# Preface

Management at companies and institutions are facing more and more problems that demand powerful, effective, efficient, and flexible methods in their computer systems in order to support making every-day decisions. It is obvious that management is critically dependent on knowledge. Without knowledge, we are unable to handle events. However, our responses are well grounded only in isolated cases because past and future of the cases cannot be considered. Knowledge gives its respect to the past and influence to the future at our decisions.

The aim of this *Special Issue on Computation and Management* is to support advancements in an important area of enterprise activity. The other purpose of this issue is to bring advanced information technology and management closer together. A common aspect of the papers collected in this issue is the application of computation methods in management-related problem solving. The papers in this special issue are divided into two different categories on the basis of the highlighted issues: 1) *Market, commerce, and resource management related issues* (seven papers) and 2) *Knowledge management related issues* (seven papers).

*In the first category*, the paper "*A formal test of asymmetric correlation in stock market returns and the relevance of time interval of returns – a case of Eurozone stock markets*" examines the asymmetry of correlation between Eurozone's stock market returns. The asymmetry of correlation is investigated pair-wise, by estimating the exceedance correlation between returns of two stock markets at a time. The paper "*Genetically Evolved Agents for Stock Prices Prediction*" is about genetically evolving agents that act as prediction algorithms to maximize their stock prices prediction ability. A newly designed stock price prediction algorithms benchmark is used as a fitness function. The next contribution, entitled "*The IT Aspect: Towards a More Efficient System-Control and Audit of EU Funds for the Cohesion Policy*", explores how deeper exploitation of the current IT background could improve the efficiency of audits and controls of EU funds for cohesion policy. From a professional point of view, direct access to core national databases would provide auditors with additional information on systems at the Member States level. The paper "*Benefits of Smart Grid Solutions in Open Electricity Market*" highlights smart grid solutions that include different technologies to improve the efficiency of electricity distribution network operation. The benefits coming from implementation of distributed automation systems are investigated and evaluated. The contribution "*An Approach for e-Commerce On-Demand Service-Oriented Product-Line Development*" emphasizes that, due to the dynamic nature of the business domain, delivering on-demand functionalities provides high flexibility in adapting to new client requirements. The reuse capabilities of a hybridization of Service Oriented Architecture (SOA) and Software Product-Line (SPL) are explored. In the paper "*Living Labs those integrate Interactive, co-creative relations and collaboration of consumers, users and producers*" a living lab is introduced as a means to integrate users in the development process of new technologies as co-creators, bridging the roles between market pull and technology push innovation, realizing some sort of concurrent innovation. The paper "*Measurement and Simulation of Energy Use in a School Building*" presents the development of a wireless temperature monitoring system and the application of measurement data for computer model validation. The results are used to develop generalized guidelines for the determination of the efficiency of energy saving measures and the evaluation of low-energy buildings.

*In the second category*, the paper "*Managing Rational and Not-Fully-Rational Knowledge*" applies the SECI model and emphasizes that knowledge management is a range of strategies and practices in organizations to identify, create, represent, distribute, and enable the adoption of insights and experiences. In knowledge management, the elements of knowledge, insights and experiences are embedded in organizational processes and practices. The contribution "*Combining co-training with ensemble learning for application on single-view natural language datasets*" introduces a learning algorithm to exploit the advantages of co-training algorithm, while being exempt from the co-training requirement for the existence of adequate feature split in the dataset. The paper "*Automatic recognition of features in spectrograms based on some image analysis methods*" presents the automatic localization, extraction, analysis, and comparison/classification of the features in signals and their spectra. The method is suitable for a variety of algorithms focused on visual data processing in applications such as the analysis of biological, acoustic, sonar, and radar signals. In the paper "*Information and Knowledge Retrieval within Software Projects and their Graphical Representation for Collaborative Programming*", a method is proposed for information and knowledge mining in the source code of medium and large enterprise projects. Developers within and outside a team can receive and utilize visualized information from the code and apply them to their projects. The contribution "*Simulation tools evaluation using theoretical manufacturing model*" confronts average values of performance indicators of the manufacturing simulation model in three well-known simulation tools. It applies inferential statistical technique after normality test and homogeneity of variances to analyze the output data of model in the different simulation tools. The aim of the paper *"Investigating the Influence of Knowledge Management Practices on Organizational Performance: An Empirical Study*" is to investigate the influence of knowledge management practices on organizational performance in small and medium enterprises (SMEs) using structural equation modeling (SEM). The paper "*Active Knowledge for Situation-driven Control of Product Definition*" introduces a new method for the application of active knowledge in the adaptive modification of product model features in the case of a changed situation or event. The objective is to achieve a higher level and more advanced active feature driven product model definition in industrial product lifecycle management (PLM) systems.

