer-facts-figures-2011-2013-page

[4] World Health Organization, "Cancer", February, 2012. Retrieved on March 29, 2012. Available: http://www.who.int/mediacentre/factsheets/fs297/en/

[5] G. M. Eisen, "Building a better colonoscope?", *Gastrointest. Endosc.,* vol. 68, pp. 711-712, 10, 2008.

[6] R. L. Barclay, J. J. Vicari, A. S. Doughty, J. F. Johanson and R. L. Greenlaw, "Colonoscopic Withdrawal Times and Adenoma Detection during Screening Colonoscopy", *N. Engl. J. Med.,* vol. 355, pp. 2533-2541, 2006.

[7] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward and D. Forman, "Global cancer statistics", *Cancer Journal for Clinicians,* vol. 61, pp. 69-90, Mar, 2011.

[8] E. J. Lai, A. H. Calderwood, G. Doros, O. K. Fix and B. C. Jacobson, "The Boston bowel preparation scale: a valid and reliable instrument for colonoscopy-oriented research", *Gastrointest. Endosc.,* vol. 69, pp. 620-625, Mar, 2009.

[9] D. K. Rex, J. L. Petrini, T. H. Baron, A. Chak, J. Cohen, S. E. Deal, B. Hoffman, B. C. Jacobson, K. Mergener, B. T. Petersen, M. A. Safdi, D. O. Faigel, I. M. Pike and ASGE/ACG Taskforce on Quality in Endoscopy, "Quality indicators for colonoscopy", *Am. J. Gastroenterol.,* vol. 101, pp. 873-885, Apr, 2006.

[10] S. Landreneau and J. Di Palma, "Update on Preparation for Colonoscopy", *Curr.Gastroenterol. Rep.,* vol. 12, pp. 366-373, 2010.

[11] Y. S. Choi, J. P. Suh, J. K. Kim, I. T. Lee, E. G. Youk, D. S. Lee, S. K. Do and D. H. Lee, "Magnesium citrate with a single dose of sodium phosphate for colonoscopy bowel preparation", *World J. Gastroenterol.,* vol. 17, pp. 242-248, Jan 14, 2011.

[12] A. Rostom and E. Jolicoeur, "Validation of a new scale for the assessment of bowel preparation quality", *Gastrointest.Endosc.,* vol. 59, pp. 482-486, 4, 2004.

[13] M. Abdellatif, "Effect of color pre-processing on color-based object detection", in *SICE Annual Conference,* pp. 1124-1129, 2008.

[14] S. Li and G. Guo, "The application of improved HSV color space model in image processing", in *Future Computer and Communication (ICFCC),* vol.2, pp.10-13, 2010.

[15] R. T. Stevens, *Computer Graphics Dictionary.* Massachusetts, USA: Charles River Media, 2002, p.112.

[16] L. J. Ubong, C.S. Tehand G. W. Ng, "A comparison of RGB and HSI color segmentation in real - time video images: A preliminary study on road sign detection", in *Information Technology. International Symposium on,* 2008, pp. 1-6.

[17] M. Riaz, G. Kang, Y. Kim, S. Pan and J. Park, "Efficient image retrieval using adaptive segmentation of HSV color space", in *Computational Sciences and its Applications. ICCSA '08. International Conference on,* 2008, pp. 491-496.

[18] G. D. Finlayson, "Color in perspective", *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 18, pp. 1034-1038, 1996.

[19] J. D. Jobson, *Applied Multivariate Data Analysis.* Virginia, USA: Springer, 1999.

[20] C. D. McDaniel and R. H. Gates, Marketing Research Essentials. Ohio, USA: South Western College Company, 1998.

Authors Information

L. Angulo-Rodríguez *is with the Department of Biomedical Engineering, University of Calgary, Calgary, AB, T2N 1N4.*

X. Gao *is with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, ABT2N 1N4, Canada.*

D. Filip *is with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada.*

C. N. Andrews *is with the Department of Medicine, University of Calgary, Calgary, AB, T2N 1N4, Canada.*

M. P. Mintchev *is with the Department of Electrical and Computer Engineering, and the Department of Medicine, University of Calgary, 2500 University Dr. NW, Calgary, AB, T2N 1N4, Canada (e-mail: mintchev@enel.ucalgary.ca).*

AN IN-DEPTH ANALYSIS AND IMAGE QUALITY ASSESSMENT OF AN EXPONENT-BASED TONE MAPPING ALGORITHM

Chika Ofili, Stanislav Glozman, Orly Yadid-Pecht

Abstract: *In order to view wide contrast details in an image scene, a wide dynamic range (WDR) image sensor is required. However, these wide dynamic range images cannot be accurately viewed on a regular display device due to its limited dynamic range. Without the proper use of a WDR image compression algorithm, the details of images will be lost. Tone-mapping algorithms are used to adapt the captured wide dynamic range scenes to the low dynamic range displays available. This paper explores the utilization of an exponent-tone mapping algorithm for colored and monochrome WDR images in lure of a regular display. The exponent-based tone mapping algorithm utilizes only the Bayer (CFA) of the WDR image to produce tone mapped image results. High quality results are achieved without the use of additional image processing techniques such as histogram clipping. The image results are then compared with other conventional tone mapping operators available.*

Keywords: *Tone mapping, Wide dynamic range, High Dynamic Range Image, Image enhancement.*

ACM Classification Keywords: *A.0 General Literature - Conference proceedings; I.4.0 Image processing and Computer Vision- General (or .3 enhancement)*

Introduction

A. Wide dynamic range Imaging:

In imaging, dynamic range can be described as the luminance ratio between brightest and darkest parts of a scene [1]. Natural sceneries have a wide dynamic range that is of five-six orders of magnitude, while commonly used display devices have a limited dynamic range. The dynamic range of most available display devices is of two orders of magnitude, which represent 2^8 (256) levels of radiance [2].

Emerging image capture devices can produce wide dynamic range images which have higher dynamic range in comparison to available limited range display devices [3]. However, when these captured wide dynamic range images are being displayed on these commonly used devices, they appear to be over-exposed in well-lit scenes or under-exposed in dark scenes (Figure 1). Hence, image details will be lost when displayed.

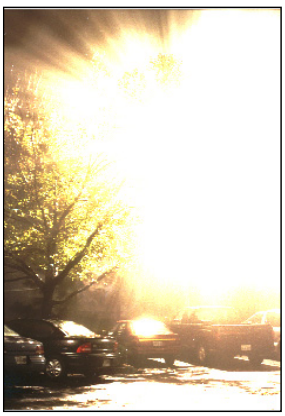


Figure 1: Under-exposed Image (left), Overexposed Image (right)

In order for the images to be accurately represented, tone-mapping algorithm is used to adapt the captured wide dynamic range scenes to the low dynamic range displays available (Figure 2).

Figure 2: Under exposed WDR image (left) and
enhanced image using Glozman et al [4] tone mapping algorithm (right)

B. Tone mapping Algorithms

There are two main categories of tone-mapping algorithms; Tone Reproduction curves (TRC) and Tone Reproduction Operators (TRO) [2].

TRC (also known as global tone mapping operator) maps all the image pixel values to a display value without taking into consideration the spatial location of the pixel in question [2]. Hence, one input pixel value corresponds to only one output pixel value. The mapping function can be a gamma, a power function, a logarithmic or a function with adaptation to the image key [5].

Conversely, TRO (also called local mapping operator) is spatial location dependent and varying transformations are applied to each pixel depending on its surrounding [2]. Hence, one input pixel value may result in different output values.

There has been a lot of progress in the development of global and local tone mapping algorithms respectively [2] [5-11]. Most tone mapping operators perform only one type of tone mapping operation (global or local). However, Meylan et al, Kats et al, Glozman et al and Shin et al [4, 5, 7, 12] have developed tone-mapping algorithms that contain both local and global mapping operators.

There are trade-offs that exist in relation to which method of tone mapping is being use used. TRC algorithms are generally less time consuming and require less computational effort but can result in loss of local contrast due to the global compression of the dynamic range [3]. In general, TRO algorithms do not result in loss of local contrast, but they require more computational effort and may result in the addition of image artifacts such as halos [4]. Consequently, TROs are less suitable for hardware implementation in comparison to TRC based algorithms.

The tone mapping algorithm developed by Glozman et al [4] achieved the goal of compressing wide range of pixel values into a smaller range that is suitable for display devices with less heavy computational effort. The simplicity of this tone mapping operator makes it a suitable algorithm that can be used as part of a system-on-chip. In this paper, an in-depth analysis of the algorithm, as well as how various modifications applied to the algorithm affects its performance will be explored. In addition, the comparison of the algorithm with other tone mapping operators in terms of computational time and image quality will be examined.

C. Exponent-based Tone mapping Operator

The algorithm implementation follows the approach proposed by Meylan et al [5] which implements the tone-mapping algorithm directly on the color filter array (CFA) data (Figure 3). This approach is different from other

traditional processing workflows that implemented rendering operations after demosaicing the CFA image. This reduces the complexity of the tone mapping algorithm needed for a colored image because the algorithm is applied to only 1/3rd of the image.

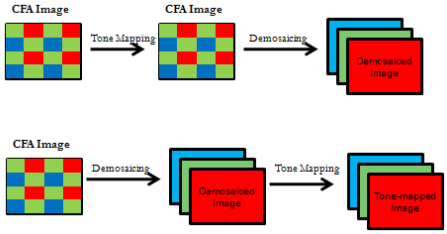


Figure 3: (Top) Approach developed by Meylan et al [5]
and traditional image processing workflow (bottom)

In the tone mapping algorithm proposed by Glozman et al [4] an inverse exponent function (1) is applied directly on the CFA.

$$Y(p) = 1 - e^{-\frac{X(p)}{X_o(p)}} \quad (1)$$

Where p is a pixel in the image, X(p) represents specific CFA pixel's input light intensity, Y(p) is the pixel's adapted signal and $X_o(p)$ is the adaptation factor of the specific CFA pixel (2). $X_o(p)$ varies for each pixel p and comprises of both a global and a local component.

$$X_o(p) = K.X_{DC} + X(P) * G_H \quad (2)$$

Where X_{DC} is the mean value of all the CFA image pixel intensities; K is the image coefficient that varies from [0 1]; "$*$" denotes the convolution operation; and G_H is a two-dimensional Low pass filter that models dependency of the $X_o(p)$ on the light intensity of the surrounding pixels.

A scaling coefficient (3) is added to equation 1 so as to ensure that all the tone mapping curves will start at the origin and meet at $X(p)=X_{max}$.

$$Y(p) = \frac{X_{max}}{\left(1 - e^{-\frac{X_{max}}{X_o(p)}}\right)} . \left(1 - e^{-\frac{X(p)}{X_o(p)}}\right) \quad (3)$$

The effects of the tone mapping parameters i.e. factor, K and the low pass filter will be discussed in the next section.

Effects of Tone mapping Parameters

K Factor

As stated above, the adaptation factor $X_o(p)$ in equation 2 consist of a global and local image processing. The global component in $X_o(p)$ comprises of factor K, and the global mean of the image X_{DC}. The amount global tonal correction done is modulated by factor K. The factor K can be adjusted between 0 and 1 depending on the image key. Low and high key images are images that have a mean intensity that is lower or higher than average [13]. K factor values closer to 0 are needed for low key images while values closer to 1 are needed for high key images. This is because the lower the K value, the higher the image contrast and overall image luminance. While

the higher the K value, the higher the compression of higher pixel values and the overall image appears less exposed (Figure 4 & 5).

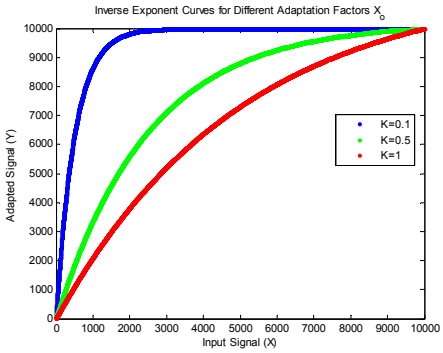


Figure 4: Inverse exponent curves for different K values

Figure 5: Memorial image (from the Devebec Library) with same low pass filter size. A. K=1, B. K=0.5 C. K=0.25 D. K=0.125

Figure 6: Dani_Cathedral image [14] with same low pass filter size. A. K=1, B. K=0.5 C. K=0.25 D.

Local Image processing: Filter

The second component in the adaptation equation (2) is the local image computation. To obtain the local component, information of the neighboring pixels is acquired. This is done by using an image filter. Various local

tone mapping operators utilize different image filters such as low pass filters, edge-preserving low pass filters, and or multi-scale pyramids to obtain the spatial localized content [14].

With tone mapping operators that involve local processing, artifacts such as halos could occur. This is because the pixels in close proximity to a specified pixel could have a very different light intensity that could result in contrast reversals [15]. Different image filters are described below and the benefits and drawbacks are examined. The aim is to find a filter that will result in less halo artifacts.

Gaussian filter:

This is a simple low pass filter that takes an average of the neighboring pixels. The average is taken in such a way that the pixels nearest to the centre pixel contribute more to the result than those further away [16]. Because it also takes into consideration the pixel values that represent edges in an image, artifacts such as halos may appear along those edges (Figure 7-8). The occurrence of contrast reversal depends on the size of the filter [11]. The kernel size in Figure 8 is smaller than the one used in Figure 7 and it has less presence of halos but also image contrast has reduced.

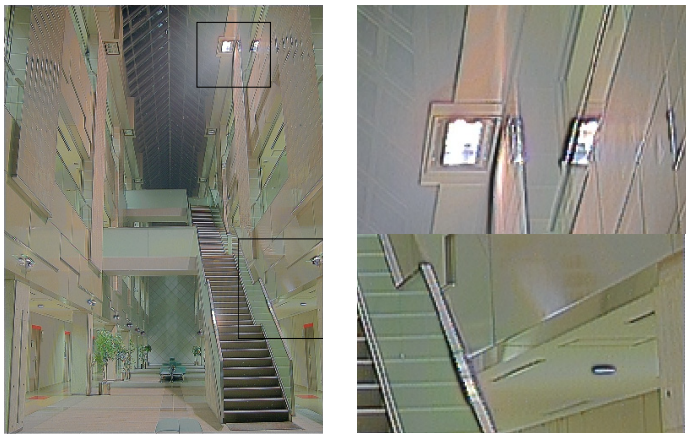


Figure 7: Tone mapped WDR image using a Gaussian filter (Kernel size=8, σ=1)

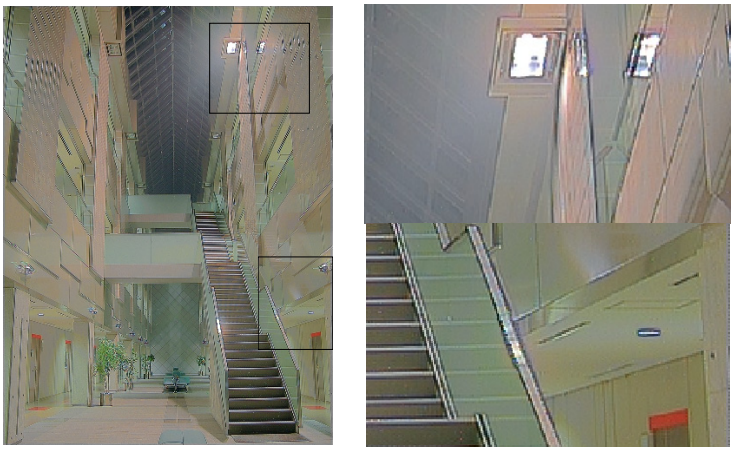


Figure 8: Tone mapped WDR image [17] using a Gaussian filter (Kernel size=3, σ=1)

Median filter:

In a median filter, the output intensity of a given pixel is the median intensity value of the pixels within a given odd sized window [18]. Based on the filter window size, it sorts the pixel values in ascending order and the median pixel value in the sorted array is stored. This makes ~~is~~ a good edge preserving filter because extremely high and low values are avoided thereby, reducing the occurrence of halos artifacts (Figure 9). Although, a key disadvantage of median filters is the loss of fine details and also presence of unexpected artifacts (Figure 10); this setback can be eliminated by reducing the size of the kernel being used (Figure 11).

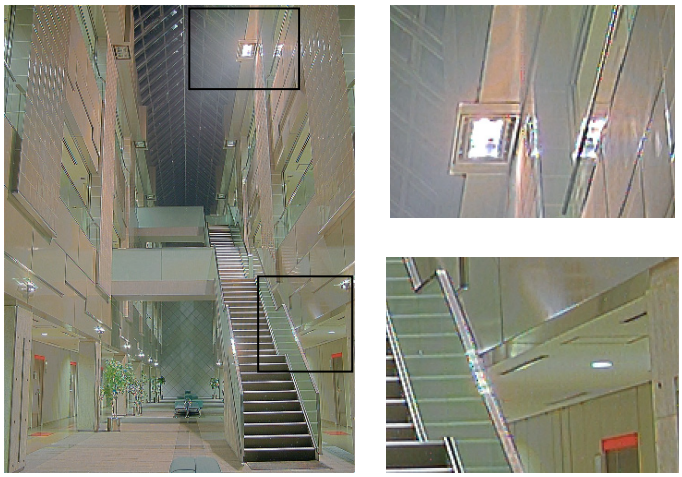


Figure 9: Tone mapped WDR image [17] using a Median filter (Kernel =7)

Figure 10: Tone mapped WDR image [19] using a Median filter (Kernel =7).

Figure 11 : Tone mapped WDR image [19] using a Median filter (Kernel =3).

Anisotropic Gaussian Filter:

This is a Gaussian-inspired filter that preserves the contrast at the edges unlike a simple Gaussian filter. It is based on the heat diffusion equation; it is modified to preserve edges by reducing the heat dissipated at the edges and at the same time, smoothen similar image [20].

It helps reduce the presence of halos and also increased the contrast as shown in Figure 12. However, it will be harder to implement on hardware in comparison to a Gaussian filter.

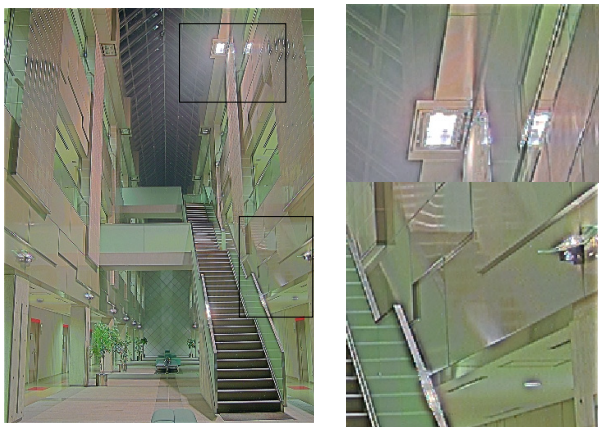


Figure 12: Tone mapped WDR Image [17] using an Anisotropic Gaussian filter.

Sigma Filter:

A sigma filter is a modified mean filter that also preserves edges. Each output pixel is an average of the surrounding input pixels that lie within an intensity range in relation to the central pixel [21]. The intensity range is described as 2*σ, where σ is an integer value. Using equation 4,

$$Y(p) = \frac{\sum_K \delta . X}{\sum_K \delta} \tag{4}$$

Where Xp is light intensity of pixels in the specified window size K, δ is used to determine which neighbouring pixels are included in the calculation of the mean. The matrix δ is obtained using equation 5

$$\delta_{i,j} = \begin{cases} 1, & \text{if } \left|X_{i,j} - X_{\text{centre pixel}}\right| \leq 2\sigma \\ 0, & \text{if } \left|X_{i,j} - X_{\text{centre pixel}}\right| > 2\sigma \end{cases} \tag{5}$$

To prevent having too few pixels involved in the calculation, the mean of the pixels within the window is used instead when the number of pixels that are within the specified intensity range is less than a specified number N [21]. Simulations done showed that decreasing the value of σ and N will reduce presence of halos.

In the results shown (Figure 13), the sigma σ, was set at 2, while the kernel size was 5. The filter preformed better in reducing the presences of halos (unlike the Gaussian filter) and did not result in unusual artefacts like the median filter. It will also be relatively easy to implement on hardware.

Figure 13: Tone mapped WDR Image [17] using a 5x5 Sigma filter (N=12 and σ=2).

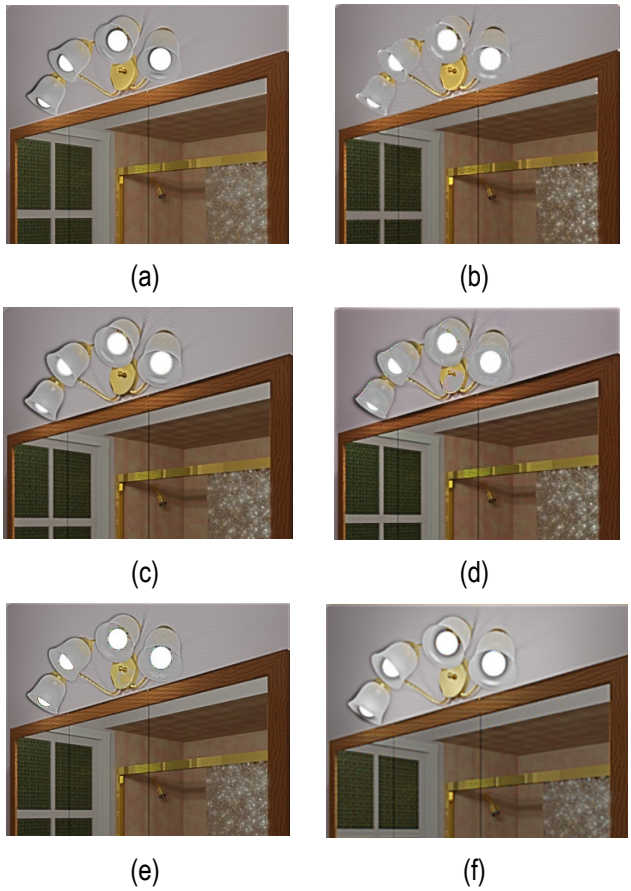
Bilateral Filter:

Bilateral filtering is a non-linear filter in which each output is a weighted average of the surrounding input pixels. Unlike the Gaussian filter, it also acts as an edge preserving filter. This is because the weight is determined from the proximity of the pixels to the pixel in question and the intensity difference between the neighboring pixels and the current pixel [22]. The weighting of the pixels deceases as the intensity difference increases [23]. Hence, this will help reduce the possibility of halo artifacts.

There is still some presence of halo artifacts but it produces a better output in comparison to the Gaussian filter (Figure 14). However, this method will not be power efficient for hardware implementation.

Figure 14: Tone mapped WDR image [17] using bilateral filter.

From simulation done on several images [11, 19], the exponent based tone mapping [4] with a sigma filter for local adaptation, was found to be the best (Figure 15-17). This is because, it had less halos, in comparison to the Gaussian filter, and it will require less computation in comparison to the anisotropic Gaussian filter and the bilateral filter. In some cases, the image produced by the anisotropic Gaussian filter had too much contrast (Figure 16). In addition, the sigma filter will correlate better with the way the eyes performs local adaptation (in comparison to the median filter). The median filter produced the best results in most cases, except in cases like Figure 10.

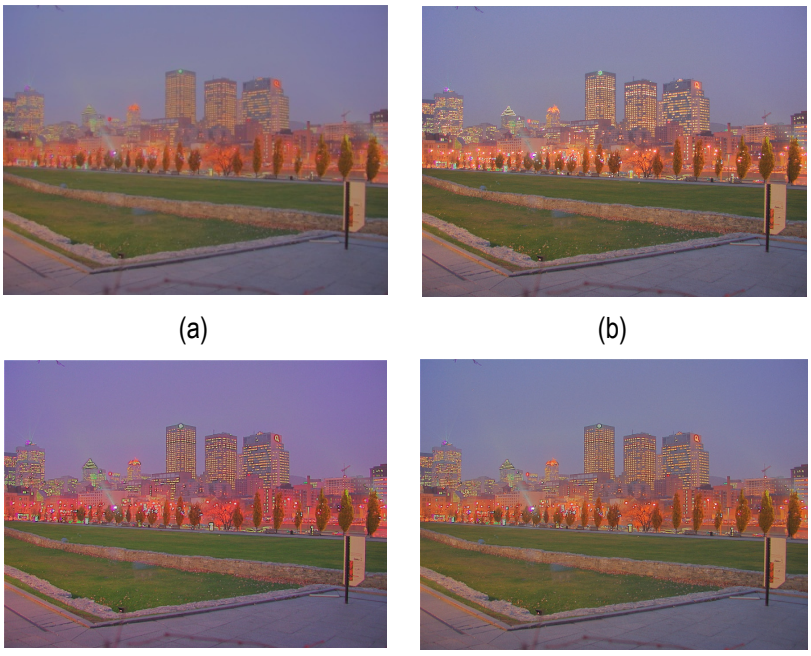


(a) (b)

(c) (d)

(e) (f)

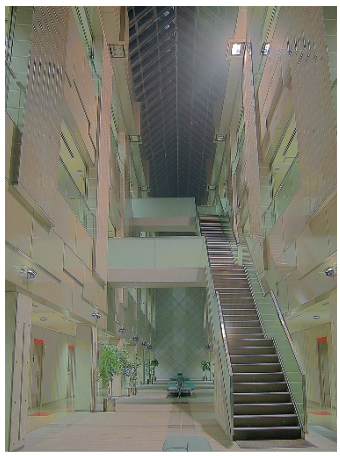
Figure 15: Tone mapped WDR image [24] using a. Gaussian filter (7x7), b. Median Filter (7x7), c. Bilateral Filter, d. Anisotropic Gaussian filter (7x7), e. Sigma Filter (7x7), f. S. Meylan method [5].



(a) (b)

(c) (d)

Figure 16: Tone mapped WDR image [24] using a. Gaussian filter (7x7), b. Median filter (7x7), c. Anisotropic Gaussian filter (7x7), d. Sigma filter (7x7).



(a)

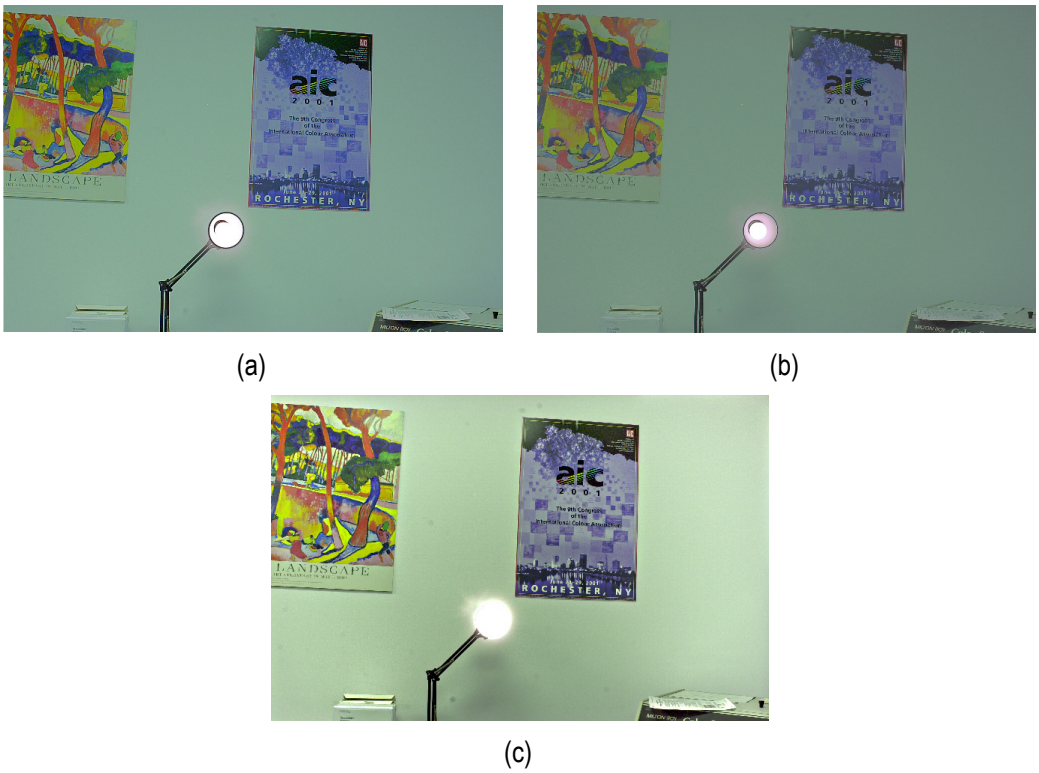
(b)

(c)

(d)

(e)

Figure 17: a. Gaussian filter (7x7), b. Median Filter (7x7), c. Bilateral Filter, d. Anisotropic Gaussian filter (7x7), e. Sigma Filter (7x7)



(a) (b)

(c)

Figure 18: Wall Image [19] using a. Meylan et al [5], b. Glozman et al [4] with Sigma filter, c. iCAM06 [8].

Results

The algorithm was implemented on MATLAB for both colored and monochrome. For colored images, demo as icing after tone mapping was performed using a simple bilinear interpolation. The colored images shown in Figure 19 show the MATLAB implementation using Gaussian filter.

Image Quality Assessment

Unlike other image processing techniques such as demosaicing, there is no ideal tone-mapped image that can be used to compare in an image quality assessment. Hence, having an objective image assessment will be difficult. In order to compensate for that, a subjective assessment was performed using 4 other available tone mapping operators [5, 6, 9, 26].

For this test, images were obtained from the Devebec library and other online sources [11, 25]. The test was aimed at checking how well the exponent-based tone mapper produced naturally good images (Figure 20). Tone mapping operators by Mantiuk et al, Fattal et al and Drago respectively [6, 9, 26], were obtained using the Windows application Luminance HDR that can be easily downloaded online [27].

The subjective study was broken into 2 separate tests that had the same image scenes except the tone mapping algorithms were different. Test 1 was done by 21 while test 2 was done by 31 people. The participants were not told which tone mapping operator was used to produce each image. For each question, the participants were asked to choose which image was the best in terms of naturalness and pleasantness. The images were randomly

arranged and the information of which tone mapping operator was used was hidden from the participants. The results are depicted in Tables 1 and 2.

After that the tone mapping operators were ranked based on the total number of votes obtained for each test. In Table 1, it can be seen that the exponent-based tone mapping algorithm [4] and Mantiuk et al's [26] algorithm had majority of the votes with the highest number of votes being for the exponent tone mapper. As can be seen in Table 2, the exponent-based algorithm by Glozman et al [4] and Meylan et al [5] were found to be amongst the best for majority of the images. It should be noted that both algorithms did not have any pre/post-processing such as histogram clipping done on the tone-mapped images produced.

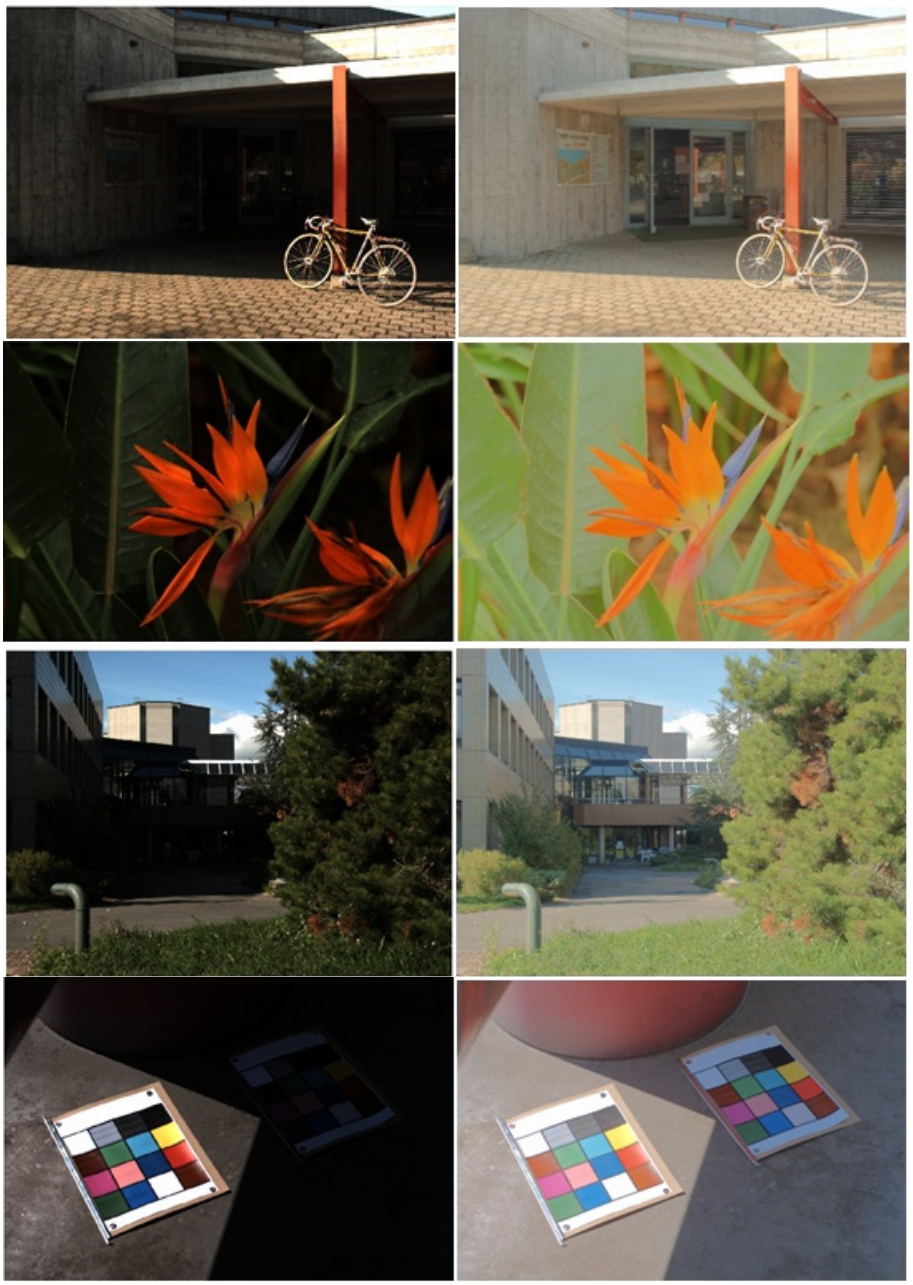


Figure 19: Tone mapped Images [25] Linear mapping (left) and Proposed automatic tuning (right).

Figure 20: Test images used in the subjective study. (Left to right): Auto, Lizard, Chapel (RAW), Synagogue

Table 1: Subjective Test results from Test 1

WDR Image	Glozman et al [4]	Drago et al [6]	Mantiuk et al [26]
Auto Image	11	2	8
Lizard	9	6	6
Chapel	5	2	14
synagogue	13	5	3
Total Votes	38	15	31

Table 2: Subjective Test Results from Test 2

WDR Image	Glozman et al [4]	Meylan et al [5]	Fattal el al [9]
Auto Image	8	15	8
Lizard	4	16	11
Chapel	18	6	7
synagogue	20	5	6
Total Votes	50	42	32

Conclusion

In this report, an in-depth analysis of an exponent-based tone mapping operator was performed. The relationships of the different components of the operators were explained. In addition, various experiments were made in order to descript the effects of various image filter needed for local adaptation processing. Median and sigma filters were found to be the best among the cases evaluated. However, the median filter resulted in loss of fine details in some tone mapped images which could be reduced by changing the kernel size. This tone mapping operator investigated will be useful for system-on-chip applications due to its simplicity.

Bibliography

[1] O. Yadid-Pecht and R. Etienne-Cummings, CMOS imagers: from phototransduction to image processing.: Springer, 2004.

[2] O. Yadid-Pecht and M. Herscovitz, "A Modified Multiscale Retinex Algorithm with an Improved Global Impression of Brightness for Wide Dynamic Range Pictures," Machine Vision and Applications, pp. 1-2, 2004.

[3] A. Belenky, A. Fish, O. Yadid-Pecht A. Spivak, "Wide Dynamic-Range CMOS Image Sensors -Comparative Performance Analysis," IEEE Trans. on Electron Devices, vol. 56, no. 11, pp. 2446-2461, November 2009.

[4] S. Glozman, T. Kats, and O. Yadid-Pecht, "Exponent Operator Based Tone Mapping Algorithm for Color Wide Dynamic Range Images," 2011.

[5] L. Meylan, D. Alleysson, and S. Susstrunk, "A Model of Retinal Local Adaptation for the Tone Mapping of Color Filter Array Images," The Journal of the Optical Society of America A, vol. 24, pp. 2807-2816, 2007.

[6] K. Myszkowski, T.Annen and N.Chiba, F.Drago, "Adaptive Logarithmic Mapping For Displaying High Contrast Scenes," EUROGRAPHICS, vol. 22, no. 2, pp. 419-426, 2003.

[7] T. Kats, S. Glozman, and O. Yadid-Pecht, "Efficient Color Filter Array luminance LOG based algorithm for Wide Dynamic Range (WDR) images compression," [Submitted].

[8] J. Kuang, Johnson G., and M. Fairchild, "iCAM06: A refined image appearance model for HDR image rendering," Journal of Visual Communication and Image Representation, vol. 18, no. 5, pp. 406-414, 2007.

[9] D. Lischinski and M. Werman R. Fattal, "Gradient domain high dynamic range compression," ACM Transactions on Graphics, vol. 21, no. 3, pp. 249-256, 2002.

[10] M. Stark, P. Shirley, and J. Ferwerda E. Reinhard, "Photographic tone reproduction for digital images," ACM Transactions on Graphics, vol. 21, no. 3, pp. 267-276, 2002.

[11] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec, High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting.: Morgan Kurfmann Publishers, 2005.

[12] H. Shin, T. Yu, Y.r Ismail, and B. Saeed, "Rendering high dynamic range images by using integrated global and local processing," Optical Engineering, vol. 50, no. 11, p. 117002.

[13] D. Tamburrino, D. Alleysson, and L., Susstrunk, S. Meylan, "Digital camera workflow for high dynamic images using a model of retinal processing," IS&T/SPIE Electronic Imaging: Digital Photography IV, vol. 6817, January 2008.

[14] K. Jiangtao, Y. Hiroshi, L. Changmeng, M. Garrett, and M. Fairchild, "Evaluating HDR Rendering Algorithms," ACM Transactions on Applied Perception, vol. 4, no. 2, 2007.

[15] L. Meylan and S. Süsstrunk, "High dynamic range image rendering with a retinex-based adaptive filter," IEEE Transactions on Image Processing, vol. 15, no. 9, pp. 2820--2830, 2006.

[16] L. O'Gorman, M. Sammon, and M. Seul, Practical algorithms for image analysis: description, examples, programs, and projects.: Cambridge University Press, 2008.

[17] K. Myszkowski F. Drago. (2011, June) Aizu University's Atrium High Dynamic Range Source Images. [Online]. http://www.mpi-inf.mpg.de/resources/atrium/atriumHdr/index.html

[18] R. Gonzalez and R. Woods, Digital Image Processing, 3rd ed.: Prentice Hall, 2008.

[19] (2011, July) RIT MCSL High Dynamic Range Image Database. [Online]. http://www.cis.rit.edu/mcsl/node/557

[20] I. Pitas, Digital image processing algorithms and applications.: Wiley-IEEE, 2000.

[21] J. Lee, "Digital image smoothing and the sigma filter," Computer Vision, Graphics, and Image Processing, vol. 24, no. 2, pp. 255-269, 1983.

[22] A. Choudhury and G. Medioni, "Perceptually motivated automatic color contrast enhancement," in Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference, 2009, pp. 1893-1900.

[23] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," ACM Transactions on Graphics, vol. 21, no. 3, pp. 257-266, 2002.

[24] G. Ward. High Dynamic Range Image Examples. [Online]. http://www.anyhere.com/gward/hdrenc/pages/originals.html

[25] L. Meylan and S. Süsstrunk, "Color image enhancement using a Retinex-based adaptive filter," in Proc. IS&T Second European Conference on Color in Graphics, Image, and Vision , 2004, pp. Vol. 2, pp. 359-363.

[26] R. Mantiuk and S. Daly and L. Kerofsky, "Display adaptive tone mapping," ACM Transactions on Graphics, vol. 27, no. 3, 2008.

[27] (2011, November) Luminance HDR. [Online]. http://qtpfsgui.sourceforge.net/

Authors' Information

C. Ofili is with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada.

S. Glozman is with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Shev, Israel.

O. Yadid-Pecht is with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada.

CROP STATE AND AREA ESTIMATION IN UKRAINE BASED ON REMOTE AND IN-SITU OBSERVATIONS

**Nataliia Kussul, Andrii Shelestov, Sergii Skakun,
Oleksii Kravchenko, Bohdan Moloshnii**

***Abstract**: This paper highlights the current state on establishing a network of test sites in Ukraine within the Joint Experiment for Crop Assessment and Monitoring (JECAM) project of the Global Earth Observation System of Systems (GEOSS). The results achieved so far on developing methods for crop state and area estimation using satellite and in situ observations are presented. The agromonitoring portal that provides access to geospatial products is described as well.*

***Keywords**: Earth remote sensing, GEOSS, JECAM, satellite data processing, agriculture, area estimation.*

***ACM Classification Keywords**: H.3.4 [Information Systems] Systems and Software - Distributed systems; I.5.1 [Computing Methodologies] Models –Neural nets; I.4.8 [Image Processing and Computer Vision] Scene Analysis - Sensor Fusion.*

Introduction

In the area of Earth observations one of the most powerful initiatives all over the world is the development of the Global Earth Observation System of Systems (GEOSS). This system is actively evolving during past ten years under the GEO Committee global activities. The main aim of the GEO Committee is availability and applicability increasing of space observations by means of coordination activities on the base of modern remote sensing possibilities for decision maker's support. Now, GEO has more than 80 country-level participants (including Ukraine), European Commission and 56 inter-governmental, international and regional organizations.

Agriculture is one of the 9 social benefit areas GEO Group. It can be mentioned that within this activities there are two global projects, Joint Experiment for Crop Assessment and Monitoring [JECAM, 2012] and, the most recent and ambitious, Global Agriculture Monitoring system (GLAM) [GLAM, 2010].

The overall goal of JECAM is to reach a convergence of approaches, develop monitoring and reporting protocols and best practices for a variety of global agricultural systems. JECAM will enable the global agricultural monitoring community to compare results based on disparate sources of data, using various methods, over a variety of global cropping systems. It is intended that the JECAM experiments should facilitate international standards for data products and reporting, eventually supporting the development of a global system of systems for agricultural crop assessment and monitoring. The JECAM initiative is developed in the framework of GEO Global Agricultural Monitoring (GEOSS Task AG0703 a) and Agricultural Risk Management (GEOSS Task AG0703 b).

To achieve the JECAM goals the initiative is bringing together the GEO Agricultural Monitoring Community of Practice to undertake an inter-comparison of monitoring and modeling methods, product accuracy assessments, and data fusion. JECAM is taking place on a finite set of regional pilot sites that are representative of a range of global agricultural systems. Data collected and shared are including time series datasets from a variety of Earth observing satellites and in-situ data with in-situ ground surveys results, in-situ soil moisture monitoring, meteorological data and other crop parameters estimation (measurement). The Community of Practice actively

works with the Committee on Earth Observing Satellites (CEOS), the space-arm of GEO, and other data providers to facilitate the acquisition of Earth Observation data and ensure a coordinated approach to space based data acquisition.

Agriculture and Agri-Food Canada (AAFC) has taken on the secretariat role of the JECAM project on behalf the GEO Agricultural Monitoring Community of Practice. One of the responsibilities as secretariat is to develop and maintain this JECAM website where government, university and non-NGO researchers can investigate and collaborate on project objectives within JECAM. Current JECAM participants are presented in Fig. 1.

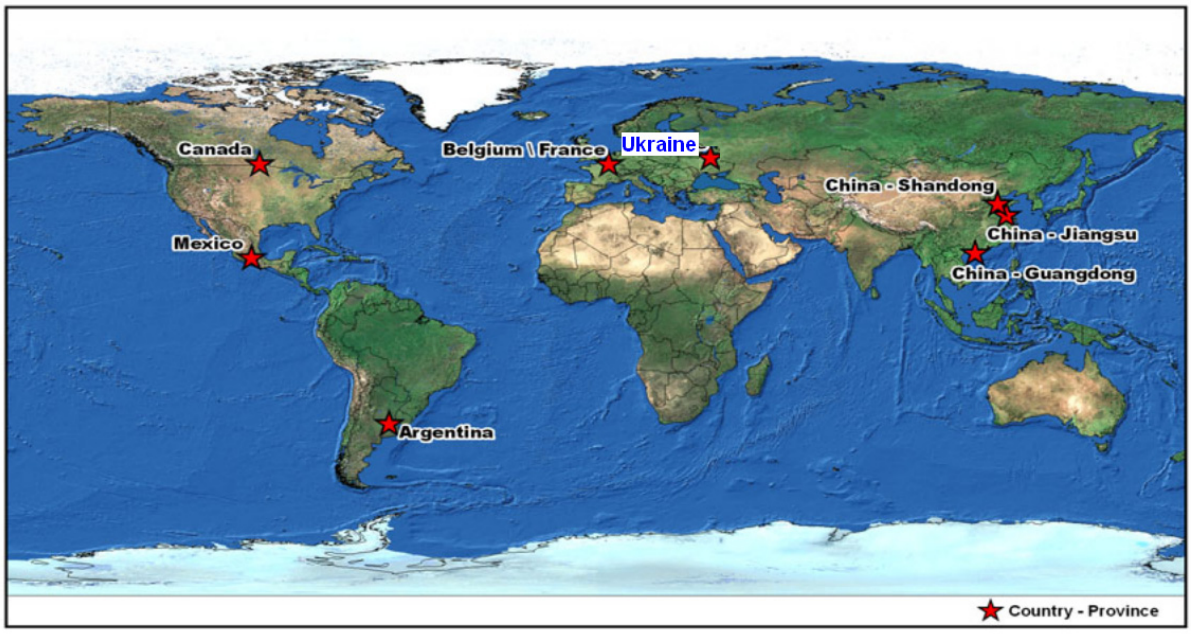


Fig. 1. International participants of the JECAM project

Last year a new global concept has been announced. (During Summit of Ministers of Agriculture G20, Paris, 22-23 June, 2011.) This initiative is related to the development of the Global Agricultural Monitoring (GLAM) for risk assessment all over the world and agrarian satellite services provision for developing countries.

Ukrainian specialists are actively involved in these international initiatives. In particular, Space Research Institute of the National Academy of Sciences and State Space Agency of Ukraine is the Ukrainian representative within the JECAM project, and holder of one of the test sites is the National University of Life and Environmental Sciences of Ukraine.

In this paper we outline the current state on establishing a network of test sites within the JECAM project and developing methods for crop state and area estimation using satellite and in situ observations [Gallego et al, 2012; Kussul et al, 2011a,b].

A Network of JECAM Test Sites in Ukraine

There are three test sites on the territory of Ukraine which are used within the JECAM project. Geographical positions, responsible organizations and scientific methodology are presented in the following sections.

As first and the most versatile test site is Kyivska oblast region (Fig. 2). Geographical location: 50°21'45.11“ of north latitude and 30°26'40.43“ of east longitude. On this area scientific investigations are provided by Space Research Institute NASU-NSAU.

The second test site has been chosen in Khmelnitsky oblast (Fig. 3). Geographical location of this agricultural region is 48°53'27.28" of north latitude и 26°50'58.50" of east longitude. The investigations on this test site were provided by Center of the Special Information Receiving and Processing, State Space Agency of Ukraine.

And, finally, third test site was countryside Pshenychne which belongs to branch of production of National University of Life and Environmental Sciences of Ukraine (Fig. 4). Geographical location: 50°7'42.99" of north latitude and 30°14'35.00" of east longitude.