The guest editors would like to thank all the authors for their valuable contributions. They would also like to thank the editors of the Acta Polytechnica Hungarica for approving this special issue.

*Jozef Kelemen, László Horváth*

Special Issue Guest Editors

# A Formal Test of Asymmetric Correlation in Stock Market Returns and the Relevance of Time Interval of Returns – a Case of Eurozone Stock Markets

**Silvo Dajčman**

University of Maribor, Faculty of Economics and Business, Razlagova 14, 2000 Maribor, Slovenia, e-mail: silvo.dajcman@uni-mb.si

*Abstract: The paper examines the asymmetry of correlation between the Eurozone's stock market returns. The asymmetry of correlation is investigated pair-wise, by estimating the exceedance correlation between returns of two stock markets at a time. As markets can be very volatile, especially in crisis periods, and because there are investors with different investment horizons, we investigate whether the results are sensitive to time interval of stock market returns. We found that the results of the exceedance correlation estimates and the asymmetric correlation test do depend on the time interval of returns. When longer time interval returns (20-day moving average returns) are used, the Eurozone stock markets' returns' dynamics are more (pair-wise) correlated in the falling markets than in the up markets, while for daily returns, the correlations in the up markets are higher for most of the investigated Eurozone's stock indices pairs. An important implication of the results of the paper is that investors in stock markets should investigate the exceedance correlations and asymmetry of correlation for those return intervals (daily, weekly, monthly, etc.) that correspond to their investment horizon.*

*Keywords: stock markets; asymmetric correlation; Eurozone*

# 1 Introduction

The study of asymmetric dependence (or asymmetric correlation) is important from the point of view of optimal portfolio allocation and risk management. Since the seminal works of [12] it is recognized that (international) diversification reduces a portfolio's total risk due to non-perfect positive co-movement between returns of the portfolio assets. However, from an investment perspective, diversification (low correlation between assets in a portfolio) is only desirable in falling (or ″down″) markets, whereas in rising (or ″up″) markets it is undesirable, and thus investors prefer assets that are highly correlated.

The benefits of diversification in stock markets have been questioned by a number of empirical studies (e.g. [11], [2], [5], [9], [13]), showing that correlations between returns of portfolios of stocks or between returns of stock indices are higher in down markets than in up markets[1]. The existence of asymmetric correlation implies that the practical benefits of diversification are substantially reduced for investors in stock markets. In [9], the authors argue that gaining information on asymmetric correlation can add substantial economic value to investors. The knowledge of asymmetric correlation in stock market returns is also important for supervisory authorities because of their implications for the stability of financial markets, and for the central banks in conducting monetary policy ([6], [3]).

The existent empirical studies on analysis of time-varying and potentially asymmetric dependence structures in stock markets predominately apply the multivariate GARCH models ([5], [13], [8]), copula functions ([10]) and the recently developed tests of asymmetric dependence ([11], [2], [9]). The measurement of asymmetry of correlation is not an easy task. In [4] and [7] it is noted that calculating correlations conditional on high or low returns, or high or low volatility, induces a conditioning bias in the correlation estimates. The recently developed tests of asymmetric dependence of [11], [2] and [9] resolve this issue as they measure correlation asymmetry by looking at behavior in the tails of the return distribution.

The novel test of correlation asymmetry of [9] has gained a lot of popularity due to its appealing features: it is a model-free test, so it can be used without having to specify a statistical model for the data. Unlike many of asymmetry tests that impose normality assumption on the data, the test allows for general distributional assumptions. The test statistic is also easy to implement and its asymptotic distribution follows a standard chi-square distribution under the null hypothesis of symmetry.