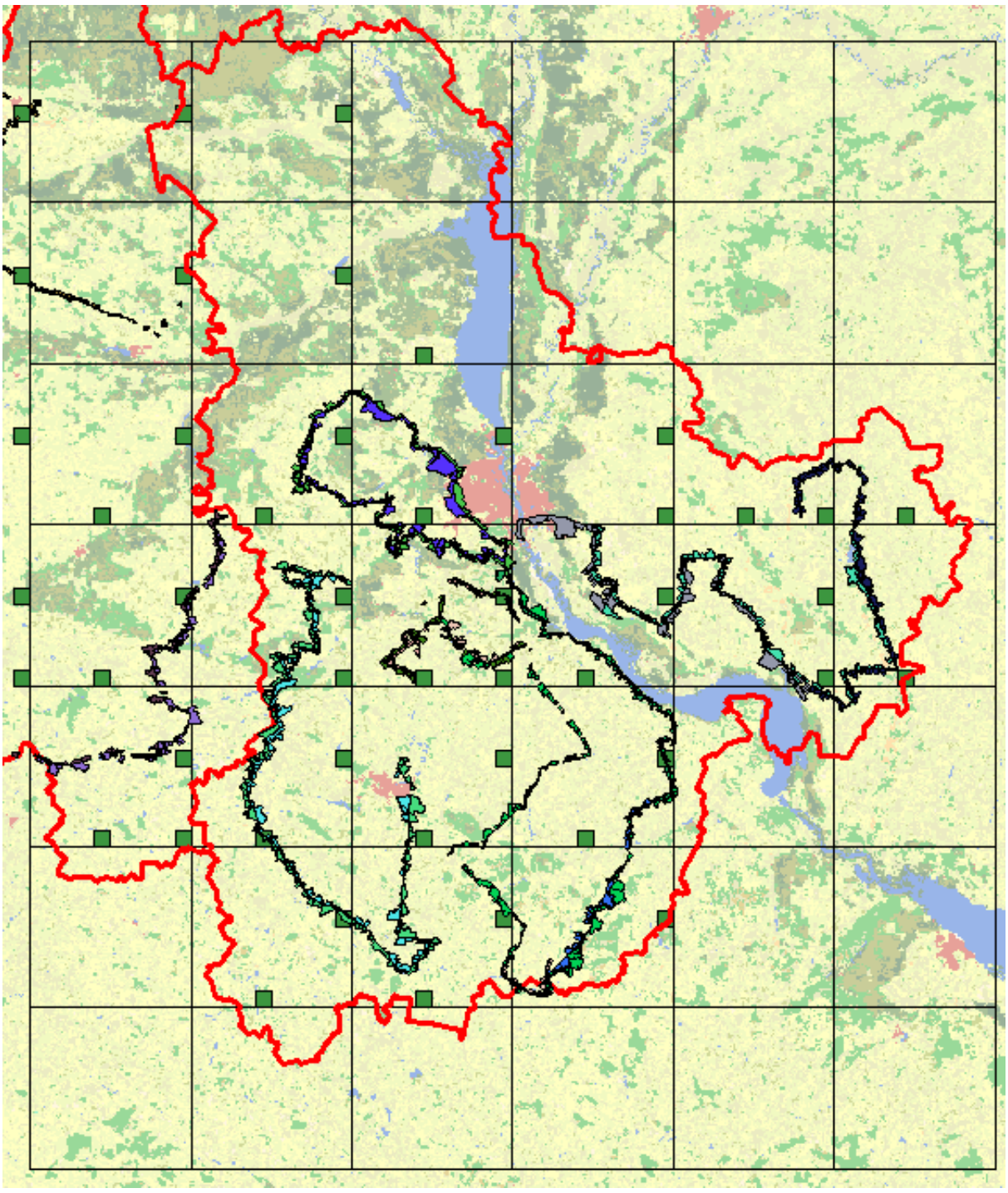


Fig. 2. Kievska oblast test site

Fig. 3. Khmelnitsky oblast test site

Fig. 4. Pshenychne countryside – third JECAM test site

Satellite Data

During last vegetation season (2011) the specialists of Space Research Institute NASU-SSAU have acquired a large amount of satellite images for target territory. This includes images that have been acquired by the following instruments.

1. **Mid-resolution Imaging Spectroradiometer (MODIS) on board Terra and Aqua satellites.** The MODIS dataset was obtained from the JRC Agri4Cast ImageServer (http://cidportal.jrc.ec.europa.eu/thematic-portals/agri4cast). The dataset contained orthorectified MODIS Normalized Difference Vegetation Index (NDVI) images. Data is stored in original Lambert Azimuthal Equal Area (LAEA) projection and cropped to the area of the selected three oblasts. Data is temporally aggregated to a decade period. The spatial resolution is 250 m. Data with heavy cloud contamination were excluded from the exercises.

2. **The Thematic Mapper (TM)** instrument onboard Landsat-5 satellite operates in 7 spectral bands with spatial resolution of 30 m and scene coverage is 185x185 km. Due to large amount of Landsat images which have been obtained during last year, this data source can be consider as main basis for agricultural monitoring.

3. **Earth Observer 1** data are provided by National Aeronautics and Space Agency of US and has 35 m spatial resolution in visible range. A couple of imageries were obtained by means of direct programming satellite via Web based informational system that is providing by NASA specialists.

4. **Ukrainian satellite Sich-2** was launched on 17th of August last year but nevertheless its data are very important for agricultural monitoring. The data have 5 spectral ranges and spatial resolution from 8 m (visible range) and 40 m (near infrared). Although Sich-2 data became available on autumn 2011 they have been used for some tasks solving.

Tasks and Scientific Methods

A number of applied agricultural monitoring tasks were solved by Space Research Institute specialists last year [Gallego et al, 2012; Kussul et al, 2011a,b; Kussul, 2011; Shelestov et al, 2011a,b]. The most important one were the crop area estimation, state vegetation estimation and winter crops area estimation, for example of winter rape beans.

Crop area estimation. For this task solving ground (in-situ) surveys were conducted when on the fields all crops were present. These included surveys along the roads and area frame sampling (AFS) surveys (segments surveys). During surveys along the roads it was collected information on crop types and geolocation data from GPS for further georeferencing of satellite images. During AFS surveys it was collected extensive information on the area being visited including land use, crop type, ground photos, type of observation, accessibility. For segment selection it has been used stratification technology from Joint Research Center of European Commission. Stratification was used to improve sampling efficiency. We followed an NASS-USDA approach where percentage of cropland area is considered as a main stratification factor. Stratification was done using ESA GLOBCOVER land cover map with resolution of 300 m. European LUCAS nomenclature was used in this investigation as a basis for land cover/land use types.

The following sampling strategy was used in the study. All the area was covered with a regular grid of sampling units of 40x40 km. Each sampling unit was further divided into segments of 4x4 km. As the field area in Ukraine is 50 to 150 ha we expected each segment to contain 15 to 20 fields in average. If some segments has more than 20 fields than area of segment was reduced up to 2x2 km.

During ground surveys surveyors were assisted with up-to-date satellite images (mostly Landsat-5, but not always). And, finally, on the base of ground measurements these satellite images were classified using different approaches such as neural networks and decision tree. Some classifications results are depicted in Fig. 5.

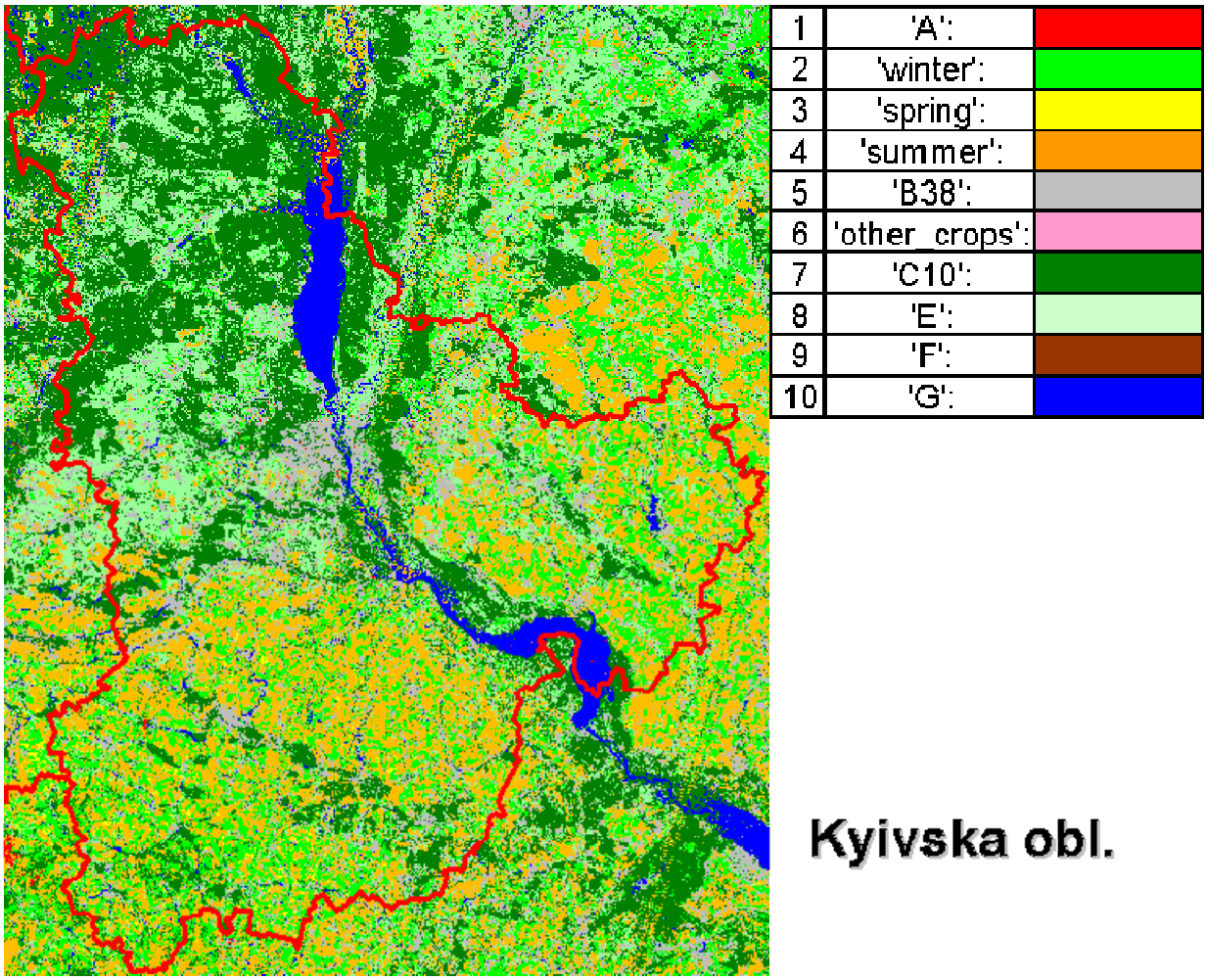


Fig. 5. Classified image for Kyivska oblast region

State vegetation estimation. During last year it was organized ground surveys in countryside Pshenychne region (another Ukrainian JECAM test site). During these in-situ observations it have been obtained a number of crop parameters such as crop height, projective cover, leaf area index, soil chemical analysis. Moreover, it were provided up-to-date satellite images. Since we have ground data as well as remote sensing data, efforts were directed to cross-validation and comparison these data. Thus, it has been estimated relationship between humus and satellite-derived biomass (Fig. 6), relationship between wheat height and biomass index (Fig. 7), relationship between crop height (ground observation) and biomass index (satellite estimation) (Fig. 8) and some other correlation dependencies.

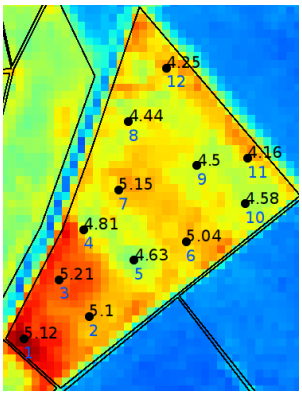


Fig. 6. Relationship between humus and satellite-derived biomass

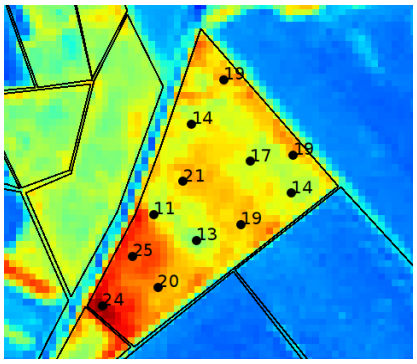


Fig. 7. Relationship between wheat height and biomass index

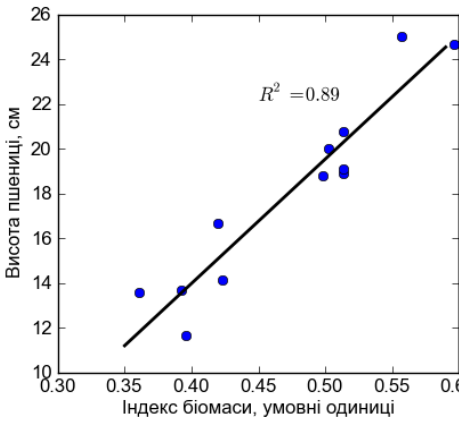


Fig. 8. Relationship between crop height and biomass index

Winter crops area estimation. European methodology of crop area estimation was used for winter crop area estimation on the base of SICH-2 data for Kyivska oblast. We used two images (1 of September and 10 of October), to distinguish winter rape and winter wheat, as well as Landsat data. These images were classified using MLP network architecture. For training set creation it were used surveys along the roads. Results are shown in Fig. 9.

Fig. 9 Classification map with winter rape fields

Information Technologies for Data Storage and Results Delivery

Since during more than 2 years we have collected a large amount of satellite images from different providers and a lot of in-situ measurements (on the fields, segments, roads etc.), including laboratory chemical soil analysis results, it was developed Web portal which provides possibilities to find, view and to do some other important GIS-operations on the multilayer basis. This set of information technologies are providing the standardized program interface for data retrieval and exchange. Portal main page is shown in Fig. 10.

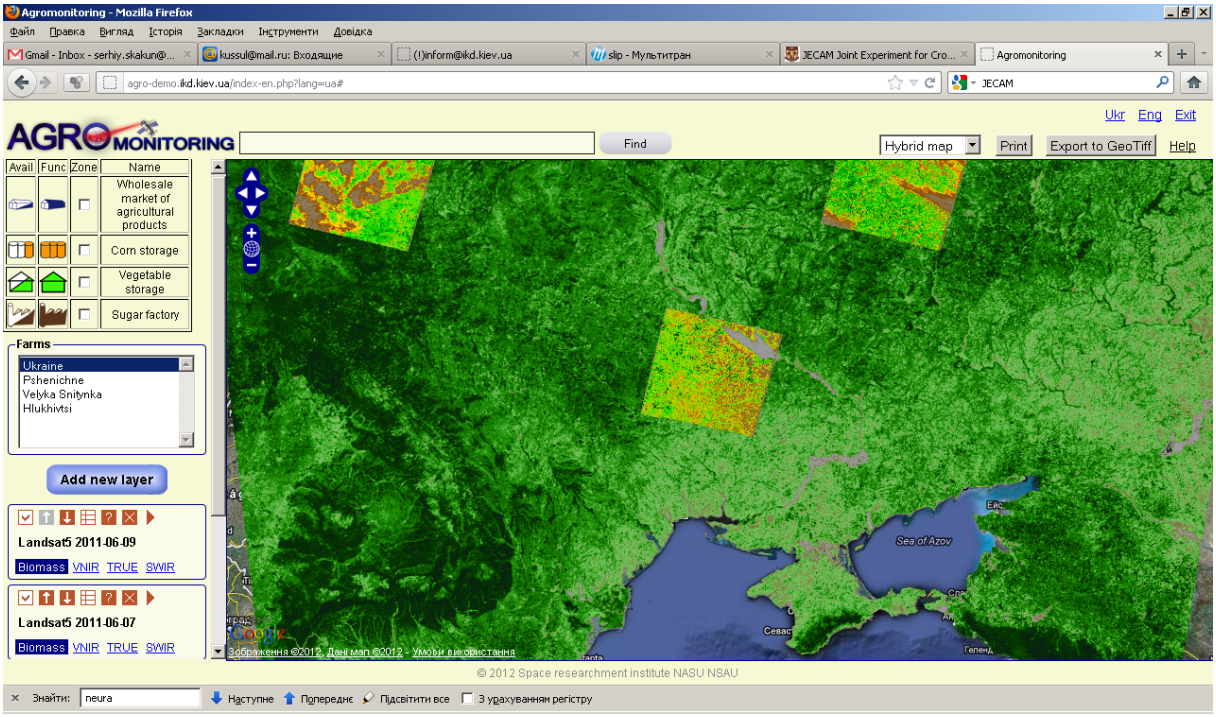


Fig. 10. Agromonitoring portal

Conclusion

In this paper we outlined the current state on establishing a network of test sites in Ukraine within the JECAM project. At present, there are 3 test sites in Kyivska and Khmelnitska oblasts. The following problems are solved using satellite and ground observation data: crop area estimation, state vegetation estimation and winter crops identification. A number of satellite images acquired by different instruments are used including MODIS, Landsat-5/TM, EO-1 and Sich-2. The satellite data and generated added value products are available at the agromonitoring portal.

Bibliography

[GLAM, 2010] Global agriculture monitoring. Global Agricultural Monitoring Community of Practice (GEO Task: AG-07-03a). (Eds) C Justice., I. Becker-Reshef, J.S. Parihar. Luxembourg: Publications Office of the European Union, 2010, doi:10.2788/82778, 32 pp.

[JECAM, 2012] Joint Experiment for Crop Assessment and Monitoring, www.umanitoba.ca/outreach/aesb-jecam.

[Gallego et al, 2012] J. Gallego, A.N. Kravchenko, N.N. Kussul, S.V. Skakun, A.Yu. Shelestov, Yu.A. Grypych. Efficiency assessment of different approaches to crop classification based on satellite and ground observations // Int. Scient. Journal “J. of Automation and Inf. Sci”. N 3. 2012. P. 123-134.

[Kussul et al, 2011a] N. Kussul, S. Skakun, O. Kravchenko. Environmental Risk Assessment Using Geospatial Data and Intelligent Methods. International Journal "Information Technologies & Knowledge" Vol.5, Number 2, 2011, pp.129-140.

[Kussul et al, 2011b] Kussul N., Shelestov A., Skakun S., Kravchenko O. Intelligent Data Processing in Global Monitoring for Environment and Security. In: High-performance Intelligent Computations for Environmental and Disaster Monitoring. ITHEA, Kiev-Sofia, 2011. P. 76-103.

[Kussul, 2011] N. Kussul. Space Information Technologies: State of Art and Prospects in Ukraine (in Russian). In: O. Fedorov (Eds.) "The Space Research Prospects of Ukraine"- 2011, Kyiv: Academperiodica, pp.148-153.

[Shelestov et al, 2011a] A. Shelestov, N. Morze, O. Kussul, Yu. Grypych. Distributed agromonitoring system (in Russian). In: Krassimir Markov, Vitalii Velychko (Eds) Applicable Information Models,-2011, ITHEA, Sofia, pp. 115-124.

[Shelestov et al, 2011b] A. Shelestov, N. Kussul, E. Zagorodny, S. Voloshyn, S. Skakun, O. Kravchenko, A. Kolotij. Geoinformational Farmer's System (in Russian). Science and Innovations. 2011, T. 7, № 3, pp. 25—29.

Authors' Information

Nataliia Kussul *– Deputy Director, Space Research Institute NASU-NSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine; e-mail: inform@ikd.kiev.ua*

Major Fields of Scientific Research: Grid technologies, design of distributed software systems, parallel computations, intelligent data processing methods, neural networks, satellite data processing, risk management and space weather.

Andrii Shelestov *– Leading Scientist at the Space Research Institute NASU-NSAU, Head of Software Development Department at the National University of Life and Environmental Sciences of Ukraine, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine; e-mail: inform@ikd.kiev.ua*

Major Fields of Scientific Research: Grid and distributed technologies, design of distributed software systems, parallel computations, intelligent data processing methods, neural networks, satellite data processing.

Sergii Skakun *– Senior Scientist, Space Research Institute NASU-NSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine; e-mail: serhiy.skakun@ikd.kiev.ua*

Major Fields of Scientific Research: Grid computing, Sensor Web, Earth observation, satellite data processing, risk analysis.

Oleksii Kravchenko *– Senior Scientist, Space Research Institute NASU-NSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine; e-mail: oleksiy.kravchenko@gmail.com*

Major Fields of Scientific Research: Earth observations, remote sensing, satellite data processing, geospatial services.

Bohdan Moloshnii *– Undergraduate student, National University of Life and Environmental Sciences of Ukraine, Heroyiv Oborony st., 15, Kyiv-03041, Ukraine; e-mail: mr.starsolo@gmail.com*

Major Fields of Scientific Research: agriculture, software engineering.

THE USE OF TIME-SERIES OF SATELLITE DATA TO FLOOD RISK MAPPING

Sergii Skakun

***Abstract**: In this paper we propose a novel approach for flood hazard mapping by processing and analyzing a time-series of satellite data and derived flood extent maps. This approach is advantageous in cases when the use of hydrological models is complicated by the lack of data, in particular high-resolution DEM. We applied this approach to the time-series of Landsat-5/7 data acquired 2000 to 2010 for the Katima Mulilo region in Namibia. We further integrated flood hazard map with dwelling units database to derive flood risk map.*

***Keywords**: flood hazard, flood risk assessment, Earth remote sensing, Earth observation, satellite data processing, UN-SPIDER.*

***ACM Classification Keywords**: H.1.1 [Models and Principles] Systems and Information Theory; I.4.8 [Image Processing and Computer Vision] Scene Analysis - Sensor Fusion.*

Introduction

Over last decades we have witnessed the upward global trend in natural disaster occurrence. Hydrological and meteorological disasters are the main contributors to this pattern [Knight, 2006; Rodriguez et al, 2009]. In 2007, hydrological disasters, such as floods and wet mass movements, represented 55% of the overall disasters reported.

It should be noted that in recent years flood management has shifted from protection against floods to managing the risks of floods (European Flood risk directive) [Mostert and Junier, 2009]. To enable flood risk assessment, corresponding flood hazard and flood risk maps should be developed. Flood risk is a function of two arguments: hazard probability and vulnerability [Mostert and Junier, 2009; Schumann and Di Baldassarre, 2010]. In other words, risk is a mathematical expectation of vulnerability (consequences) function [Jonkmana et al, 2003; Hoes and Schuurmans, 2006; Kussul et al, 2010]. Flood probability density is to be estimated in order to produce flood hazard maps. Usually, this is done through hydraulic modeling of a peak flow []. But running such models faces many uncertainties [Horritt, 2006] due to the lack of hydrological and other required data, their incompleteness and imperfection [Mostert and Junier, 2009]. The use of space-borne remote sensing data to flood risk mapping is a complement approach to the existing flood modeling techniques [Schumann and Di Baldassarre, 2010; Bates et al, 1997; Bates 2004; Horritt 2006; Lecca et al, 2011].

In [Schumann and Di Baldassarre, 2010], a novel approach for rapid flood risk mapping is proposed based on the use of radar satellite data. An event-specific weighted hazard map was generated based on plausible flood area observations from an aggregation of widely applied image-processing techniques. The map is further augmented to an event-specific fuzzy flood risk map by fusing the multialgorithm ensemble map with vulnerability-weighted land cover vector data. In [See and Abrahart, 2001], an ensemble approach to hydrological forecasting is exploited. River level is predicted by fusing outputs from different models, in particular fuzzy neural network, statistical model, and hydrological model TOPMODEL. It was shown that accuracy of the ensemble model is higher than for separate models.

In this paper we propose a novel approach for flood hazard mapping by analyzing a time-series of satellite data. In particular, satellite images are processed in order to derive flood extent maps and the latter are used to estimate flood probability density. In particular, each pixel of the flood extent maps can be one of the following values: 0 - «No water», 1 - «Water», 2 - «NoData». The specific value «NoData» is used to mark pixel that contain no valuable information due to the cloud cover and shadows, or specifics of the satellite instrument (for example, Landsat-7/ETM+ SLC-off pixels). Therefore, if «NoData» pixels are excluded from considerations, each pixel is binary with Bernoulli distribution. Maximum likelihood method is used to derive a parameter of Bernoulli distribution, a *success probability*, from sampling set. This parameter shows probability of inundation, and can be viewed as flood probability density function.

Study area and available data description

The study area is the Katima Mulilo region in Namibia (Fig. 1). Since 2009, Namibia has experienced a surge of flooding in the Northern portion of the country. It was estimated that during 2009, 700,000 of the approximately 2 million people in Namibia were impacted by the floods of 2009, furthermore around 50,000 people were displaced and 102 people lost their lives. During the 2000-2011 periods, each year, except 2005, was characterized by floods that usually occurred from the month of February through May. Three floods from this period were in top 10 water level records historically, and 8 floods were in top 20.

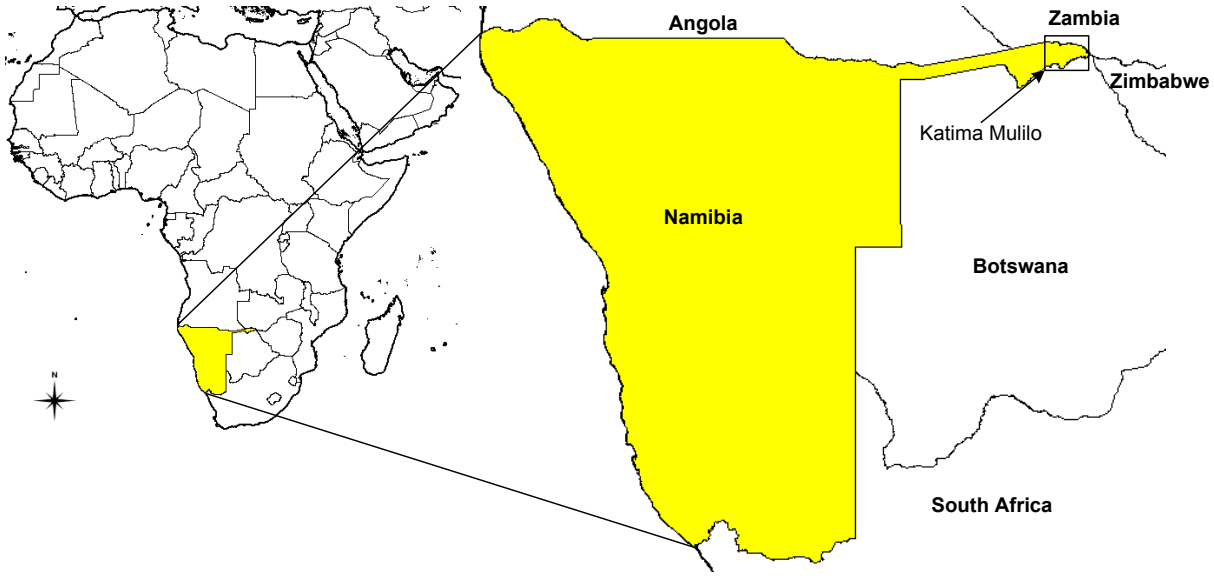


Figure1. Location of the study area

As a follow-up to the assistance provided by the United Nations Platform for Space-based Information for Disaster Management and Emergency Response (UN-SPIDER), National Aeronautics and Space Agency (NASA), German Space Agency (DLR), Ukraine Space Research Institute (USRI), and other space agencies in 2009, an international multi-disciplinary initiative, titled Namibia SensorWeb Pilot Project, was established. The Pilot aims at developing an operational trans-boundary flood management decision support system for the Southern African region to provide useful flood information and water-borne disease forecasting tools for local decision makers.

One of main tasks within the Pilot is flood risk assessment. In order to provide comprehensive hydrological modeling, a high-resolution (approximately 1 m) digital elevation model (DEM) is required since the topography is very flat in the study area. At present, such DEM is not available, and a 90 m resolution SRTM DEM is not enough to accomplish that task.

To tackle this problem we exploit different data sets in order to obtain flood hazard map. In particular, we benefit from large number of freely available images that were acquired by Landsat-5/7 satellite over 2000-2010 years.

The following are characteristics of the geospatial data that were used in the study:

-- 44 images acquired by Landsat-5/TM and Landsat-7/ETM+ from 2000 to 2010 (Fig. 2).

-- River gauge data: water level and discharge from 1965 to 2010 (provided by Hydrological Services Namibia, (Fig. 3).

-- The Tropical Rainfall Measuring Mission (TRMM) rainfall estimates and global flood potential forecast [Yilmaz et al, 2010]. Rainfall estimates are 3-, 24-, 72- and 169-hour rainfall accumulation. Flood potential forecasts are provided for 24-, 72- and 128-hour in advance. Real-time global estimation of flood areas using satellite-based rainfall and a hydrological model are run globally, every three hours at 0.25° resolution. Real-time product are produced within 6 h after observations made by TRMM.

-- Namibia dwelling unit database.

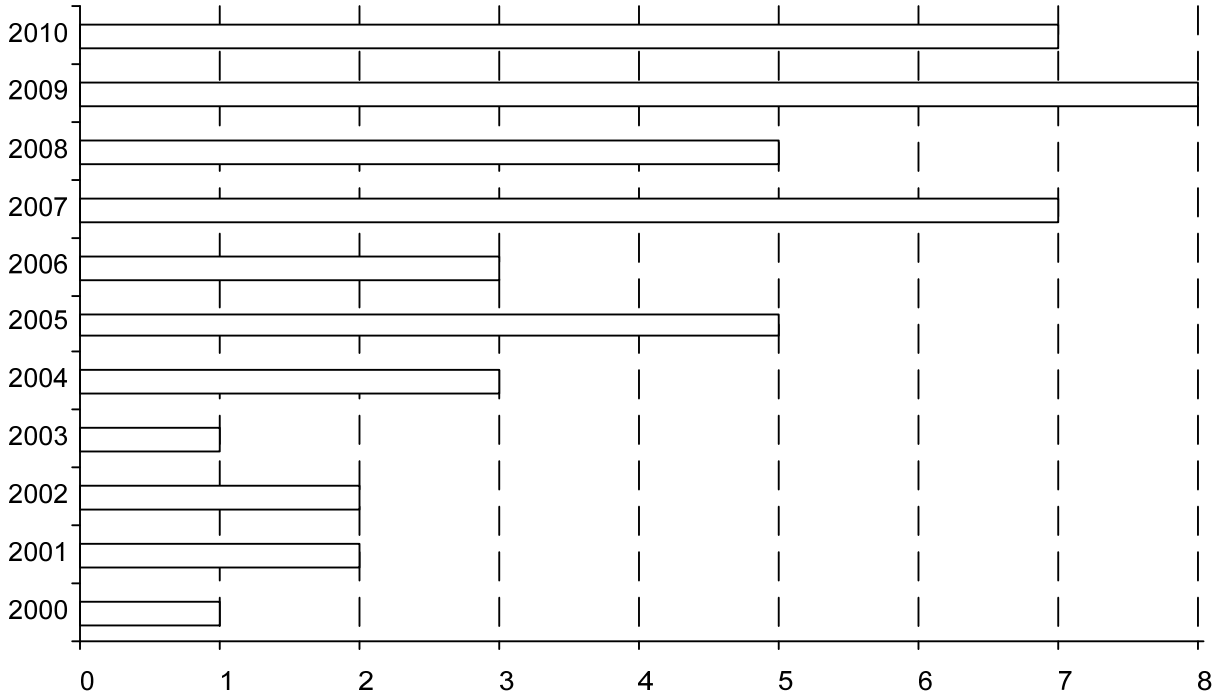


Figure 2. Distribution of the number of Landsat-5/7 images over 2000-2010 years

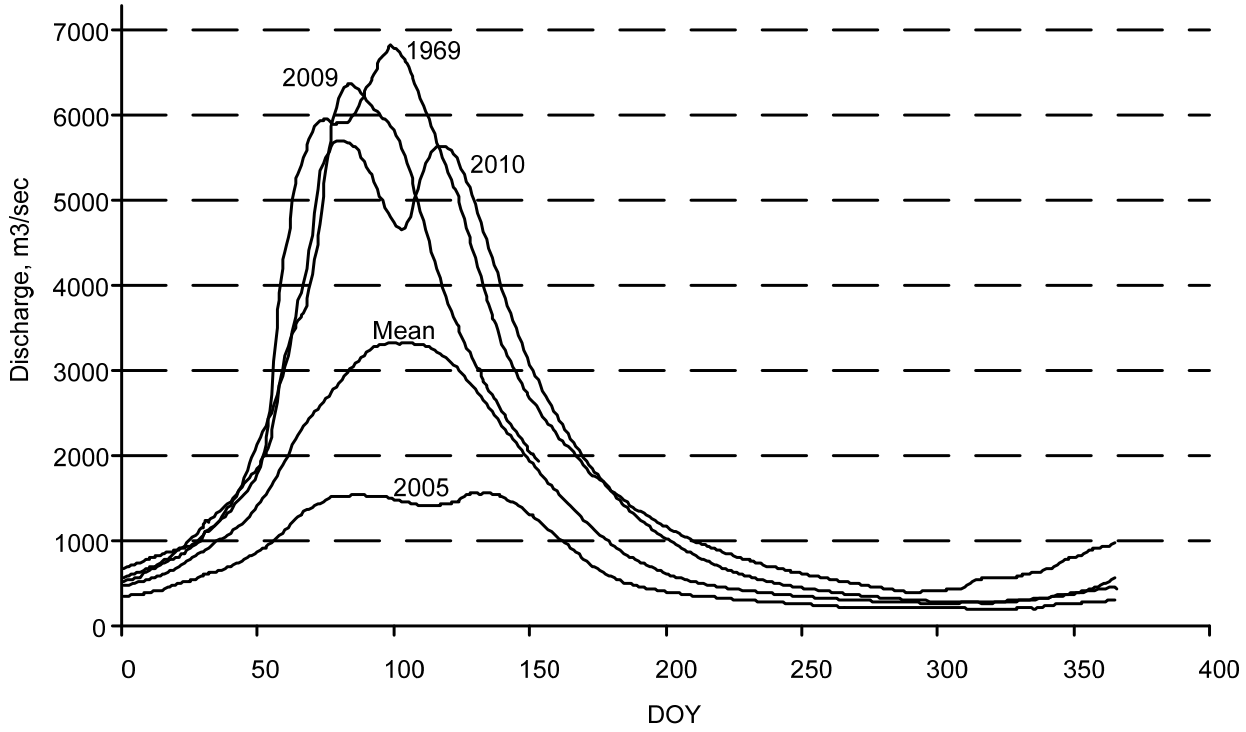


Figure 3. Hydrograph for 1969, 2009, 2010 and 2005 years along with the mean averaged for 1965 to 2010

Recurrence interval estimation

Based on river gauge data we estimated recurrence interval. For each year a maximum discharge was calculated, and these maximum values were sorted in descending order. The year with the maximum discharge was given m=1 magnitude, the year with the second maximum discharge was given m=2 magnitude and so on. These values along with total number records, n=46, were used to calculate recurrence interval (Weibull equation):

$$R = (n + 1)/m. \quad (1)$$

Table 1 shows recurrence interval for the top 10 years with maximum discharge.

A polynomial fit was constructed to predict discharge from recurrence interval. The following 3-order polynomial dependence was obtained:

$$y = 2969.8x^3 - 9567.7x^2 + 11163x + 1181, \quad (2)$$

where y is discharge, and $x = \log_{10}(R)$.

The obtained coefficient of determination was 0.99. Figure 4 shows the plot along with 95% confidence interval.

Table 1. Recurrence interval of floods for the Katima Mulilo region in Namibia

Magnitude, *m*	Year	Discharge, m^3/sec	*R*
1	1969	6817	47.0
2	2009	6365	23.5
3	1978	6251	15.7
4	2010	5704	11.8
5	1979	5675	9.4
6	1976	5568	7.8
7	2007	5564	6.7
8	1975	5409	5.9
9	1968	5312	5.2
10	1966	5276	4.7

Table 2 gives 10-, 50- and 100-year floods values.

Table 2. 10-, 50- and 100-year floods

n, year	Discharge, m^3/sec	Confidence interval
10	5746	[5419; 6073]
50	7093	[6654; 7532]
100	8993	[8131; 9855]

That is, probability of the flood with discharge exceeding 8993 m^3/sec in any given year is equal to 0.01.

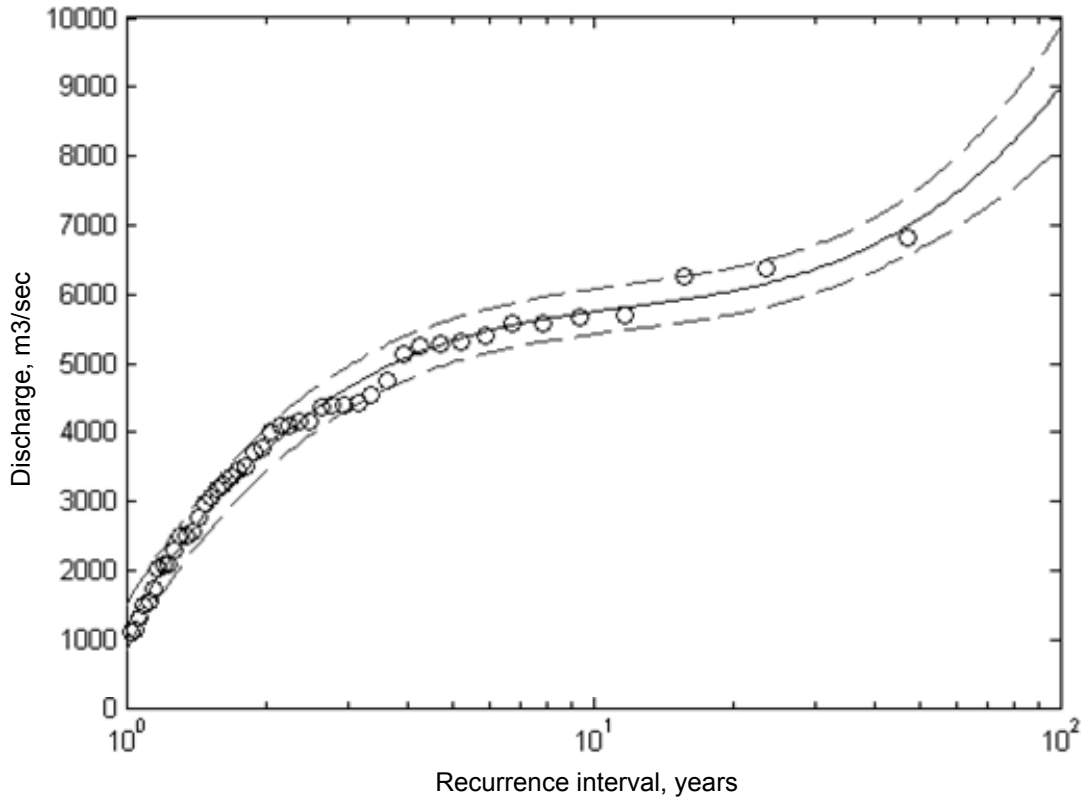


Figure 4. Dependency between maximum discharge values from recurrence interval

Flood hazard mapping using satellite data

We used a time-series of 44 Landsat TM and ETM+ images at 30 m spatial resolution to estimate flood probability density. Firstly, clouds and shadows were identified on Landsat images using the Automated Cloud-Cover Assessment (ACCA) Algorithm [Irish et al, 2006]. Also SLC-off pixels on ETM+ images were marked and removed from consideration. All these pixels were assigned with the "NoData" value, and were removed from the further analysis. At second, water bodies were detected using a density sliding method [Frazier and Page, 2000]. Therefore, each pixel in the image was assigned one of the following values: 0 - «No water» class, 1 - «Water» class, 2 - «No Data». Figure 5 shows the original and processed Landsat-5 image.

Two approaches were used to estimate probability density function. They were different in how images were integrated within the single year. Let A be the set of all satellite images, i.e. $A=\bigcup\{a\}$, where the image a is characterized by the following tuple:

$$a = \{y, doy, (i, j)\}, \tag{3}$$

where $a.y$ and $a.doy$ are the year and day of the year the satellite image was acquired, and $a(i,j) \in \{0, 1, 2\}$ is the value of image pixel with coordinates i and j.

Within the first approach, for each year we selected an image that was closest to the DOY with maximum discharge, and then aggregated these images into a probability of inundation map.

$$A_1 = \left\{ \forall y \in \{2000,\ldots,2010\} : a^* = \min_{a \in A, a.y = y} \left| a.doy - \text{doy\_min\_discharge(y)} \right| \right\}, \tag{4}$$

$$PI_1(i,j) = \frac{1}{\left| \{ a \in A_1 : a(i,j) \neq 2 \} \right|} \sum_{\{a \in A_1 : a(i,j) \neq 2\}} a(i,j). \tag{5}$$

Within the second approach, for each year we aggregated all available images into the single image by assigning a pixel the "Water" class, if at least on one of the images it was identified as the "Water" class. The same was for the "No Water" class. Therefore, the pixel was assigned the "NoData" value only if on all images it had the "NoData" value. These yearly images were then into a probability of inundation map:

$$A_2 = \left\{ \forall y \in \{2000,\ldots,2010\} : a^*(i,j) = 1 \text{ if } \exists a \in \bigcup \{ a \in A : a.y = y \} : a(i,j) = 1 \right\}, \tag{6}$$

$$PI_2(i,j) = \frac{1}{\left| \{ a \in A_2 : a(i,j) \neq 2 \} \right|} \sum_{\{a \in A_2 : a(i,j) \neq 2\}} a(i,j). \tag{7}$$

The main difference in these two approaches is the following. The image generated using the former approach will show flooded areas on the day of a peak discharge (or the closest day). Since the satellite images were not acquired on the day of maximum discharge we, obviously, will miss some flooded areas. Moreover, in the case of Katima Mulilo region there is latency between river flow at the gauge and flow coming to Liambezi Lake. In contrast, the latter approach allows us to identify all the pixels that were flooded during the flood season. Figure 6 and Figure 7 show two maps that were generated using both approaches. We can see that main differences between images are in the south-west part of the area. First approach does not allow us to capture latency in

flood wave that is coming through the region to the Liambezi Lake and to the south, while the second approach does.

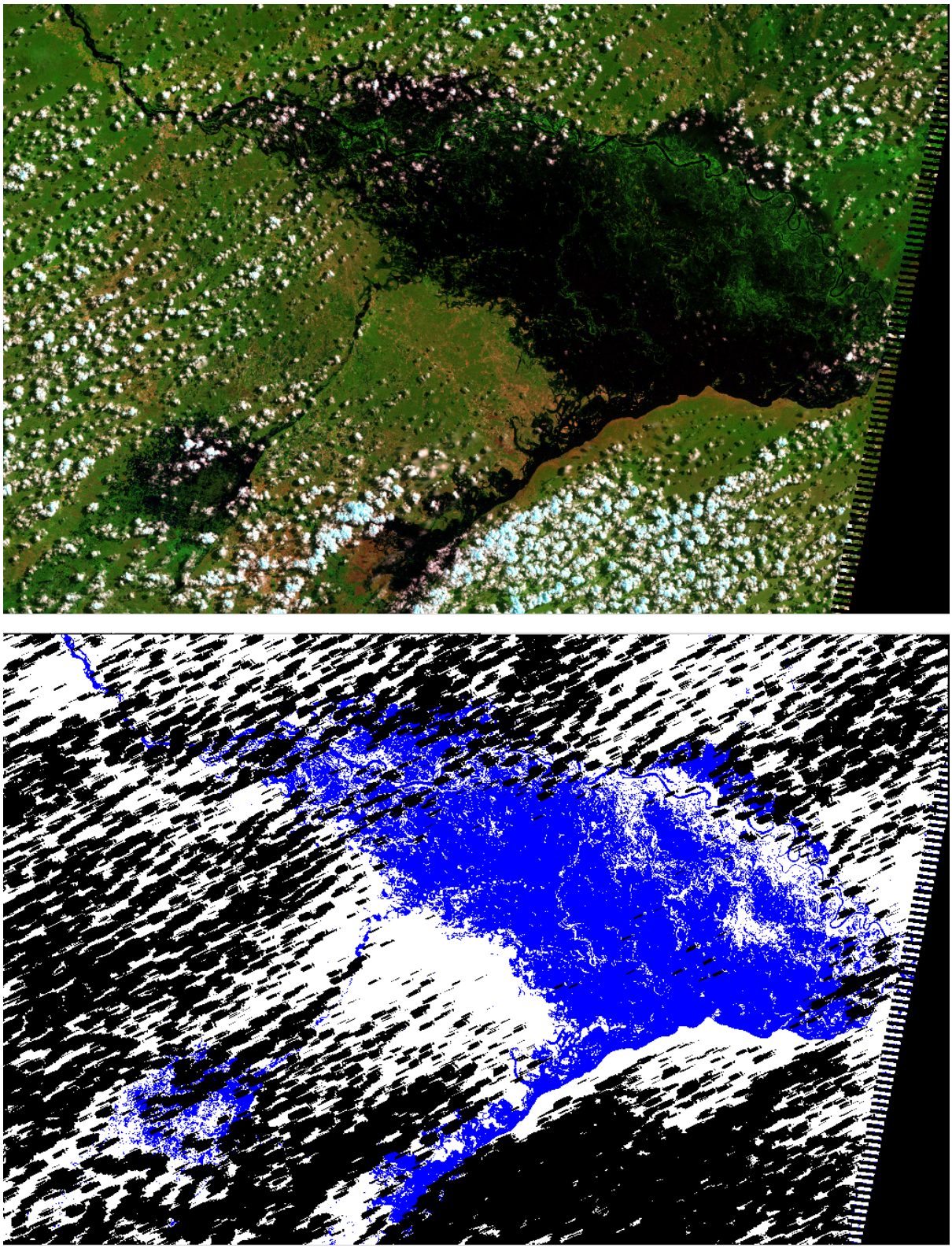


Figure 5. Original (top) and processed (bottom) Landsat-5 image acquired in 2010, DOY=81

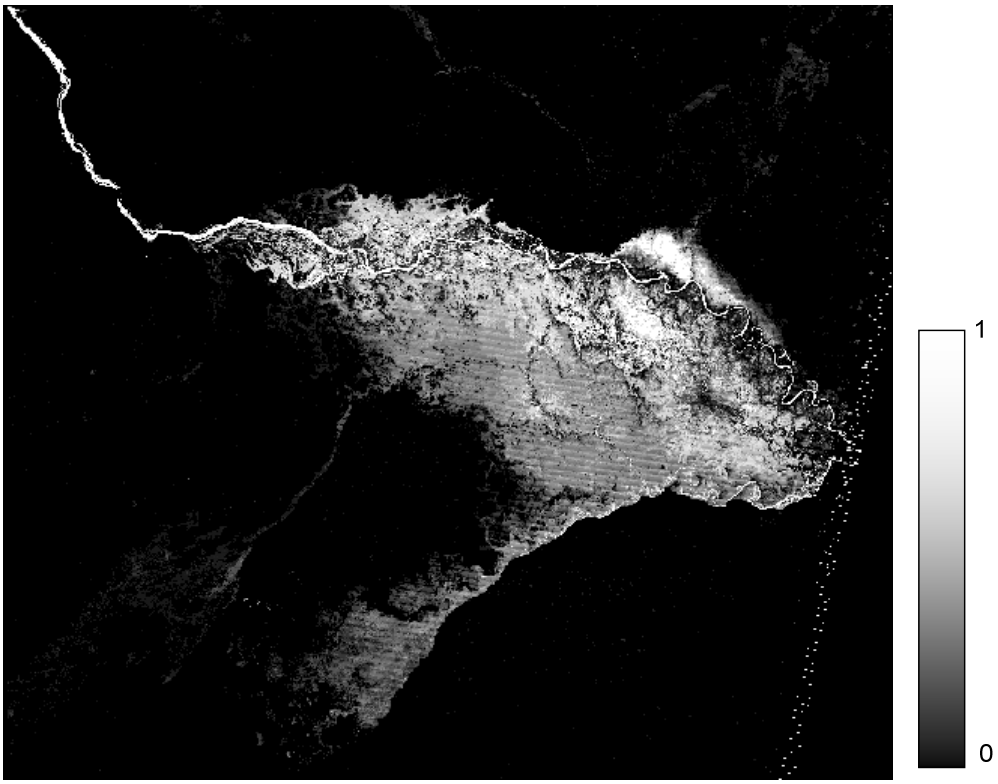


Figure 6. PI for the Katima Mulilo region, Namibia, obtained using Eq. (4)-(5)

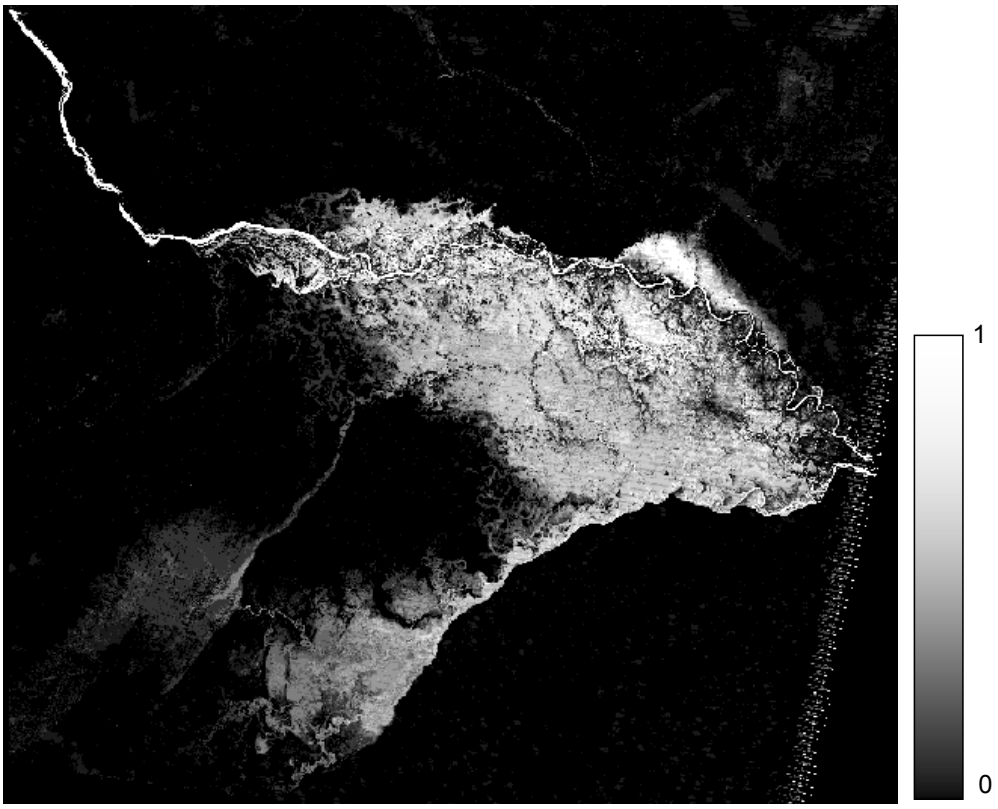


Figure 7. PI for the Katima Mulilo region, Namibia, obtained using Eq. (6)-(7)

Flooded area estimation

Based on the yearly flood extent maps obtained from Eq. (6) we calculated the area of flooded territories. In order to predict the expected flooded area we built a regression that predicts the flooded area in dependence of discharge (Fig. 8). The figure shows a good correspondence between the flooded area and discharge with coefficient of determination of 0.83. The only outlier is the values for the year of 2000. It could be probably explained that for this year a single satellite image was only available, and no maximum discharge was recorded on the date of image acquisition (there was a 1 day difference). In contrast, in 2003, when a single image was also available, the satellite image was acquired on the day when maximum discharge was recorded.