This paper provides the most recent evidence of asymmetry of correlation between Eurozone stock market returns for the period between December 3, 2003 and January 27, 2012, applying the test of Hong et al. (2007). Asymmetry of correlation is investigated pair-wise, by estimating the exceedance correlations between two stock market returns at a time. [11], [2], and [9] use monthly returns in their tests of asymmetry of correlation. Many investors in stock markets, however, have investment horizons that are shorter than one month. In this paper we examine whether the time interval of returns impacts the estimates of tests of asymmetry of correlation of [9] by comparing the results for 20 trading days (approximately one calendar month) moving average returns and daily returns.

---

1 For a review of factors that can influence the strength of comovement during up and down markets see [1].

# 2 Methodology

The asymmetry of correlation between stock market returns is investigated by estimating the exceedance correlations between two time series of stock market returns at a time. According to [11], exceedance is defined as a stock market return of a particular country that is above or below a certain exceedance level, $c$. Let $\{r_{1t}, r_{2t}\}$ be the returns of two stock markets (represented by two stock market indices) in period $t$. The exceedance correlation between the stock market returns at an exceedance level $c$ is the correlation between the returns series, on condition that the returns of both stock indices are either at least $c$ standard deviations above or at least $c$ standard deviations below the mean value of the respective stock market returns:

$$\rho^+(c) = corr(r_{1t}, r_{2t} | r_{1t} > c, r_{2t} > c), \quad (1a)$$

$$\rho^-(c) = corr(r_{1t}, r_{2t} | r_{1t} < -c, r_{2t} < -c), \quad (1b)$$

where the returns are standardized to have zero mean and unit variance so that the mean and variance do not appear explicitly in the right-hand side of the equations (1a) and (1b).

Following [11], [2], and [9], we test whether the correlation between the upper-tail returns (i.e., returns at least $c$ standard deviations above the mean return) and the lower-tail returns (i.e., returns at least $c$ standard deviations below the mean return) of the two stock markets are the same. Thus, the null hypothesis of symmetric correlation is

$$H_0 : \rho^+(c) = \rho^-(c), \quad (2)$$

for all $c \geq 0$.

In [9], the authors provide a model free test for the null of symmetric exceedance correlations. They show that for the $m$ chosen number of different exceedance levels, the vector

$$\rho^+(c) - \rho^-(c) = [\hat{\rho}^+(c_1) - \hat{\rho}^-(c_1), \ldots, \hat{\rho}^+(c_m) - \hat{\rho}^-(c_m)], \quad (3)$$

after being scaled by $\sqrt{T}$ ($T$ is the size of a random sample of the returns series), under the null hypothesis of symmetry has an asymptotic normal distribution with zero mean and positive definite variance-covariance matrix, $\Omega$, for all possible true distributions of the data.

Under condition that $r_{1t}$ and $r_{2t}$ are larger than $c$ simultaneously, the sample means and variance of these conditional time series are computed

$$\hat{\mu}_1^+(c) = \frac{1}{T_c^+} \sum_{t=1}^{T} r_{1t} \Big| (r_{1t} > c, r_{2t} > c) , \tag{4a}$$

$$\hat{\mu}_2^+(c) = \frac{1}{T_c^+} \sum_{t=1}^{T} r_{2t} \Big| (r_{1t} > c, r_{2t} > c) , \tag{4b}$$

$$\hat{\sigma}_1^+(c)^2 = \frac{1}{T_c^+ - 1} \sum_{t=1}^{T} [r_{1t} - \hat{\mu}_1^+(c)]^2 \Big| (r_{1t} > c, r_{2t} > c) , \tag{4c}$$

$$\hat{\sigma}_2^+(c)^2 = \frac{1}{T_c^+ - 1} \sum_{t=1}^{T} [r_{2t} - \hat{\mu}_2^+(c)]^2 \Big| (r_{1t} > c, r_{2t} > c) , \tag{4d}$$

where $\hat{\mu}_1^+(c)$ and $\hat{\mu}_2^+(c)$ are the estimated conditional means of the series and $\hat{\sigma}_1^+(c)^2$ and $\hat{\sigma}_2^+(c)^2$ the estimated conditional variance of the series.