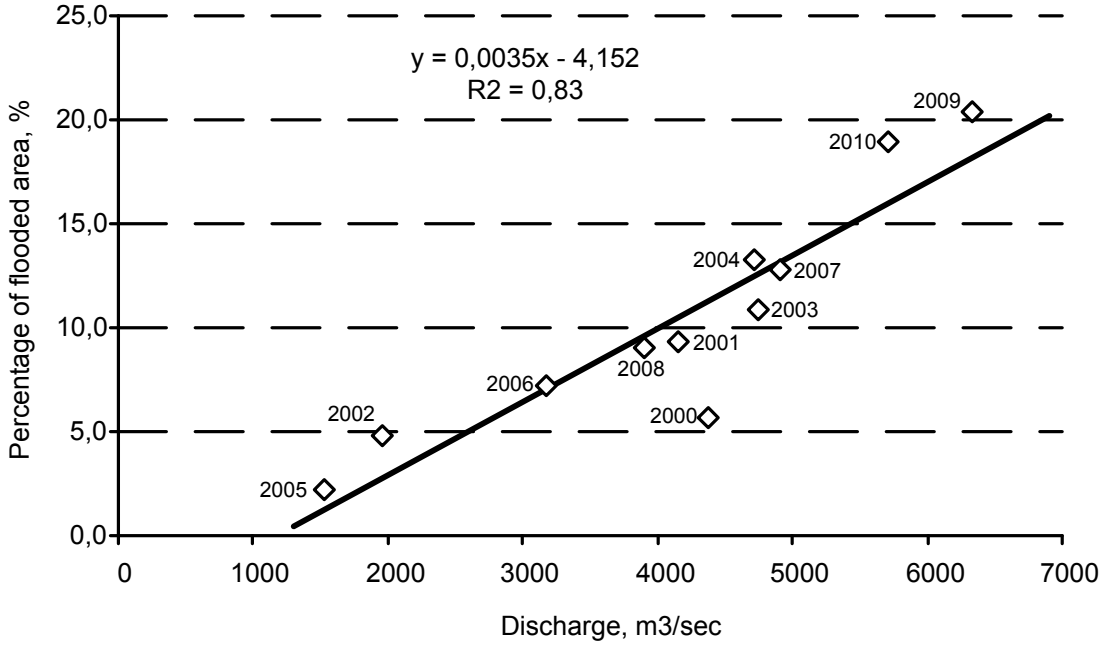


Figure 8. Dependence of the area of flooded territories from discharge

Another issue to be addressed is that we used a pixel counting method for area estimation. Since our classification algorithm does not provide 100% accuracy, the pixel counting method is usually downward biased [Gallego, 2004]. Counting pixels and multiplying by the area of each pixel will result in biased area estimates and should be considered raw numbers needing bias correction. One way to tackle this problem is to provide area frame sampling (AFS) data and then use a regression estimator to improve estimates [Carfagna and Gallego, 2005]. Since it is impractical to provide AFS using ground observations, AFS could be done by photointerpretation. Flooded waters can be reliably identified by visual inspection.

Flood risk mapping

The obtained flood hazard map was integrated with dwelling unit database to provide flood risk mapping. Such analysis allows us to identify the dwellings that are more likely to be inundated during the flood season, and specify probability of being flooded under different scenarios. Figure 9 shows integration flood hazard map with dwellings.

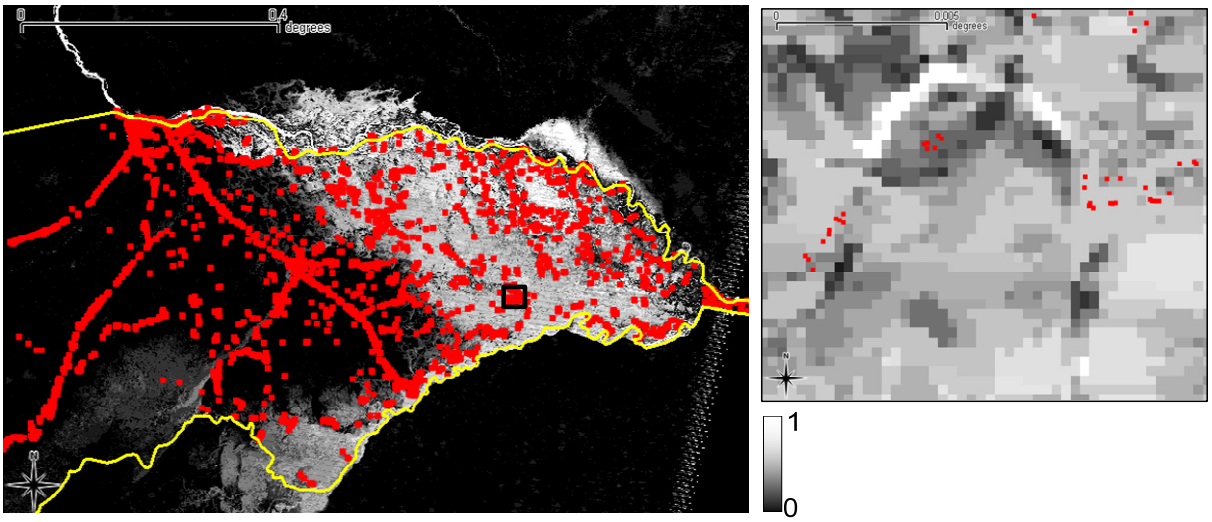


Figure 9. Flood risk map. Red squares show dwelling units

Conclusion, discussion and future works

In this paper we proposed a novel approach for flood hazard mapping by processing and analyzing a time-series of satellite data and derived flood extent maps. This approach is advantageous in cases when the use of hydrological models is complicated by the lack of data, in particular high-resolution DEM. Two approaches were investigated for generating flood extent maps for each year: by selecting an image with date of acquisition closest to the day when the maximum discharge was recorded, and integrating all flood extent maps available for the year. Due to the cloud cover and shadows the former method tends to miss areas that were flooded during the flood season, while the latter accounts for all areas that were flooded. Each pixel of the yearly flood extent map is viewed as Bernoulli distribution value, and maximum likelihood method was applied to estimate a *success probability* from sampling set. This parameter shows probability of inundation, and can be viewed as flood probability density function. Also, we believe that the derived flood extent maps will be very valuable in validating hydrological models once high-resolution DEM is available.

The future works should be directed: (1) to account for uncertainties in pixels with the “NoData” value (which can be either “Water” or “No water”); (2) to build a model that based on flood extent maps, gauge records and low-resolution DEM will predict flooded areas; (3) to provide flood risk mapping based on infrastructure facilities (e.g. roads, enterprises, etc); (4) to integrate radar satellite images with optical ones to reduce the effect of cloud cover.

Acknowledgement

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Bates, 2004] P.D. Bates. Invited commentary: remote sensing and flood inundation modelling. Hydrological Processes, 2004, 18, pp. 2593–2597.

[Bates et al, 1997] P.D. Bates, M.S. Horritt, C.N. Smith, D.C. Mason. Integrating remote sensing observations of flood hydrology and hydraulic modelling. Hydrological Processes, 1997, 11, pp. 1777–1795.

[Carfagna and Gallego, 2005] E. Carfagna, F.J. Gallego. The use of remote sensing in agricultural statistics, International Statistical Review, 2005, 73, 3, pp. 389-404.

[Frazier and Page, 2000] P.S. Frazier, K.J. Page. Water Body Detection and Delineation with Landsat TM Data. Photogrammetric Engineering & Remote Sensing, 2000, Vol. 66, N 12, pp. 1461-1467.

[Gallego, 2004] F.J. Gallego. Remote sensing and land cover area estimation. International Journal of Remote Sensing. 2004, Vol. 25, n. 15, pp. 3019-3047.

[Hoes and Schuurmans, 2006] O. Hoes, W. Schuurmans. Flood standards or risk analyses for polder management in the Netherlands. Irrig. Drain., 2006, 55, pp. 113-119.

[Horritt, 2006] M.S. Horritt. A methodology for the validation of uncertain flood inundation models. Journal of Hydrology, 2006, 326, pp. 153-165.

[Irish et al, 2006] R. Irish, J. Barker, S. Goward, T. Arvidson. Characterization of the Landsat-7 ETM+ Automated Cloud-Cover Assessment (ACCA) Algorithm. Photogrammetric Engineering & Remote Sensing, 2006, Vol. 72, N 10, pp. 1179–1188.

[Jonkmana et al, 2003] S.N. Jonkmana, P.H.A.J.M. van Gelder, J.K. Vrijling. An overview of quantitative risk measures for loss of life and economic damage. Journal of Hazardous Materials, A99, 2003, pp. 1-30.

[Knight, 2006] D.W. Knight. Introduction to flooding and river basin modelling. In: River Basin Modelling for Flood Risk Mitigation, D. Knight and A. Shamseldin (Eds), 2006, pp. 1–20.

[Kussul et al, 2010] N.N. Kussul, B.V. Sokolov, Y.I. Zyelyk, V.A. Zelentsov, S.V. Skakun, A.Yu. Shelestov. Disaster Risk Assessment Based on Heterogeneous Geospatial Information. Journal of Automation and Information Sciences, 2010, Volume 42, Issue 12, pp. 32-45.

[Kussul et al, 2012] N. Kussul, D. Mandl, K. Moe, J.-P. Mund, J. Post, A. Shelestov, S. Skakun, J. Szarzynski, G. Van Langenhove, M. Handy. Interoperable Infrastructure for Flood Monitoring: Sensor Web, Grid & Cloud. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (J-STARS), 2012, doi: 10.1109/JSTARS.2012.2192417.

[Lecca et al, 2011] G. Lecca, M. Petitdidier, L. Hluchy, M. Ivanovic, N. Kussul, N. Ray, V. Thieron. Grid computing technology for hydrological applications. Journal of Hydrology, 2011, Volume 403, Issues 1-2, pp. 186-199.

[Mostert and Junier, 2009] E. Mostert, S.J. Junier. The European flood risk directive: challenges for research. Hydrology and Earth System Sciences Discussions, 2009, Vol. 6, N 4, pp. 4961-4988.

[Rodriguez et al, 2009] J. Rodriguez, F. Vos, R. Below, D. Guha-Sapir. Annual Disaster Statistical Review 2008: The numbers and trends. Centre for Research on the Epidemiology of Disasters, Jacoffset Printers, Melin (Belgium), 2009, 33 p.

[Schumann and Di Baldassarre, 2010] G. Schumann, G. Di Baldassarre. The direct use of radar satellites for event-specific flood risk mapping. Remote Sensing Letters, 2010, Vol. 1, N 2, pp. 75-84.

[See and Abrahart, 2001] L. See, R.J. Abrahart. Multi-model data fusion for hydrological forecasting. Computers & Geosciences, 2001, Vol. 27, N 8, pp. 987-994.

[Yilmaz et al, 2010] K.K. Yilmaz, R.F. Adler, Y. Tian, Y. Hong, H.F. Pierce. Evaluation of a satellite-based global flood monitoring system. International Journal of Remote Sensing, 2010, vol. 31, no. 14, pp. 3763–3782.

Authors' Information

Sergii Skakun *– Senior Scientist, Space Research Institute NASU-NSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine; e-mail: serhiy.skakun@ikd.kiev.ua*
Major Fields of Scientific Research: Grid computing, Sensor Web, Earth observation, satellite data processing, risk analysis.

THEORETICAL ANALYSIS OF EMPIRICAL RELATIONSHIPS FOR PARETO-DISTRIBUTED SCIENTOMETRIC DATA

Vladimir Atanassov, Ekaterina Detcheva

Abstract*: In this paper we study some problems involved in analysis of Pareto-distributed scientometric data (series of citations versus paper ranks). The problems include appropriate choices of i) the distribution type (continuous, discrete or finite-size discrete) and ii) statistical methods to obtain unbiased estimates for the power-law exponent (maximum likelihood procedure or least square regression.). Since relatively low magnitudes of the power exponent (less than 2), are observed massively in scientometric databases, finite-size discrete Pareto distribution (citations, distributed to finite number of paper ranks) appears to be more adequate for data analysis than the traditional ones. This conclusion is illustrated with two examples (for synthetic and actual data, respectively). We also derive empirical relationships, in particular, for the maximum and the total number of citations dependence on the Hirsch index. The latter generalize results of previous studies.*

Keywords*: Scientometrics, Hirsch index, Pareto distributions, data analysis, empirical relationships*

ACM Classification Keywords*: H. Information Systems, H.2. Database Management, H.2.8. Database applications, subject: Scientific databases; I. Computing methodologies, I.6 Simulation and Modeling, I.6.4. Model Validation and Analysis*

Introduction

It is hard to deny that, although not always and not everywhere welcome, *scientometrics* has entered the life of scientific community worldwide. The reasons for that are in much extent the tools it provides for assessment of quantity and quality of scientific output. Nowadays *scientometric indicators* are widely used for variety of purposes, including decision making in projects approval and evaluation, team building, scientific careers promotion, development of university and educational programs, scientific journals ranking, to mention just a few. Almost all of these indicators are based on two primary quantities, namely *number of publications*, considered as a measure of *productivity* and *number of citations* (appearing in the form of references to these publications), as a measure of *impact*, or popularity among the scientists. The attempts to find out a single score of scientific activity have brought to the world a variety of (secondary) scientometric indicators, the Hirsch index [Hirsch, 2005] being probably one of the most popular among them. Based on a simple model, this index suggests some compromise between productivity and impact.

Due to their common origin, scientometric indicators are mutually related *via* theoretically derived or empirically obtained relationships. Studying these relationships is important not only for understanding which indicators indicate what and how (*i.e.* from pure academic point of view), but also for appropriate choice of indicators in various applications. The relations between the indicators essentially depend on the (form of the) ranked citations-papers distribution. The empirical dependence of the total number of citations on the Hirsch index appearing in [Hirsch, 2005] has been supported by the linear negative-slope distribution resulting from the model, while other theoretical studies [*e.g.* Glänzel, 2006; Egghe and Rousseau, 2006] are concentrated on Pareto-distributed scientometric quantities.

In this paper we study several problems associated with the analysis of Pareto-distributed data (citations to paper ranks), suggest solutions to these problems and obtain relationships among indicators that are easy to obtain, or provided by the scientometric databases: total number of citations, number of publications, number of citations of the most cited paper, Hirsch's *h*-index *etc*. The problem of fitting power-law data has been addressed, too. This problem is of crucial importance, in particular with the recent development in that field ([Goldstein *et al*, 2004], [Clauset *et al*, 2009]).

The paper is organized as follows: in the first section we discuss the essential properties of Hirsch's index and the associated model; the second section critically reviews the application of Pareto distributions in scientometrics and the problems that arise in fitting to power-law data. In the next two sections we use the discrete finite-size Pareto distribution for scientometric data analysis and obtain some useful relationships, including the dependence of the total number of citations on the Hirsch's index.

Hirsch's model considerations

We recall the well known *h*-index definition: '*A scientist has index h if h of his or her N papers have at least h citations each and the other N-h papers have no more than h citations each',* and continue with a brief description of the model [Hirsch, 2005] used as its basis. Under the assumption that an individual publishes N_{ppy} papers per year and each published paper is cited N_{cpppy} times every (subsequent) year, the total number of citations N_c earned after N_y ($N_y \gg 1$) years is:

$$N_c \approx \frac{1}{2} N_{ppy} N_{cpppy} (1/N_{ppy} + 1/N_{cpppy})^2 h^2 , \quad (1)$$

where the *h*-index depends linearly on N_y:

$$h \approx (1/N_{ppy} + 1/N_{cpppy})^{-1} N_y . \quad (2)$$

It is worth noting that the scientometric indicators involved in Eqs. (1) and (2) are provided by most bibliometric databases, *e.g.* Web of Knowledge (Thomson Reuters), Scopus (Elsevier), Google Scholar *etc*.

For a time interval of N_y years this simple model yields a linear negative-slope distribution $C(P)$ of citation number C to paper rank P, papers being arranged in descending order of number of citations, *i.e.* the most cited placed first. It follows from Eq. (1) that N_c depends quadratically on h:

$$N_c = Ah^2 . \quad (3)$$

It is easy to see that $A \geq 2$; it reaches its minimum for a straight line of slope -1 (Fig. 1). Moreover, for quite general negative slope concave type of $C(P)$ distribution (Fig. 2) it can be shown that

$$A > A_0 = 1 - \frac{1}{2}(C' + 1/C') \geq 2 , \quad (4)$$

where $C' = (dC/dP)_{P=h}$ is the slope of the distribution at $P = h$. Values of $A < A_0$ indicate negative slope convex type of distributions. It should be noted that Hirsch has empirically found $A \approx 3 - 5$ as a typical value.

Further on it is worth mentioning another (although rather artificial) distribution $C(P)$ – the uniform distribution (Fig. 3 a,b) that could be considered as a *limiting case* of convex negative-slope type distribution. It clearly

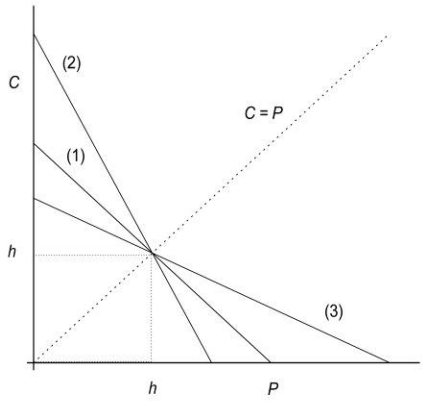


Fig. 1. Illustration of linear negative-slope distributions: (1) slope $= -1$, $A = 2$; (2) slope < -1, $A > 2$ (excess of citations); (3) slope > -1, $A > 2$ (excess of papers).

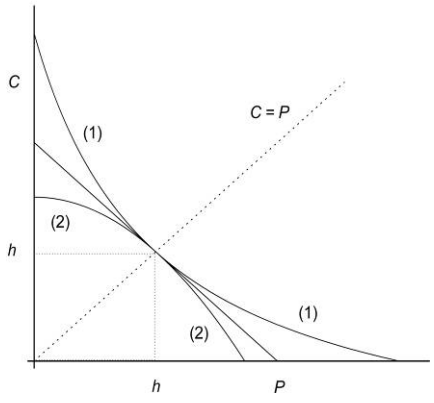


Fig. 2. Examples of negative slope concave (1) and convex (2) type of $C(P)$ distributions

demonstrates the limitations to *h, i.e. h* cannot exceed the maximum number of citations $C_{\max}$ and the maximum number of papers $P_{\max}$:

$$h = \min(C_{\max}, P_{\max}), \tag{5}$$

and the total number of citations is:

$$N_c = \max(C_{\max}, P_{\max})h. \tag{6}$$

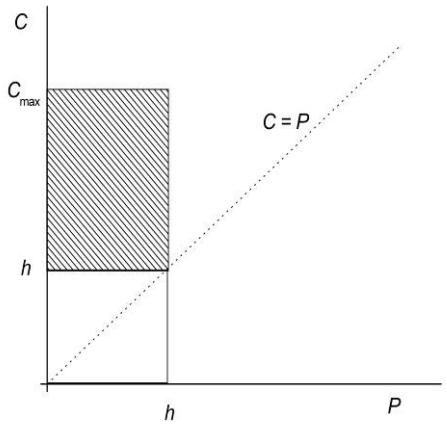


Fig. 3 a. Uniform $C(P)$ distribution example: excess of citations, *h*-index limited by maximum number of papers $P = h$

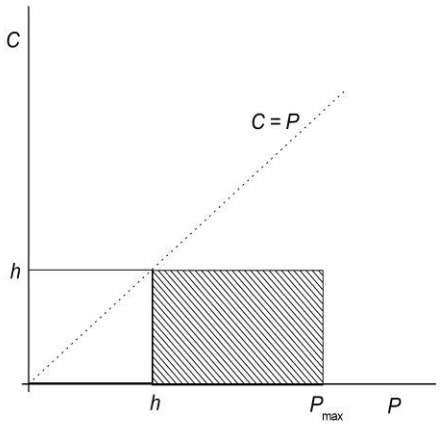


Fig. 3 b. Uniform $C(P)$ distribution example: excess of papers, *h*-index limited by maximum number of citations $C = h$

Pareto distributions in scientometrics

The distributions of number of citations to papers' rank $C(P)$ considered in the previous section concern rather sparse examples of data sets; the bulk of scientometric data is mainly characterized by negative slope concave type probability densities. One of the simplest types of such distributions has been used by *Alfredo Pareto* (1848-1923) for solving problems in economics, as allocation of wealth *etc*. *Pareto distribution* is a special case of *power-law distribution* with negative exponent, defined as strictly zero below some (positive) number that we shall assume to be 1:

$$PDF(X) = (\alpha - 1)X^{-\alpha}, CPF(X) = 1 - X^{1-\alpha}, X \geq 1 \quad (7)$$

Three examples for (continuous) Pareto distribution as well as one for its discrete version (called also *Zeta distribution*, defined for positive integers):

$$Probability(I) = (1/\zeta(\alpha))I^{-\alpha}, \alpha > 1, I = 1, 2, 3, \ldots \quad (8)$$

are demonstrated on Fig. 4 a,b. In Eq. 8 $\zeta(\alpha)$ is the Riemann zeta-function, tabulated in [Walther, 1926], [Janke *et al*, 1960] *etc*.

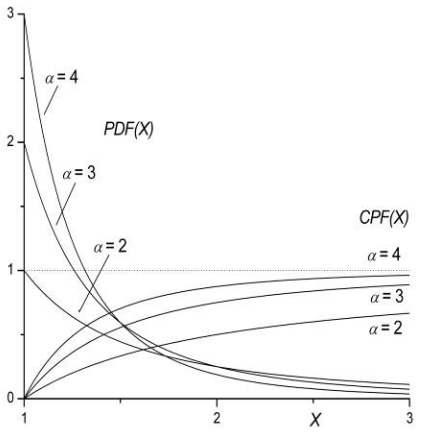


Fig. 4a. Pareto distribution examples: probability density $PDF(X) = (\alpha - 1)x^{-\alpha}$ and cumulative probability function $CPF(x) = 1 - x^{-\alpha}$ for $\alpha = 2, 3$ and 4.

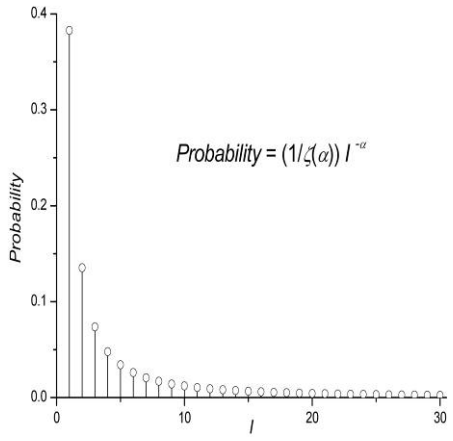


Fig. 4b. Discrete Pareto (*Zeta*) distribution example: $Probability(I) = (1/\zeta(\alpha))I^{-\alpha}, \alpha = 1.5$

Both (continuous and discrete) Pareto distributions should be handled with care bearing in mind that all moments of order *k* and above simply *do not exist* for $k \geq \alpha - 1$. For instance, a (rather popular, see [Goldstein *et al*, 2004] and [Clauset *et al*, 2009]) Pareto distribution with $\alpha = 2.5$ has infinite dispersion and does not meet the requirements of the *Central Limit Theorem*. Probability densities with $1 < \alpha \leq 2$ have no mean, and those with $\alpha \leq 1$ are no distributions at all.

There are some questions and problems that appear in connection with Pareto distribution applications in scientometrics and we address them further on. The first one is: are citation numbers really Pareto-distributed to paper ranks? An acceptable answer to this important question could be found in [Clauset *et al*, 2009]: it states that *power law* is a plausible choice, just as *log-normal* and *stretched exponential* (see, *e.g.* [Hirsch, 2005]) are;

all other statistical hypotheses could be rejected. Testing of power-law hypothesis is a tough problem that remains beyond the scope of this study and further on we assume that data is *apriori* Pareto distributed.

Another point of Pareto distributed data analysis, namely data fit and derivation of unbiased power-law exponent estimate seems to be a 'hot potato', too (see [Goldstein *et al*, 2004], [Clauset *et al*, 2009]). The analyses of synthetic data described in above mentioned papers reveal that the straightforward log-log *ordinary least square regression* (OLSR, see *e.g.* [Weisberg, 2005]) gives essentially distorted estimates for the power-law exponent; the authors recommend the use of *maximum likelihood estimate* (MLE) instead. For continuous Pareto-distributed data the latter looks like:

$$\alpha_{est} = 1 + (\frac{1}{n}\sum_{i=1}^{n} \ln x_i)^{-1}, \tag{9}$$

while for discrete Pareto-distributed data one has to solve a transcendental equation:

$$\zeta'(\alpha_{est})/\zeta(\alpha_{est}) = -(1/n)\sum_{i=1}^{n} \ln x_i \ . \tag{10}$$

In Eqs. (9) and (10) $\zeta'(\alpha)$ is the first derivative of the Riemann zeta-function $\zeta(\alpha)$, (see Fig. 5), α_{est} is the MLE for the power-low exponent and $\{x_i; i = 1,2,...,n\}$ represents a *sample* containing *n independent data*. Again we refer to [Clauset *et al*, 2009] for estimates of the MLE standard error.

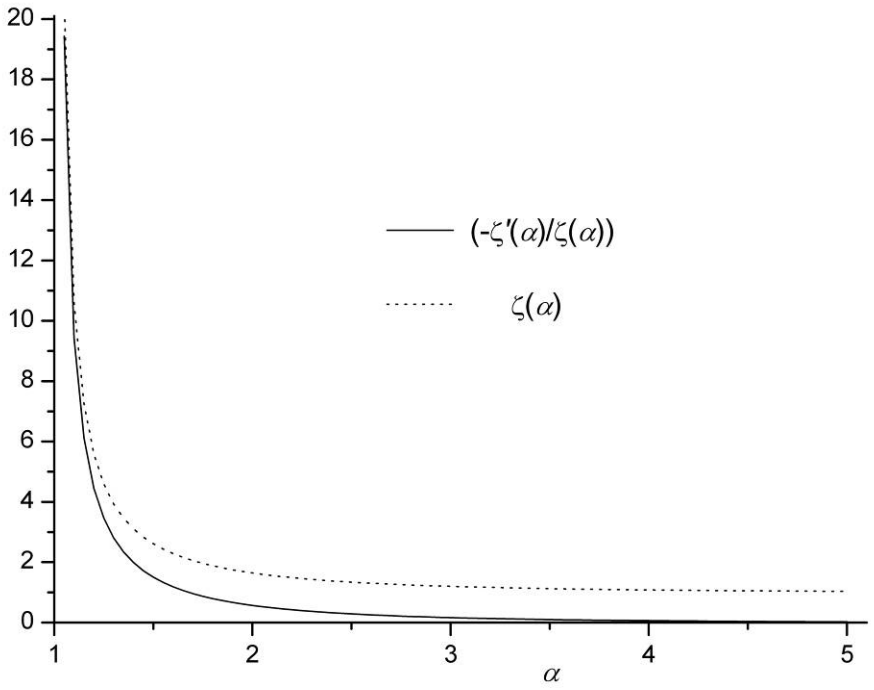


Fig. 5. Riemann zeta-function $\zeta(\alpha)$ and its logarithmic derivative $-\zeta'(\alpha)/\zeta(\alpha)$.

The last but not least important item of our list of questions and problems is the (estimated) value of the power-law exponent α **itself.** Theoretical studies on Pareto-distributed scientometric data (*e.g.* [Glänzel, 2006]) consider samples with essentially infinite number of papers *N* and finite mean (*i.e.* first moment, $k = 1$). As we have

already mentioned, this imposes a restriction $\alpha > 2$ for most results obtained there. The point is, whether citation data resulting from such distributions exist in any extent of importance. One could argue that data sets with 90 percent of citations concentrated in the first four papers (as it follows from a distribution with $\alpha = 2$) are extremely rare to find in bibliometric databases. Usually $\alpha < 2$, and values close to (and even less than) 1 could be found, too. For example, Schreiber's dataset [Schreiber, 2007] used for analysis of self-citations effect on the *h*-index is characterized by $\alpha = 1.35$ (MLE) and $\alpha = 0.70$ (log-log OLSR estimate with $R^2 = 0.98$). Hence it is necessary to find out a scheme that can be used to analyze scientometric data in the 'twilight zone' $1 \le \alpha \le 2$, evading the pole of $\zeta(\alpha)$ at $\alpha = 1$ and moments that diverge even for the lowest orders.

Finite size discrete Pareto distributions

Introducing some kind of upper limit (cut-off) for all summations or integrations is usually considered as universal remedy ('painkiller') for the problems similar to those we discussed in the previous section. It is not a secret for anyone trying to compute directly the zeta function that millions of terms must be summed up to achieve somewhat acceptable accuracy. This, however, does not correspond to the simple fact that individual scientist's production normally does not exceed several hundred papers. Therefore, it does not make much sense in merging slowly converging tails to the *actual* probability density by approximating it with discrete infinite-sized Pareto distribution. The latter could be replaced by a finite size distribution:

$$Probability(I,N) = (1 / S(\alpha,N))I^{-\alpha}, I = 1,2,...,N, \quad (11)$$

where N is the number of papers *cited at least once* (or its estimate) and the incomplete zeta function (Fig. 6) is:

$$S(\alpha,N) = \sum_{I=1}^{N} I^{-\alpha} \quad (12)$$

Since $S(\alpha,N)$ depends slowly (within an interval from several tens to several hundreds) on N, we can perform the standard maximum likelihood optimization to obtain the MLE for α:

$$S'(\alpha_{est},N)/S(\alpha_{est},N) = -(1/n)\sum_{i=1}^{n} \ln x_i, \quad (13)$$

where $S'(\alpha,N) = \partial S(\alpha,N) / \partial\alpha$ (Fig. 7.). Figs 6 and 7 give a notion for why and where the standard discrete Pareto analysis fails to give acceptable (unbiased) estimates for the power exponent α.

In order to illustrate how this scheme works we have chosen two examples: a synthetic data one (Fig. 8) and one with actual data (Fig. 9). Synthetic data have been computed by using

$$I_c(I) = \mathrm{nint}(I_{c\max} I^{-\alpha}), I = 1,2,...,N, \quad (14)$$

where the maximum number of citations, *i.e.* the number of citations gained by the first rank paper is estimated for given total number of citations N_c and power-law exponent α as

$$I_{c\max} = I_c(1) = N_c / S(\alpha,N). \quad (15)$$

Thus rounding error was the only one introduced by the computation. The actual data have been adopted from Thomson Reuters Web of Knowledge database. Four kinds of analysis have been performed in both cases, as follows: 1. C+U MLE denotes maximum likelihood estimate for continuous, infinite-size (*i.e.* unlimited) argument Pareto distribution, by using Eqs. (7) and (9); 2. D+U MLE denotes MLE for discrete infinite-size (unlimited) argument discrete Pareto distribution, obtained *via* Eqs. (8) and (10); 3. D+L MLE corresponds to MLE for discrete finite size (limited argument) Pareto distribution according to Eqs. (11)-(13) and 4. log-log OLSR stays for

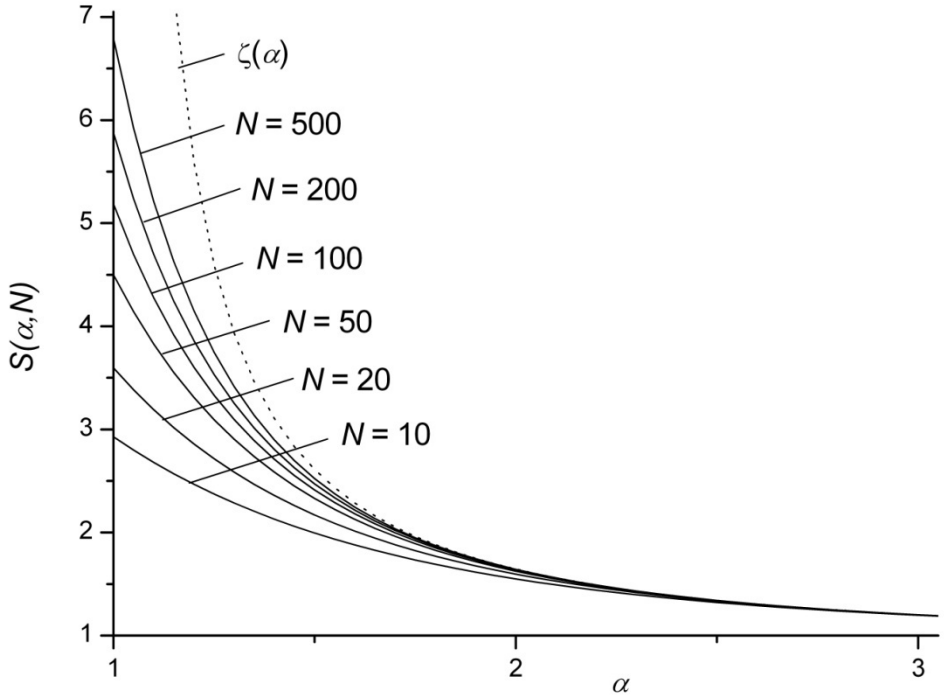


Fig. 6. The function $S(\alpha,N)$ versus α for various N. Note that it closely approaches $\zeta(\alpha)$ for $\alpha > 2$.

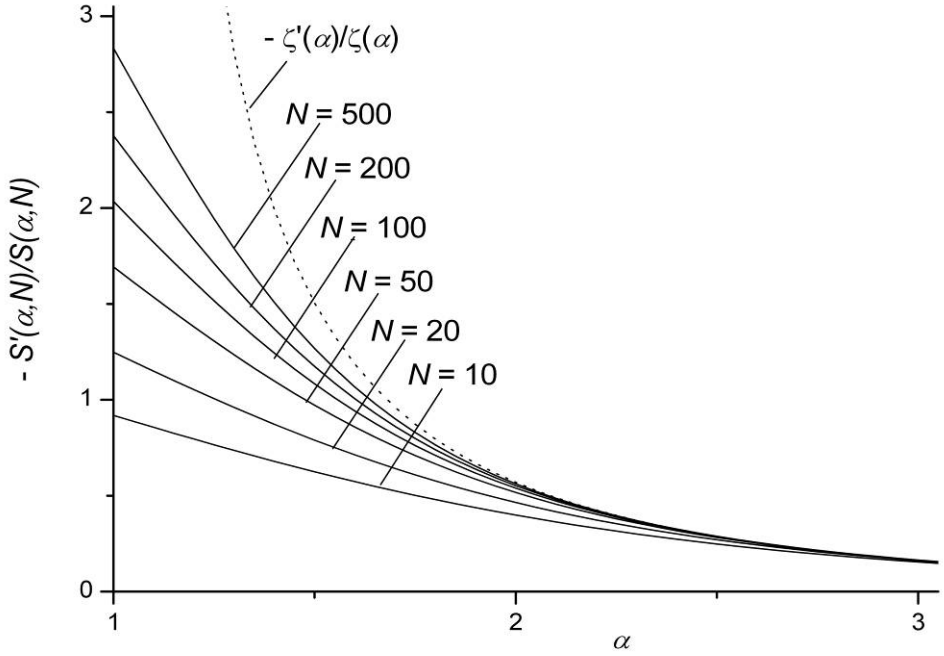


Fig. 7. The function $-S'(\alpha,N)/S(\alpha,N)$. Note that it closely follows $-\zeta'(\alpha)/\zeta(\alpha)$ for $\alpha > 2.5$.

log-log ordinary linear regression. The results of these analyses are summarized in Tables 1 and 2 for the synthetic and actual data, respectively.

Table 1. Estimates of α and I_{cmax} for synthetic data $I_c = \text{nint}(I_{cmax} I^{-\alpha})$, $\alpha = 1.5$, $I_{cmax} = 213$, $N = 56$ (Fig. 8)

	C+U MLE	D+U MLE	D+L MLE	log-log OLSR
α_{est}	1.98	1.67	1.48	1.37
I_{cmax}	495	233	210	158
n	506	506	506	$(R^2 = 0.97)$

Table 2. Estimates of α and I_{cmax} for real data $I_C(I)$ with $I_{cmax} = 62$, $N = 16$, $N_c = 234$, $h = 8$ (Fig. 9.)

	C+U MLE	D+U MLE	D+L MLE	log-log OLSR
α_{est}	1.94	1.65	1.09	1.37
I_{cmax}	220	108	76	112
n	234	234	234	$(R^2 = 0.85)$

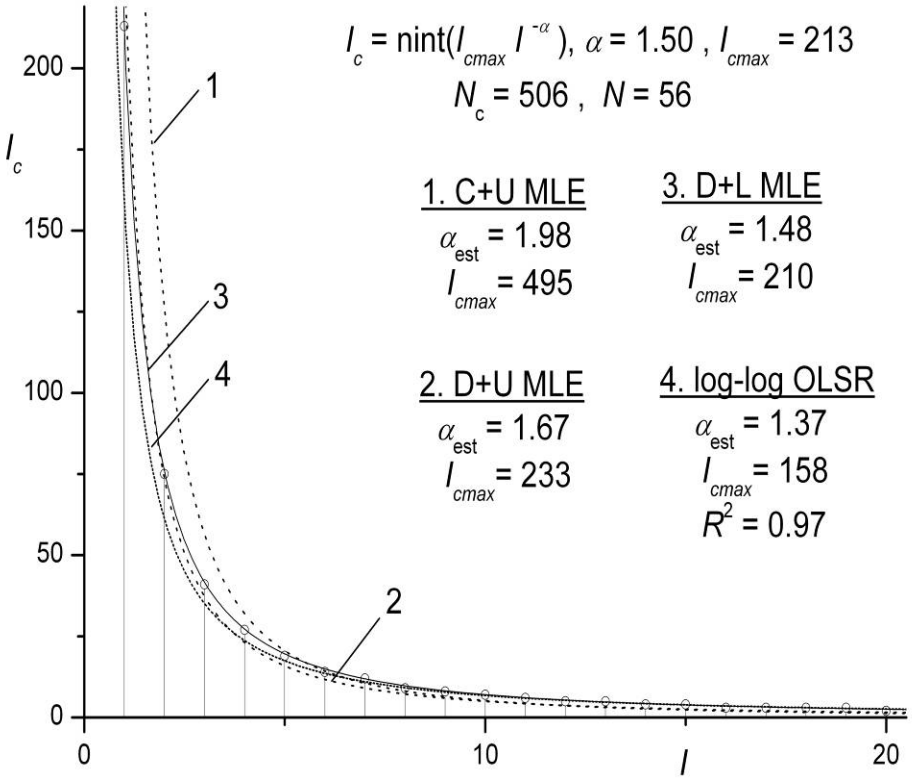


Fig. 8. Synthetic data example: number of citations I_c versus paper rank I.

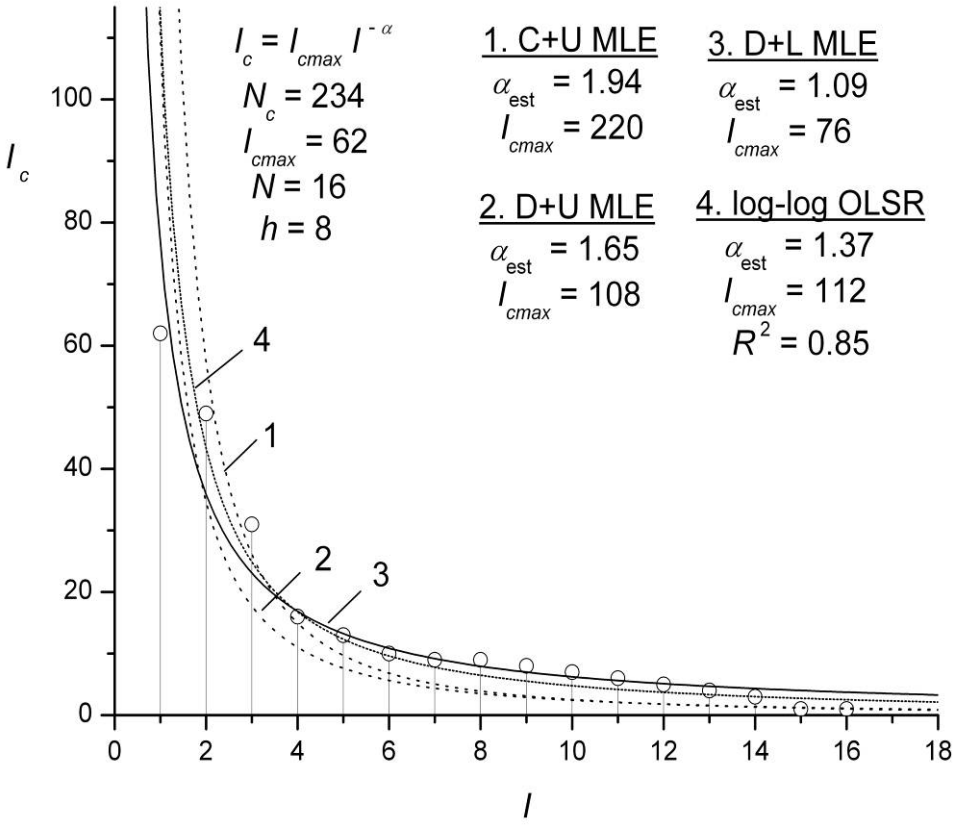


Fig. 9. Real world data example: number of citations I_c versus paper rank I.

Both examples demonstrate the complete failure of the continuous infinite-sized Pareto distribution analysis (Eqs. (7) and (9)) to provide an acceptable unbiased estimate for α and $I_{c\max}$ (for the synthetic data example) and $I_{c\max}$ (for the actual data). The MLE, resulting from an infinite-size discrete Pareto distribution -Eqs. (8) and (10), is usually considered to be the best one and this might be true for $\alpha > 2$. In our synthetic data example, however, it tends to overestimate the prescribed power-law exponent ($\alpha = 1.5$), while the log-log ordinary least square regression underestimates it. The latter phenomenon has been reported earlier in the simulation experiments of [Goldstein *et al*, 2004] and [Clauset *et al*, 2009]. We point out that, without any doubt, the discrete finite size MLE (Eqs. (11)-(13)) performs best for both synthetic and real life data.

Empirical relationships for Pareto-distributed scientometric data

Let us now consider some relationships between scientometric parameters following from the assumption for citations that are (discrete) Pareto distributed to paper ranks. From the definition of *h*-index $I_c(h) \approx h$ we obtain

$$I_{c\max} \approx h^{1+\alpha}. \qquad (16)$$

This relation holds for all types of Pareto distributions – continuous, discrete, infinite or finite-sized; it also matches quite well the actual scientometric data (Fig. 9). Note that in the Hirsch's model I_{cmax} depends *linearly* on *h:*

$$I_{cmax} \approx (1 + N_{cpppy} / N_{ppy})h . \quad (17)$$

By replacing I_{cmax} in (15) we arrive at

$$N_c \approx S(\alpha, N)h^{1+\alpha} . \quad (18)$$

Eq. (18) relates the total number of citations N_c the $(1+\alpha)$-th power of the *h*-index *via* the (slowly depending on the power law exponent α and on the number of publications cited at least once, N) function $S(\alpha, N)$. It represents a *generalization* of Hirsch's relationship (3) for Pareto-distributed scientometric data. Moreover, one can see on Fig. 10 that for reasonable choice of N (50-350) and α (1.05-1.50) the coefficient $S(\alpha, N)$ in (17) varies between 2.5 and 5.5, *i.e.* close to the Hirsch's empirical result $A = 3 - 5$ [Hirsch, 2005].

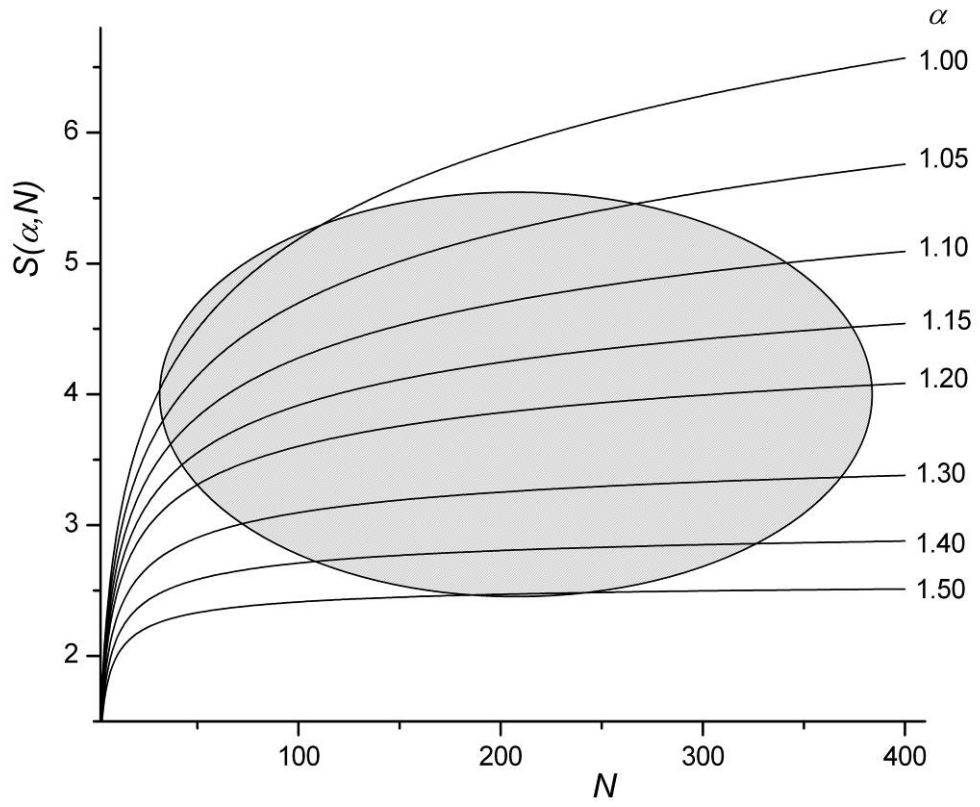


Fig. 10. $S(\alpha, N)$ versus number of publications N for various α .The area in grey roughly pictures the range of Hirsch's empirical evaluation of A .