The sample conditional correlation $\hat{\rho}^+(c)$ is then given by:

$$\hat{\rho}^+(c) = \frac{1}{T_c^+ - 1} \sum_{t=1}^{T} [\hat{X}_{1t}^+(c) \hat{X}_{2t}^+(c) \Big| (r_{1t} > c, r_{2t} > c) , \tag{5}$$

where $\hat{X}_{1t}^+(c) = \frac{r_{1t} - \hat{\mu}_1^+(c)}{\hat{\sigma}_1^+(c)}$, and $\hat{X}_{2t}^+(c) = \frac{r_{2t} - \hat{\mu}_2^+(c)}{\hat{\sigma}_2^+(c)}$. The same computations apply also for $\hat{\rho}^-(c)$.

The authors of [9] prove that the null hypothesis of symmetric correlation can be tested by the $J_p$ statistics which under the null hypothesis and under certain regularity conditions is asymptotically chi-square distributed with $m$ degrees of freedom

$$J_p = T(\hat{\rho}^+(c) - \hat{\rho}^-(c)) \hat{\Omega}^{-1} (\hat{\rho}^+(c) - \hat{\rho}^-(c)) \sim \chi_m^2 , \tag{6}$$

where $\hat{\rho}^+(c) - \hat{\rho}^-(c)$ is defined by equation (3) and $\hat{\Omega}$ is consistent estimate of the asymptotic covariance matrix of $\hat{\rho}^+(c) - \hat{\rho}^-(c)$. The variance-covariance matrix is given by

$$\hat{\Omega} = \sum_{1-T}^{T-1} k(\frac{1}{p}) \hat{\gamma}_1 , \tag{7}$$

where $\hat{\gamma}_1$ is an $N \times N$ matrix with $(i, j)$-th element

$$\hat{\gamma}_1(c_i c_j) = \frac{1}{T}\sum_{t=|l|+1}^{T} \hat{\varsigma}_t(c_i)\hat{\varsigma}_{t-|l|}(c_j), \tag{8}$$

and

$$\hat{\varsigma}_t(c) = \frac{1}{T_c^+}[\hat{X}_{1t}^+(c)\hat{X}_{2t}^+(c) - \hat{\rho}^+(c)]\Big|(r_{1t} > c, r_{2t} > c) - \frac{1}{T_c^-}[\hat{X}_{1t}^-(c)\hat{X}_{2t}^-(c) - \hat{\rho}^-(c)]\Big|(r_{1t} < -c, r_{2t} < -c), \tag{9}$$

where $k(\cdot)$ is a Bartlett kernel function that assigns a suitable weight to each lag of order $l$; $p$ is the smoothing parameter or lag truncation order.

# 3 Data and Empirical Results

The symmetry of correlation between returns of Eurozone's stock markets, listed in Table 1, is analyzed for the period from December 3, 2003 to January 27, 2012. The main stock indices returns were used as proxies for stock market returns of particular countries. The returns were calculated as the differences in the logarithms of the daily closing prices of indices ( $\ln(P_t) - \ln(P_{t-1})$, where $P$ is an index value). The stock indices included are: the ATX (for Austria), CAC40 (for France), DAX (for Germany), FTSE100 (for the U.K.), FTSEMIB (for Italy), ISEQ (for Ireland), and IBEX35 (for Spain). Days with no trading in any of the observed market were left out. The data for stock indices is Yahoo! Finance. Table 1 presents descriptive statistics of the data.

Table 1

Descriptive statistics of stock indices returns

| | Min | Max | Mean | Std. deviation | Skewness | Kurtosis | Jarque-Bera statistics |
|---|---|---|---|---|---|---|---|
| ATX | -0.1637 | 0.1304 | 0.00018 | 0.0187 | -0.2891 | 11.4275 | 5705.57*** |
| CAC40 | -0.0947 | 0.1059 | -0.00003 | 0.0158 | 0.1642 | 10.4452 | 4440.75*** |
| DAX | -0.0743 | 0.1080 | 0.00028 | 0.0153 | 0.1162 | 9.4645 | 3345.76*** |
| FTSE100 | -0.0927 | 0.1079 | 0.00014 | 0.0135 | 0.1914 | 12.3669 | 7027.12*** |
| ISEQ | -0.1396 | 0.0973 | -0.00024 | 0.0173 | -0.5573 | 9.8403 | 3840.51*** |
| FTSEMIB | -0.0997 | 0.1087 | -0.00028 | 0.0162 | -0.1569 | 9.6681 | 3563.07*** |
| IBEX35 | -0.116 | 0.1348 | -0.00008 | 0.0160 | 0.0099 | 12.0721 | 6580.83*** |

*Notes: The Jarque-Bera statistics: *** indicate that the null hypothesis (of normal distribution) is rejected at a 1% significance level.*

All series display significant leptokurtic behavior as evidenced by the large kurtosis with respect to the Gaussian distribution. The Jarque-Bera test rejects the hypothesis of normally distributed time series. We also tested for the stationarity of time series by the Augmented Dickey-Fuller (ADF) test, the Phillips-Perron (PP) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, and the results led to conclusion of no unit root in returns series[2].

Table 2 reports pair-wise Pearson's correlation coefficients for the returns of stock indices. Notably, the greatest correlation is observed between the DAX-CAC40, CAC40-FTSE100 and CAC40-FTSEMIB, while the lowest is between ISEQ and IBEX35.

Table 2
Pearson's correlation of Eurozone stock market returns

| | ATX (Austria) | CAC40 (France) | DAX (Germany) | FTSE100 (U.K.) | ISEQ (Ireland) | FTSEMIB (Italy) | IBEX35 (Spain) |
|---|---|---|---|---|---|---|---|
| ATX | 1 | | | | | | |
| CAC40 | 0.8187 | 1 | | | | | |
| DAX | 0.7867 | 0.9332 | 1 | | | | |
| FTSE100 | 0.7994 | 0.9262 | 0.8717 | 1 | | | |
| ISEQ | 0.7179 | 0.7619 | 0.7093 | 0.7528 | 1 | | |
| FTSEMIB | 0.7933 | 0.9171 | 0.8633 | 0.8395 | 0.7097 | 1 | |
| IBEX35 | 0.7819 | 0.8926 | 0.8350 | 0.8230 | 0.6900 | 0.8880 | 1 |

*Notes: All the correlation coefficients are significantly different from zero.*

Next, we report the estimates of the exceedance correlation and the results of the test of correlation symmetry for the particular pair-wise investigated stock markets for 20-day moving average returns, computed at singleton exceedance level $c = \{0\}$ (Table 2). The second column of Table 3 reports the p-value of the $J_p$ statistics and the third column the difference between the upper-tail and lower-tail correlation, $\hat{\rho}^{+}(c) - \hat{\rho}^{-}(c)$. Following [8] the test of correlation asymmetry is computed also for a set of exceedance levels, $c = \{0, 0.5, 1, 1.5\}$. The results for the set of exceedance levels and the corresponding p-values are given in columns four through eight.

Notably, for a singleton exceedance level, $c = \{0\}$, as well as for a set of excedance levels, $c = \{0, 0.5, 1, 1.5\}$, the difference between the upper-tail and the lower-tail correlations, $\hat{\rho}^{+}(c) - \hat{\rho}^{-}(c)$, is negative. Stock markets are thus more correlated in down than in up markets, which is in accordance with the findings of

2 The results are not presented here, but can be obtained from the author.

[11], [2], and [9]. The test of correlation asymmetry, however, does not significantly (at 5% significance level) reject the null hypothesis of symmetry of the upper- and lower-tail correlations for any observed stock indices pairs.

Table 3

Results of the correlation asymmetry test for 20-day simple moving average returns of stock indices