In the limit $N \to \infty$ we have $S(\alpha, N) \to \zeta(\alpha)$. A brief inspection of Fig. 5 allows us to conclude that $\zeta(\alpha)$ might be considered as slowly varying for $\alpha > 2$. This however, implies dependence stronger than $N_c \sim h^3$. In addition, the continuous version of Eq. (17):

$$N_c = (\alpha - 1)^{-1} h^{1+\alpha}, \tag{19}$$

has severe problems for α close to unity and is obviously not applicable to analyze scientometric data.

The Pareto assumption allows us to obtain another relation that might be useful when the sample under consideration contains publications, all of them with nonzero number of citations, *i.e.* $I_c(I) \neq 0$ for I=1,2,...,N. This means that the actual number of papers cited at least once is $N_a \geq N$. In order to estimate the actual upper limit of the discrete finite-size Pareto distribution one could take into account the fact that by definition the number of citations I_c is always a positive integer. Bearing in mind the **nint** (nearest integer) convention as well as Eqs. (14) and (16) we obtain

$$N_a = (2I_{c\max})^{1/\alpha} = 2^{1/\alpha} h^{(1+\alpha)/\alpha}, \tag{20}$$

as a crude estimate for the actual number of publications with at least one citation. One could compare it with the maximum number of publications that follows from the Hirsch's model:

$$N_{\max} = (1 + N_{ppy} / N_{cpppy}) h \quad , \tag{21}$$

and linearly depends on the *h*-index.

Summary and conclusions

In this paper we have studied some problems appearing in the analysis of Pareto-distributed scientometric data in the form of series of citations versus paper ranks. These problems include appropriate choice of the distribution type (continuous, discrete or finite-size discrete) and of statistical methods to obtain unbiased estimates for the power-law exponent (maximum likelihood procedure or least square regression). Further on, we have theoretically derived relationships, in particular, the total number of citations dependence on the Hirsch index, that generalize results of previous studies and may be proved empirically.

Our conclusions are summarized as follows:

- Pareto-distribution analysis of citations versus paper rank data requires use of probability densities with power-law exponent **of** magnitude 2 or less. Therefore it is necessary to use finite-sized (*i.e.* defined for a finite number of papers) discrete Pareto distribution;
- The maximum likelihood estimate of the power-law exponent for the finite-size discrete Pareto distribution seems to provide best fit to the data, while those resulting from the maximum likelihood procedure for infinite series discrete distribution and from the log-log ordinary least square regression give overestimated and underestimated values. The continuous Pareto analysis proves to be inappropriate for this kind of data;
- The maximum number of citations gained by a paper (this is the paper of rank 1) is a power function of Hirsch's index; the same holds for the total number of citations, however, with constant of proportionality weakly depending on power exponent and number of papers;
- Papers of zero citation count are irrelevant for Pareto distribution analysis of scientometric data. The number of papers with at least one citation is power function of the Hirsch's index, too.

Acknowledgments:

The authors wish to acknowledge the kind invitation and support of Dr K. Markov and Dr S. Poryazov. This paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Clauset et al, 2009] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data, J. SIAM Review 51 (4) 661-703 (2009)

[Egghe and Rousseau, 2006] L. Egghe and R. Rousseau. An informetric model for the Hirsch-index, Scientometrics 69 (1) 121-129 (2006)

[Glänzel, 2006] W. Glänzel. On the h-index – A mathematical approach to a new measure of publication activity and citation impact, Scientometrics 67 (2) 315-321 (2006)

[Goldstein et al, 2004] M. L. Goldstein, S. A. Morris, and G. G. Yen. Problems with fitting to the power-law distribution, Eur. Phys. J. B – Condensed Matter and Complex Systems 41 (2) 255-258 (2004)

[Hirsch, 2005] J.E. Hirsch. An index to quantify an individual's scientific research output, Proc. Nat. Acad. Sci. 102 (46) 16569-16572 (2005)

[Janke et al, 1960] E. Janke, F. Emde and F. Lösh, Tafeln Höherer Funktionen, B. G. Teubner Verlagsgesellschaft, Stuttgart, 1960

[Schreiber, 2007] M. Schreiber. Self-citation corrections for the Hirsch index, Eur. Phys. Lett. 78 30002 (2007)

[Walther, 1926] A. Walther. Anschauliches zur Riemannischen Zetafunktion, Acta Mathematica 48 (3-4) 393-400 (1926)

[Weisberg, 2005] S. Weisberg, Applied Linear Regression, Wiley-Interscience, 2005

Authors' Information

Vladimir Atanassov *– Institute of Electronics, Bulgarian Academy of Sciences, 1784 Sofia, Bulgaria; e-mail: v.atanassov@abv.bg*

Major Fields of Scientific Research: Plasma Physics and Gas Discharges, Radars & Ocean Waves, Nonlinearity & Chaos, Scientometrics

Ekaterina Detcheva *– Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,1113 Sofia, Bulgaria; e-mail: detcheva@math.bas.bg*

Major Fields of Scientific Research: Web-based applications, Image processing, analysis and classification, Knowledge representation, Business applications, Applications in Medicine and Biology, Applications in Psychology and Special Education, Computer Algebra.

ANALYSIS AND JUSTIFICATION FOR SELECTION PARAMETERS OF WIRED ACCESS SYSTEMS

Svetlana Sakharova

Abstract*: The executed researches belong to area of design of perspective access networks. Work is devoted to the analysis of parameters of access networks and a choice of the most significant among them. Results of researches for wire decisions of the organization of a network are given.*

Keywords*: access network, parameters of access networks.*

ACM Classification Keywords*: C.2. Computer-communication networks, H. Information Systems - H.1 Models and Principles, K. Computing Milieux - K.6 Management of computing and information system.*

Introduction

Development of telecommunications leads to emergence of new infocommunicatons services (ICS) and at the same time the problem of access to them becomes complicated. At introduction of each new service possibly change of requirements to a network and the network equipment due to the need of ensuring capacity, types of a traffic and the procedures of service specific to this service.

The subject of research is connected with one of the main problems in the field of telecommunications which creation of *Next Generation Network* (*NGN*), intended for granting to users of all set IKS. Component of NGN are access networks (AN), providing to users access to all ICS on the common line of access (LA). The LA parameters define quality and the nomenclature to ICS available to users of NGN. Need of creation AN is specified in documents of the International Telecommunication Union (ITU) as one of the most important problems of 21 eyelids [1], and prospects of their development are widely discussed at conferences and seminars the infocommunication devoted to development. Various aspects of AN creation are covered in works of Sokolov N. A., Goldstein B. S., Baklanov I.G., Krendzel A.V., Hilenko V. V., Mikhaylov V. F., Gayvoronska G. S., Balashov V.A., Zyablov S. V. and others.

Creation of AN now got a special urgency as the site of access is that segment of a telecommunication network (TN) which introduction of broadband high-quality IKS at the expense of which the operator can have considerable profit brakes. Besides, modernization of user's networks and creation on their base of perspective AN is the third stage of the TN transformation, the first which two stages is replacement of analog systems of transfer and switching nodes on digital. These stages are carried out by increasing rates, and time of the third final stage which is directly connected with a work subject now came.

At the present time design of AN is based on methods of calculation of user's telephone systems that is inadmissible as An differ both structure and functions which they carry out, and as a set of parameters therefore AN should be created on other principles. As the concept of AN [2] is developed rather recently, today there are no approved methods of their design therefore there was a need for development of sequence of process of creation of AN, and for this purpose it is necessary to carry out the analysis of the AN parameters influencing process of design.

Research problem statement

Now a huge variety of technologies of access to IKS is applied, there is a set of views and approaches to creation of AN [3-5]. At creation of AN it is necessary to consider a set of the AN parameters and the requirements which are put forward to it [6, 7]. As studied AN have a set of parameters, distinct from parameters of existing user's networks, it is necessary to carry out the analysis and to make the characteristic of these parameters, to reveal correlation between them and to bring them into a form convenient for modeling. Thus, directed by a task need of the solution of a task of the analysis therefore the list of the parameters being initial AN for creation is defined, and what not essentially influence process of creation of a network and which account can be neglected is noted. Research objective is increase of efficiency of design of perspective AN, decrease in expenses for their creation and increase of efficiency of operation.

The structural composition of AN

Many approaches it is possible to carry out classification, group and the description of access networks. However more often the traditional method is applied. It consists of two components: text description and graphic display. Feature of the text description is that such method gives the chance to the reader to familiarize with studied object in full. But the text description doesn't provide an objective image of classification structure of construction. Therefore the text description is often supplemented with graphic display of classification model.

Graphic display, in turn, is divided into two forms of representation: in the form of tabular structure and treelike structure. The tabular structure is used for representation of the general classification elements of analyzed objects more often. Usually, the tabular structure contains numerical values for possibility of comparison of studied objects. Application of this way leads to the compressed representation.

The treelike structure provides most volume visual representation about object of research. Step-by-step movement on branches of classification model allows to track and see distinctive features in creation of objects at one level of a tree.

As a textual description and graphical representation are mutually beneficial, then the possibilities as far as possible to characterize AN, involved in both methods. The graphic part of work displays treelike structure of classification model of creation of AN. For the general acquaintance with structure of model the treelike scheme displaying the general elements of classification for systems of wire access is provided. But there are hidden all subtleties of construction for each method of realization of AN.

The separate treelike structure gives deeper idea of objects. In work it is presented as addition to the general classification model of creation of AN.

By consideration and the analysis of a large number of existing decisions on AN realization at present, they can be grouped in the following signs: in a form of a transferred signal; to destination access networks; on capacity; on structural construction; on applied technologies; on management; by the form access; on a class of the served district (figure 1).

Group of the AN parameters on a served class of the district

For distribution of parameters on their importance in the course of creation of AN signs on which it is possible to make classification of parameters are allocated. As an example, it is possible to carry out classification of parameters for AN on a class of the served district since at design of a network it is necessary to consider features of that district for which it is under construction.

The following parameters are taken into account.

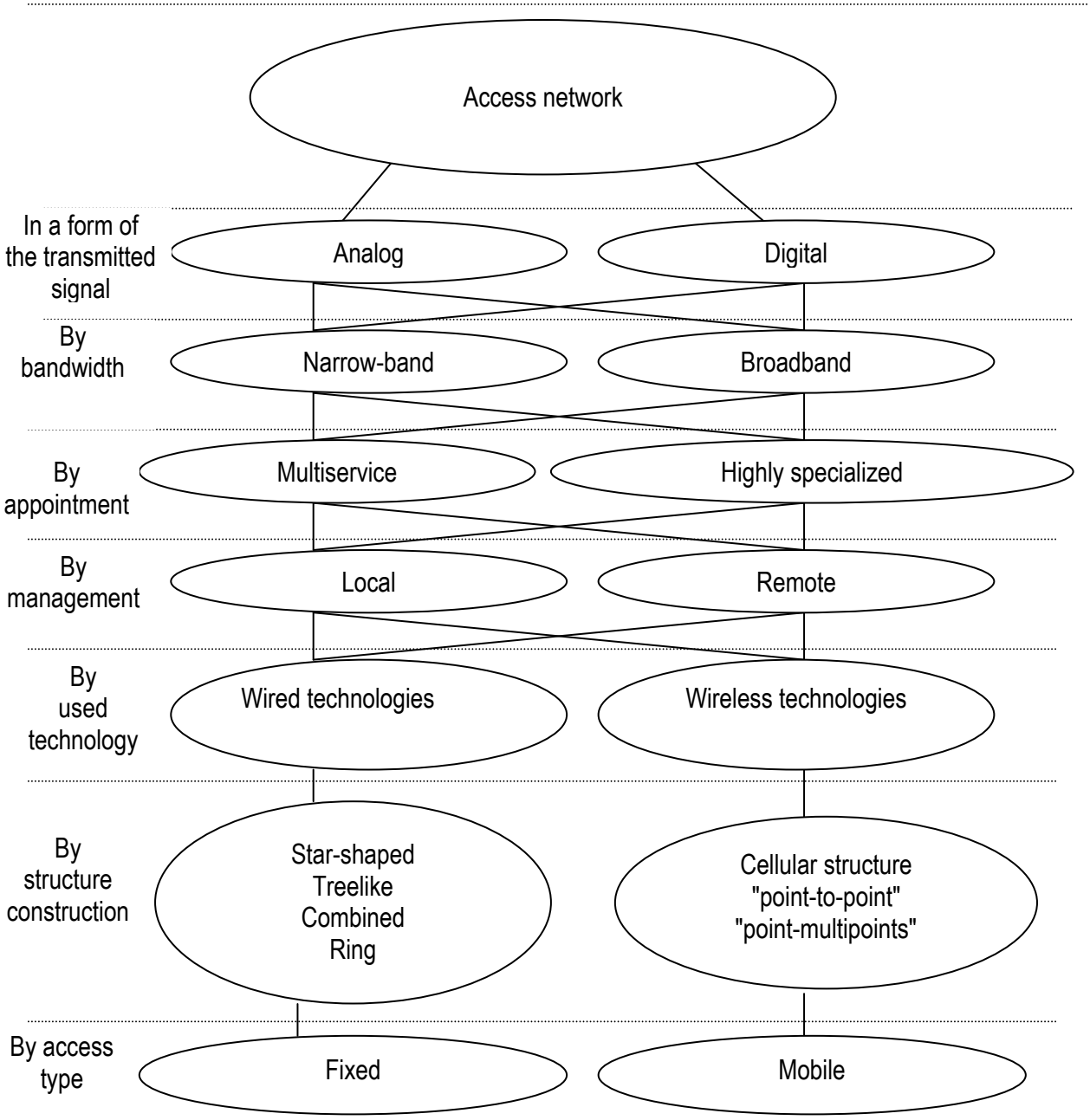


Fig.1 The structural composition of AN

1. Geographical position of the district. The territory on which creation of AN is planned, can settle down in the mountain district, on the island, near the objects influencing functioning of a network etc. The choice of the environment of transfer, structure of creation of a network depends on it.

2. Superficial population density – number of inhabitants per unit of area. Distinctions on served population density are allocated:

- serving the district with high population density;
- serving the district with average population density;
- serving the district with low population density.

High superficial population density is often observed in the central part of the city. Usually this population of multi-storey buildings, workers at the large enterprises. For service of these areas access networks with big loading ability are established.

Average superficial population density generally has the district being outside of the central part of the city. These are districts of the city are built usually up with low buildings (generally one – and five-floor constructions). However, high level of transport movement can be observed. Systems of access networks serving these areas, usually have average loading ability, quite sufficient for normal work of a network.

Low superficial population density is generally observed outside the city, these are usually rural areas. It is enough to apply access systems to service of these areas with low loading ability, and big range of communication.

3. Also important characteristic is the superficial density of users – number of users of ICS per unit of area.

In the large cities and the regional centers usually with concentration the greatest number of users of IKS being the most exacting to provided services and to the AN parameters. Rural areas with dominating number of inhabitants of aged age on the contrary differ lack of a large number of users of IKS, respectively there is no need to put large sums in creation of AN of big capacity. On parameter of superficial density of users, the district can be divided into districts with high, average and low density of users.

4. Financial possibilities of users. Creation of AN should be economic. It is not rational to put big money in the network creation which users have no financial possibility to pay all range of services provided by this network. On the contrary, in the areas which inhabitants have unlimited financial possibilities, the probability of the appeal to IKS range with the requirement of high speed of information transfer raises. Thus, by financial possibilities of users the served district can be divided into the district with high, average and low financial possibilities of users.

5. It is also necessary to consider type of the district or the area is there can be a business center of the city, the dormitory area, the cottage settlement, a private sector, a military camp, the campus, an elite housing estate, a housing estate class business, country sector, a residential suburb, a resort zone, a trading zone, the plant territory etc.

This group of parameters, aren't directly the AN parameters (parameters of the equipment of AN), but have essential influence on its structure, cost and operation process.

The grouping of the parameters of wired AN

The organization of a wire access network assumes existence of the physical environment for transfer of information, the equipment of access, transfer systems, and also the additional resources providing normal work.

Existing AN on technical parameters are grouped in the following signs: as the used environment of transfer; on versions of technologies; till speed of information transfer; as the termination of interfaces; on a way of division of channels of reception and transfer; on a modulation method in the channel; on an operating mode; on range of communication; on character of a traffic and services.

It is required to present the chosen parameters in shape, convenient for modeling. All set of parameters shares on two categories: qualitative and quantitative parameters for which ranges and gradation of accepted values are defined. Directed by a research problem we will be limited to questions of creation of AN only on the basis of wire technologies of access. An example of display of results of the analysis of parameters of the wire AN, presented in the form of a tree with software use Concept Draw Office MINDMAP.

Table 1 – The grouping of the parameters of wired AN

Systems of wire access					
By type of transmission medium used		Optical fiber			
	Copper cable				
				Coaxial cable	
According to the species of Technology	*Modems, SRM-modems, xDSL-modems*	*FTTCab FTTB/C WDM*	*FTTH*	*FTTC HFC*	*DVB*-modems *CATV*- modems
Bt information transfers	low-speed				low-speed
	medium-rate			medium-rate	
	high speed		high speed		
			very high speed		
By type of interface terminations	2-x wire		1 or 2 optical fibers	1 Coaxial cable	
	4-x wire				
By way of separating the transmit and receive channels	Frequency				
	Temporal				
	In the direction		Spectral		
According to the method of modulation in the channel	CAP, DMT, 2B1Q, HDB3 FM, RPM etc.	AM, FM, RPM etc.	AM, IM etc.	64QAM, QPSK, AM, IM, etc.	AM, FM, RPM, 64QAM, QPSK
On an operating mode	Asymmetric				
	Symmetric				
On range of communication	Small				
	Medium				Medium
		Big			
On character of a traffic	Speech				
	Data				
	Text				
	Image				
	Multimedia				

Considering that by consideration of systems of access to IKS rather large number of parameters is allocated, for convenience the chosen parameters were divided into groups. The research structure, is presented on figure 2.

In work one of ways of division of parameters on groups is offered.

1 The parameters which are not the AN parameters, but influencing modeling process. For example, financial possibilities of users, district type, its geographical position concern them etc.

2 The AN parameters being basic data for modeling. Among them are allocated:

requirements of users on which created AN, such as speed of information transfer, range of communication, factor of mistakes, delay time, a delay variation, etc. is focused;

the parameters based on requirements of users, but not being important for them, such as a modulation method in the channel, applied technology of access, a method of division of the channel of reception and transfer etc.

3 Depending on a set of requirements and parameters such parameters, as cost of services, a communication quality are formed.

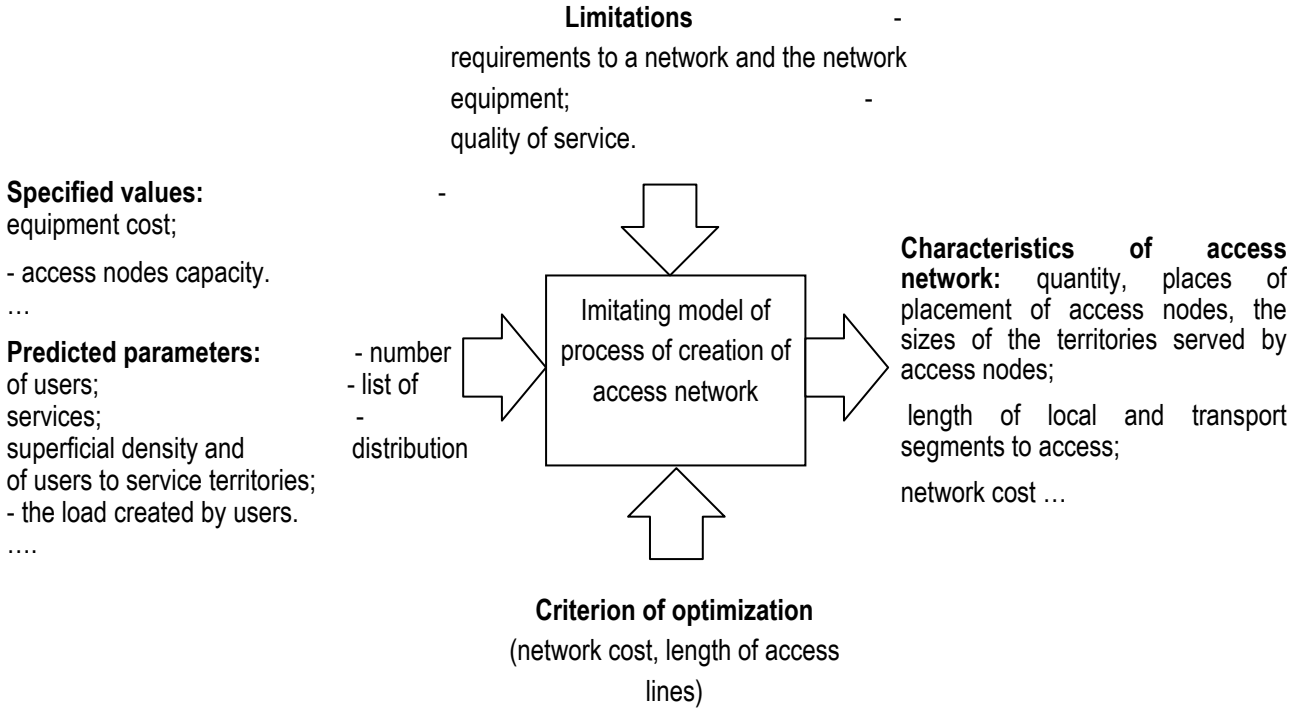


Fig. 2 Model of group of the AN parameters

Conclusion

At statement of a research problem need of the solution of a task of the analysis as a result of which the list of the parameters being initial AN for creation is defined is noted.

The analysis of initial parameters of access networks allowed to allocate what not essentially influence process of creation of a network and which account can be neglected, and the parameters being essential at creation of access networks, and influencing their structure and cost.

The characteristic of each of considered parameters is made. For distribution of parameters on their importance in the course of creation of access networks signs on which it is possible to make classification of parameters are allocated.

Considering that by consideration of systems of access rather large number of parameters is allocated, for further research the chosen parameters are divided into groups.

All set of parameters of networks of access is divided into two categories: qualitative and quantitative for which ranges and gradation of values accepted by them are defined. Results of the analysis of parameters of access networks, are presented in a tabular look and in the form of a tree with software use Concept Draw Office MINDMAP.

Results of the analysis of the AN parameters are presented by the author in works [8-11].

Acknowledgments

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

"Международный союз электросвязи (ITU)", официальное Интернет-представительство, [Электронный ресурс] http://www.itu.int. (дата обращения 12.09.2009). — Режим доступа: \www/ http://www.itu.int / — 01.03.2012 г. — Загл. с экрана.

Рекомендация ITU-T G.902 Framework Recommendation on functional access networks (AN) [Электронный ресурс] / ITU: Committed to connecting the world.— Режим доступа: \www/ http://www.itu.int / — 10.03.2012 г. — Загл. с экрана.

Гайворонская Г.С. Основные задачи модернизации сетей пользовательского доступа / Г.С Гайворонская., А.И. Котова // Зв'язок. – 2010. – №1 (89). –С. 18-24;

Гайворонская Г.С. Концепция пользовательского доступа: Учебник для ВУЗов [Текст] / Г.С Гайворонская – Одесса: ОГАХ, 2008. – 408 с.

Соколов Н. А. Сети абонентского доступа [Текст] / Н.А. Соколов – Пермь: ИПК Звезда, 1999. – 154 с.

Гайворонская Г. С. Оценка влияния некоторых факторов на процесс развития телекоммуникационных сетей / Г.С Гайворонская // Холодильна техніка і технологія. – 2006. –№2 (100). – С. 95-100.

Гайворонская Г. С. Анализ влияния вариации исходных параметров на результаты сетевого планирования / Г.С Гайворонская // Тр. УНИИРТ. – 2006. – №3 (47). – С. 102-106;

Гайворонская Г.С. Дослідження впливу помилок прогнозу вихідних даних на процес планування мереж доступу / Гайворонська Г.С., Сахарова С.В. // Збірник наукових праць ВІТІ НТУУ „КПІ". – 2010. – № 2., с 23-29.

Сахарова С.В. Задача выбора параметров сети доступа / С.В. Сахарова // Сборник тезисов Третьей международной конференции «Проблеми телекомунікацій», КПІ, Киев, 21-24 апреля 2009. – С.61.;

Гайворонская Г.С. Классификация параметров сетей доступа / Г.С. Гайворонская, С.В. Сахарова // Сборник тезисов Пятой международной НТК «Современные информационно-коммуникационные технологии», Крым, Ялта, Ливадия, 05-09 октября 2009. – С.77-78.;

Сахарова С.В. Исследование параметров сетей абонентского доступа / С.В. Сахарова // Материалы VIII МНТК «Математическое моделирование и информационные технологии» /ММИТ-2008/ - Одесса: ОГАХ. – 2008.–С. 29

Authors' information

Svetlana Sakharova *– Information technologies' institute of Odessa state academy of refrigeration, technical science's candidate, lecturer of the information-communication technologies' department; Dvoryanskaya str., 1/3, Odessa-26, 65026, Ukraine; tel. (048)-720-91-48, e-mail: switchonline@rambler.ru*

Major fields of scientific research: problems of perspective access networks' design.

HTML VALIDATION THROUGH EXTENDED VALIDATION SCHEMA

Radoslav Radev

***Abstract**: The paper presents extensible software architecture and a prototype and an implementation of a highly configurable system for HTML validation. It is based on validation rules defined in an XML document called "extended validation schema". It serves as an extended validation schema beside the official HTML specification, because the browsers' and other web clients' differences in HTML visualization makes the HTML specification insufficient and it is perfectly possible an HTML document to be syntax valid and yet not well visualized in some browser or mail-client. The extended validation schema allows definition of custom and specific validation rules in three levels - document rules, element (or tag) rules and attributes rules. The correctness of the validation schema is checked via a predefined XSD schema. The paper defines a prototype of a validation engine that consists of HTML parser, HTML validator, Storage module and Statistics module. The HTML parser parses the HTML file and breaks it into corresponding elements. The HTML validator applies the custom validations defined in the extended validation schema for every single element and attribute along with document-level validations, and also automatically corrects the errors wherever possible. The Storage module saves the validation results to a persistent storage. They can be considered for unit tests and used by the Statistics module to create additional statistics, analyses, quality assurance and bug tracking. A comparison is made with other HTML validation services and solutions. The results of an implementation of the prototype system in a software company are also presented.*

***Keywords**: HTML validation, XML schema, quality assurance, unit tests, bugs tracking.*

***ACM Classification Keywords**: D.m Software – Miscellaneous.*

Introduction

The final output of a typical real world web-application today is an HTML/CSS/Java Script code. Web browsers interpret this code and visualize it to the final user. Because of the different manner they do it, in practice it is difficult to write a HTML/CSS/Java Script code that will be visualized in a correct way. If the code is meant to be visualized not only in a browser but by an e-mail or a mobile client also, this means additional limitations and troubles. Actually the perfect HTML/CSS/Java Script code must match the intersection of the supported functionality of all the clients that are expected to visualize it.

The experienced front-end developers know good patterns for avoiding potential problems. They are not documented in a single global reference but exist only as a personal experience and knowledge. The beginners usually learn them with practice, causing a lot of errors, effort, and time. So it is good and even required for a front-end developer or team to have automatic validation functionality.

There are many HTML validation services, for example W3C Markup Validation Service [W3C Service]. They validate the HTML code against the rules of the HTML specification [HTML Specification]. They identify syntax errors in the HTML code [Chen, 2005]. But usually there are many other "unofficial" rules, defined only as a common agreement between the members of the development team. Practically the HTML specification is only the solid bases that must be followed. On top of it there are many other specific rules depending on the software project, client, task, team organization and other factors. So it is perfectly possible an HTML file to be valid

according to HTML specification and still not visualized as intended by the browsers and other clients. These specific factors in validation make it impossible to create a universal validator.

The approach suggested in this paper is to propose a configurable validation schema defined as an XML file. It is subsequently executed against the HTML code. In this way every developer or team can easily define its own specific validation rules and track their satisfaction for a period in time. The most common mistakes can be identified and corrected, and for example a specific training course for developers can be organized according the results.

Objectives

The first objective of this paper is to create a validation schema that allows developers to define specific validation rules beside the official HTML specification. For this reason three types of rules are defined:

1. Document-level rules:
   - The document must have a specified encoding.
   - The document must have a specified document type.
   - The document must contain a specific JavaScript code.
   - All special symbols in the document must be encoded.
   - The document must not contain any comments.
2. Element (tag) level rules:
   - The document must contain a meta tag with a specified content.
   - The document must not contain a meta tag with a specified content.
   - A specified tag must not be used in the HTML document.
   - A specified tag must not be empty (must have at least one child).
   - A specified start tag and its corresponding end tag must be on the same line in the code.
   - A specified tag to be treated as correct even if it does not exists in the HTML specification.
3. Attribute level rules:
   - A specified tag must contain a specified attribute.
   - A specified tag must contain a specified attribute and its value must not be empty.
   - A specified tag must contain a specified attribute and its value must match a specified value.
   - A specified tag must contain a specified attribute and it must not contain spaces around its value.
   - A specified tag must contain a specified attribute and it must not contain spaces around its equal sign.
   - A specified tag must contain a specified attribute and it must not contain spaces in its value.
   - A specified tag must contain a specified attribute and its value must match a specified regular expression.
   - A specified tag must not contain a specified attribute.
   - A specified tag must not contain any attributes except the specified ones.

The second objective of this paper is to create a prototype of a system that:

- Receives a validation schema as an XML file and the HTML document to be validated.
- Performs the validations defined in the schema against the document.
- Performs automatic correction of the errors wherever possible.

- Visualize the validation results.
- Saves the validation results to a persistent storage.
- Creates and visualizes statistics and analysis based on the validation results in the persistent storage.

Validation Schema

The validation schema proposed in this paper is an XML file. Here is a real-world example of a validation schema with full functionality:

```
<?xml version="1.0" encoding="utf-8" ?>
<ValidationManager xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xsd="http://www.w3.org/2001/XMLSchema" DocType="HTML 4.01 Transitional" Encoding="UTF-8">
 <Tags>
  <ValidationTag Tag="a">
   <Attributes>
    <ValidationAttribute Name="target" IsRequired="true" Value="_blank" />
    <ValidationAttribute Name="title" IsRequired="true" AllowSpaceAroundEqualSign="false"
AllowSpaceAroundValue="false" />
    <ValidationAttribute Name="rilt" IsRequired="true" AllowEmpty="false" AllowSpaceInValue="false"
AllowSpaceAroundEqualSign="false" AllowSpaceAroundValue="false" />
    <ValidationAttribute Name="href" AllowSpaceAroundValue="false" AllowSpaceInValue="false">
     <Expressions>
      <ValidationExpression Expression="^http:\/\/" Message="Link href does not start with http://." />
     </Expressions>
    </ValidationAttribute>
   </Attributes>
  </ValidationTag>
  <ValidationTag Tag="img">
   <Attributes>
    <ValidationAttribute Name="alt" IsRequired="true" AllowSpaceAroundEqualSign="false"
AllowSpaceAroundValue="false" />
    <ValidationAttribute Name="display" IsRequired="true" Value="block" IsStyle="true" />
   </Attributes>
  </ValidationTag>
  <ValidationTag Tag="p" IsAllowed="false" />
  <ValidationTag Tag="table">
   <Attributes>
    <ValidationAttribute Name="width" AllowEmpty="false" />
   </Attributes>
```

```
 </ValidationTag>
 <ValidationTag Tag="td">
  <Attributes>
   <ValidationAttribute Name="width" AllowEmpty="false" />
  </Attributes>
 </ValidationTag>
 <ValidationTag Tag="th">
  <Attributes>
   <ValidationAttribute Name="width" AllowEmpty="false" />
  </Attributes>
 </ValidationTag>
 <ValidationTag Tag="tr" AllowEmpty="false" />
</Tags>
<MetaTags>
 <ValidationMetaTag IsRequired="true">
  <Attributes>
   <ValidationAttribute Name="http-equiv" Value="Content-Type" />
    <ValidationAttribute Name="content">
     <Expressions>
      <ValidationExpression Expression="^text/html;(\s*)charset=utf-8;(\s*)width=device-width,(\s*)initial-
scale=0.75$" />
     </Expressions>
    </ValidationAttribute>
   <ValidationAttribute Name="name" Value="viewport" />
  </Attributes>
 </ValidationMetaTag>
</MetaTags>
 <SpecialChars ErrorMessage="Not encoded special symbol.">¡ ¢ £ ¤ ¥ ¦ § ¨ © ª « ¬ - ® ¯ ° ± ² ³ ´ µ ¶ • ¸ ¹ º » ¼
½ ¾ ¿ À Á Â Ã Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ð Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã ä å æ ç è é ê ë ì í î ï ð ñ ò ó ô
õ ö ÷ ø ù ú û ü ý þ ÿ Œ œ Š š Ÿ ƒ ˆ ˜ Α Β Γ Δ Ε Ζ Η Θ Ι Κ Λ Μ Ν Ξ Ο Π Ρ Σ Τ Υ Φ Χ Ψ Ω α β γ δ ε ζ η θ ι κ λ μ ν ξ
ο π ρ ς σ τ υ φ χ ψ ω ϑ ϒ ϖ – — ‘ ’ ‚ “ ” „ † ‡ • … ‰ ′ ″ ‹ › ‾ ⁄ € ℑ ℘ ℜ ™ ℵ ← ↑ → ↓ ↔ ↵ ⇐ ⇑
⇒ ⇓ ⇔ ∀ ∂ ∃ ∅ ∇ ∈ ∉ ∋ ∏ ∑ − ∗ √ ∝ ∞ ∠ ∧ ∨ ∩ ∪ ∫ ∴ ∼ ≅ ≈ ≠ ≡ ≤ ≥ ⊂ ⊃ ⊄ ⊆ ⊇
⊕ ⊗ ⊥ · ⌈ ⌉ ⌊ ⌋ ◊ ♠ ♣ ♥ ♦ </SpecialChars>
</ValidationManager>
```

Here is the XSD schema that any validation schema must match:

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema id="ValidationManager" xmlns:xs="http://www.w3.org/2001/XMLSchema">
 <xs:element name="ValidationManager">
  <xs:complexType>
   <xs:choice minOccurs="0" maxOccurs="unbounded">
    <xs:element name="Tags">
     <xs:complexType>
      <xs:sequence>
       <xs:element name="ValidationTag" minOccurs="0" maxOccurs="unbounded">
        <xs:complexType>
         <xs:sequence>
          <xs:element name="Attributes" minOccurs="0" maxOccurs="1">
           <xs:complexType>
            <xs:sequence>
             <xs:element name="ValidationAttribute" minOccurs="0" maxOccurs="unbounded">
              <xs:complexType>
               <xs:sequence>
                <xs:element name="Expressions" minOccurs="0" maxOccurs="1">
                 <xs:complexType>
                  <xs:sequence>
                   <xs:element name="ValidationExpression" minOccurs="0" maxOccurs="unbounded">
                    <xs:complexType>
                     <xs:attribute name="Expression" type="xs:string" use="required" />
                     <xs:attribute name="Message" type="xs:string" use="required" />
                    </xs:complexType>
                   </xs:element>
                  </xs:sequence>
                 </xs:complexType>
                </xs:element>
               </xs:sequence>
               <xs:attribute name="Name" type="xs:string" use="required" />
               <xs:attribute name="IsStyle" type="xs:boolean" />
               <xs:attribute name="IsRequired" type="xs:boolean" />
               <xs:attribute name="Value" type="xs:string" />
               <xs:attribute name="AllowEmpty" type="xs:boolean" />
               <xs:attribute name="IsAllowed" type="xs:boolean" />
```

```
          <xs:attribute name="AllowSpaceAroundEqualSign" type="xs:boolean" />
          <xs:attribute name="AllowSpaceAroundValue" type="xs:boolean" />
          <xs:attribute name="AllowSpaceInValue" type="xs:boolean" />
         </xs:complexType>
        </xs:element>
       </xs:sequence>
      </xs:complexType>
     </xs:element>
    </xs:sequence>
    <xs:attribute name="Tag" type="xs:string" use="required" />
    <xs:attribute name="IsOnTheSameLine" type="xs:boolean" />
    <xs:attribute name="AllowEmpty" type="xs:boolean" />
    <xs:attribute name="IsAllowed" type="xs:boolean" />
    <xs:attribute name="AllowOtherStyleAttributes" type="xs:boolean" />
   </xs:complexType>
  </xs:element>
 </xs:sequence>
</xs:complexType>
</xs:element>
<xs:element name="SpecialChars" nillable="true">
 <xs:complexType>
  <xs:simpleContent>
   <xs:extension base="xs:string">
    <xs:attribute name="ErrorMessage" type="xs:string" use="required" />
   </xs:extension>
  </xs:simpleContent>
 </xs:complexType>
</xs:element>
<xs:element name="Comments">
 <xs:complexType>
  <xs:attribute name="AllowComments" type="xs:boolean" />
  <xs:attribute name="ErrorMessage" type="xs:string" use="required" />
 </xs:complexType>
</xs:element>
<xs:element name="AllowedTags">
 <xs:complexType>
  <xs:sequence>
   <xs:element name="Tag" type="xs:string" minOccurs="0" />
```

```
 </xs:sequence>
</xs:complexType>
</xs:element>
<xs:element name="MetaTags">
 <xs:complexType>
  <xs:sequence>
   <xs:element name="ValidationMetaTag" minOccurs="0" maxOccurs="unbounded">
    <xs:complexType>
     <xs:sequence>
      <xs:element name="Attributes" minOccurs="1" maxOccurs="1">
       <xs:complexType>
        <xs:sequence>
         <xs:element name="ValidationAttribute" minOccurs="1" maxOccurs="unbounded">
          <xs:complexType>
           <xs:sequence>
            <xs:element name="Expressions" minOccurs="0" maxOccurs="1">
             <xs:complexType>
              <xs:sequence>
               <xs:element name="ValidationExpression" minOccurs="0" maxOccurs="unbounded">
                <xs:complexType>
                 <xs:attribute name="Expression" type="xs:string" use="required" />
                </xs:complexType>
               </xs:element>
              </xs:sequence>
             </xs:complexType>
            </xs:element>
           </xs:sequence>
           <xs:attribute name="Name" type="xs:string" use="required" />
           <xs:attribute name="Value" type="xs:string" />
          </xs:complexType>
         </xs:element>
        </xs:sequence>
       </xs:complexType>
      </xs:element>
     </xs:sequence>
     <xs:attribute name="IsAllowed" type="xs:boolean" />
     <xs:attribute name="IsRequired" type="xs:boolean" />
    </xs:complexType>
```

```
      </xs:element>
     </xs:sequence>
    </xs:complexType>
   </xs:element>
  </xs:choice>
  <xs:attribute name="DocType" type="xs:string" />
  <xs:attribute name="Encoding" type="xs:string" />
 </xs:complexType>
</xs:element>
</xs:schema>
```

Validation Engine

The program performing the validation of the HTML document against the validation schema is called validation engine. It consists of the following modules:

1. User interface to receive the HTML document to be validated with following standard possibilities:

1.1. Enter the document as text.

1.2. Upload the document as a file.

1.3. Read the document from an URI.

2. HTML parser:

2.1. Parses the HTML document.

2.2. Breaks it into corresponding parts.

3. HTML validator:

3.1. Performs the validation against the official HTML specification.

3.2. Loads the specified validation schema from an XML file.

3.2. Validates the validation schema against the XSD schema that it must match.

3.3. Iterates through the parts of the HTML document and applies the validation rules defined for any element and attribute.

3.4. If an automatic correction of some errors is possible, it is performed accordingly.

3. Storage module – it saves the errors found in a persistent storage, for example a relational database and in a log file. Automatic corrected errors are also saved to the persistent storage.

4. User interface to display the errors found, each of them consisting of:

4.1. User-friendly error message.

4.2. Line and column in the HTML code.

5. Statistics module – it displays tables, diagrams and charts, for example:

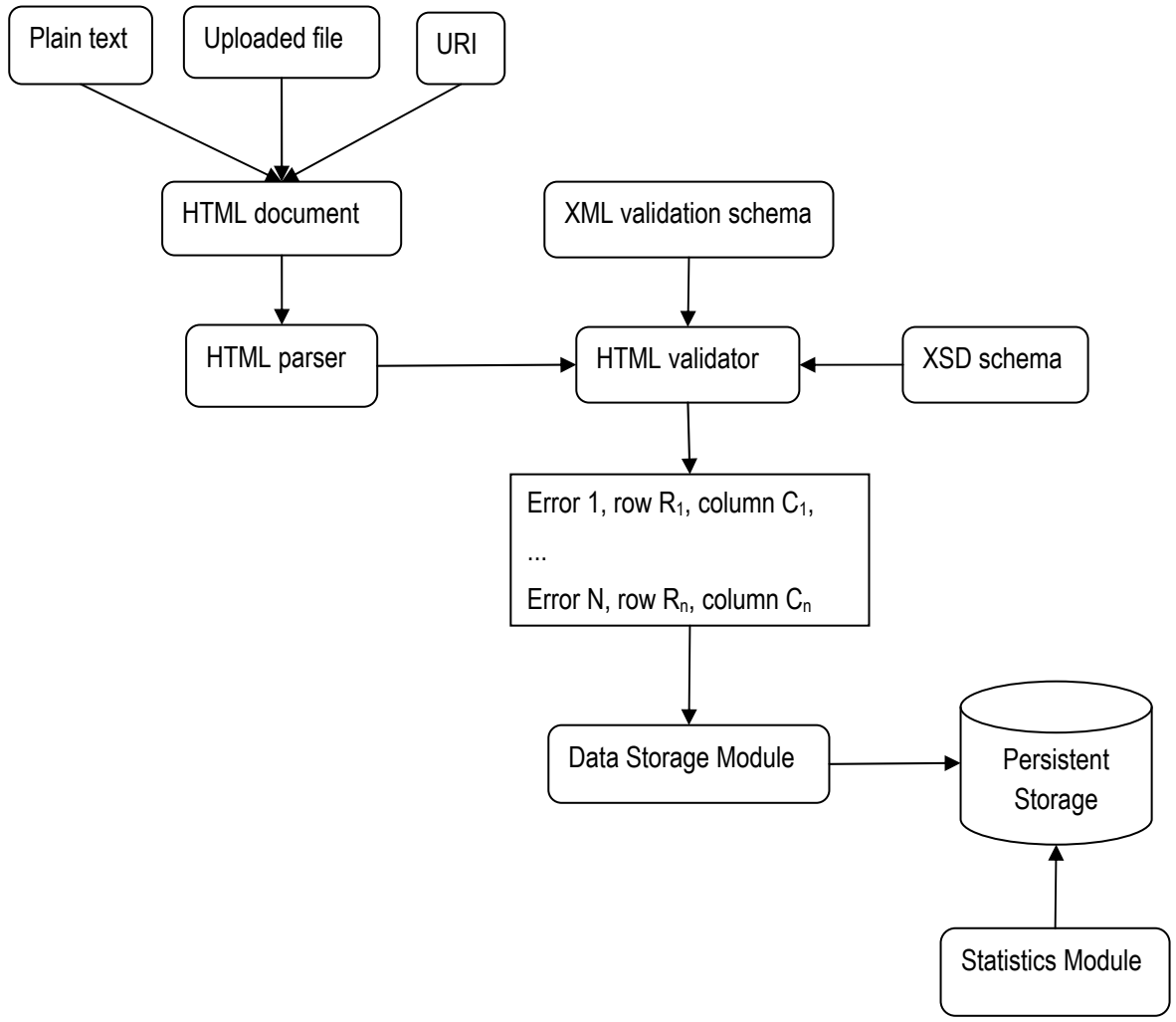
5.1. Most common validation errors (most common incorrect elements and attributes).

5.2. Average number of validations performed for a HTML document.

5.3. The trend of a specific error for a period of time – i.e. bug tracking.

All these statistics may be created for a single developer or for the team as a whole. This allows measuring of creativity and advance of the front-end developers for a period of time.

The following diagram illustrates the validation process:



Comparison with other HTML validation services

There are many HTML validation services and tools, for example:

- W3C Markup Validation Service [W3C Service]
- WDG HTML Validator [WDG]
- CSE HTML Validator [CSE]
- Validome Validation Service [Validome]

They provide syntax checking of the HTML code but do not allow custom validation rules to be added. The solution proposed in this paper seems unique in this area, although may be many companies have implemented some specific internal and not documented solutions according their needs. This paper tries to generalize the extended HTML validation and to serve as a basis for future solutions and implementations.

Conclusion

The proposed validation schema and validation process was integrated in the software company "Stanga" Ltd. in 2010 with the following technologies:

- Programming language: C# 4.0.
- User interface: ASP .NET 4.0.
- Integrated Development Environment: Visual Studio 2010.
- Database: Microsoft SQL Server 2008.

In practice it turned out to be a very useful solution with a perspective for extension in the following directions:

- More validation rules to be added.
- To create an application programming interface (API) to allow third party integration and further customization of the validation process.
- Performance measurements of the validation process and additional logging features.

A test version of the system can be found at: http://qatool.dev.provisionsofia.com/

Bibliography

[W3C Service] W3C Markup Validation Service, http://validator.w3.org/

[HTML Specification] HTML 4.01 Specification, http://www.w3.org/TR/html4/

[Chen, 2005] Chen, Sh., Hong, D., Shen, V., An Experimental Study on Validation Problems with Existing HTML Webpages, Proceedings of the 2005 International Conference on Internet Computing, June 27-30, Las Vegas, USA

[WDG] WDG HTML Validator, http://htmlhelp.com/tools/validator/

[CSE] CSE HTML Validator, http://www.htmlvalidator.com/

[Validome] Validome Validation Service, http://www.validome.org/

Acknowledgments

This paper is partially supported by Plovdiv University NPD grant NI2011-FMI-004.

Authors' Information

Radoslav Radev *– Plovdiv University, Faculty of Mathematics and Informatics, 24, Tsar Asen Str., 4000 Plovdiv, Bulgaria; e-mail: radoslav_radev@gbg.bg*

Major Fields of Scientific Research: software engineering, object-relational mapping, aspect-oriented programming.

TABLE OF CONTENT

Polynomial Approximation Using Particle SWARM Optimization of Linear Enhanced Neural Networks with no Hidden Layers
Luis F. de Mingo, Miguel A. Muriel, Nuria Gómez Blas, Daniel Triviño G. 203
Solving Diophantine Equations with a Parallel Membrane Computing Model
Alberto Arteta, Nuria Gomez, Rafael Gonzalo 220
Automated System for Quantifying the Level of Preparation in Colonoscopy
Leticia Angulo-Rodríguez, Xuexin Gao, Dobromir Filip, Christopher N. Andrews and Martin P. Mintchev 226
An In-depth Analysis and Image Quality Assessment of an Exponent-based Tone Mapping Algorithm
Chika Ofili, Stanislav Glozman, Orly Yadid-Pecht 236
Crop State and Area Estimation in Ukraine Based on Remote and In-situ Observations
Nataliia Kussul, Andrii Shelestov, Sergii Skakun, Oleksii Kravchenko, Bohdan Moloshnii 251
The Use of Time-series of Satellite Data to Flood Risk Mapping
Sergii Skakun 260
Theoretical Analysis of Empirical Relationships for Pareto-Distributed Scientometric Data
Vladimir Atanassov, Ekaterina Detcheva 271
Analysis and Justification for Selection Parameters of Wired Access Systems
Svetlana Sakharova 283
HTML Validation Through Extended Validation Schema
Radoslav Radev 290
Table of content 300