| Stock market indices | $c=\{0\}$ | | $c=\{0,0.5,1,1.5\}$ | | | | |
|---|---|---|---|---|---|---|---|
| | p-value | $\hat{\rho}^{+}(c)-\hat{\rho}^{-}(c)$ | p-value | $\hat{\rho}^{+}(c)-\hat{\rho}^{-}(c)$; $c=0$ | $\hat{\rho}^{+}(c)-\hat{\rho}^{-}(c)$; $c=0.5$ | $\hat{\rho}^{+}(c)-\hat{\rho}^{-}(c)$; $c=1$ | $\hat{\rho}^{+}(c)-\hat{\rho}^{-}(c)$; $c=1.5$ |
| ATX-CAC40 | 0.2073 | -0.2924 | 0.4299 | -0.2924 | -0.2036 | -0.2014 | -0.3531 |
| ATX-DAX | 0.2934 | -0.2638 | 0.5247 | -0.2638 | -0.1732 | -0.1471 | -0.2940 |
| ATX-FTSE100 | 0.1872 | -0.3479 | 0.3648 | -0.3479 | -0.3266 | -0.4320 | -0.1359 |
| ATX-FTSEMIB | 0.4659 | -0.1978 | 0.1408 | -0.1978 | 0.0001 | 0.1827 | 0.3955 |
| ATX-ISEQ | 0.4755 | -0.1954 | 0.8696 | -0.1954 | -0.2167 | -0.2347 | -0.0128 |
| ATX-IBEX35 | 0.3866 | -0.1884 | 0.4262 | -0.1884 | -0.0761 | 0.2006 | 0.2253 |
| CAC40-DAX | 0.7461 | -0.0743 | 0.9954 | -0.0743 | -0.0723 | -0.0859 | -0.0850 |
| CAC40-FTSE100 | 0.6231 | -0.1107 | 0.9479 | -0.1107 | -0.1054 | -0.1706 | -0.1874 |
| CAC40-ISEQ | 0.5537 | -0.1247 | 0.5050 | -0.1247 | -0.0711 | -0.2439 | -0.3445 |
| CAC40-FTSEMIB | 0.6509 | -0.1037 | 0.7332 | -0.1037 | -0.1155 | -0.2291 | -0.3686 |
| CAC40-IBEX35 | 0.4144 | -0.1484 | 0.6594 | -0.1484 | -0.1283 | -0.0196 | -0.1419 |
| DAX-FTSE100 | 0.4321 | -0.1879 | 0.7667 | -0.1879 | -0.1894 | -0.2837 | -0.3368 |
| DAX-ISEQ | 0.7764 | -0.0637 | 0.5081 | -0.0637 | -0.0459 | -0.1836 | 0.1378 |
| DAX-FTSEMIB | 0.7285 | -0.0902 | 0.5067 | -0.0902 | -0.0204 | -0.0575 | -0.2808 |
| DAX-IBEX35 | 0.6623 | -0.0882 | 0.5015 | -0.0882 | 0.0650 | 0.1717 | 0.0042 |
| FTSE100-FTSEMIB | 0.3275 | -0.2299 | 0.5743 | -0.2299 | -0.2121 | -0.3097 | -0.1834 |
| FTSE100-ISEQ | 0.2692 | -0.2768 | 0.4039 | -0.2768 | -0.2855 | -0.4956 | -0.5434 |
| FTSE100-IBEX35 | 0.4629 | -0.1327 | 0.6741 | -0.1327 | -0.0328 | -0.0779 | 0.0662 |
| ISEQ-FTSEMIB | 0.9093 | -0.0281 | 0.2965 | -0.0281 | 0.1179 | -0.0116 | 0.1391 |
| ISEQ-IBEX35 | 0.6658 | -0.0814 | 0.4716 | -0.0814 | 0.0906 | 0.2512 | -0.0377 |
| FTSEMIB-IBEX35 | 0.7441 | -0.0772 | 0.9300 | -0.0772 | 0.0025 | 0.0754 | 0.1302 |

*Notes: Two sets of exceedance levels were used to perform the test. The first is the singleton $c=\{0\}$ and the second is $c=\{0,0.5,1,1.5\}$. $\hat{\rho}^{+}(c)-\hat{\rho}^{-}(c)$ represents the difference in correlation for the exceedance level $c$. P-value is a significance level of rejecting the null hypothesis of symmetric correlation. The lower the p-value, the more asymmetric are the lower- and the upper-tail correlations.*

As noted by [9], a large difference of correlation of returns between up and down markets does not mean that there is necessarily a genuine difference in the population parameters. There are always differences in the sample estimates simple due to sample variations. We can observe that for some stock indices pairs $\hat{\rho}^{+}(c)-\hat{\rho}^{-}(c)$ gets more negative when exceedance level increases (for instance CAC40-FTSE100, DAX-FTSE100 or FTSE100-ISEQ), whereas for some others the difference between the upper- and lower-tail correlation reduces or even becomes positive (for instance ATX-FTSEMIB, ATX-IBEX35, FTSEMIB-IBEX35).

This striking difference in exceedance correlations between stock markets has important implications for investors in investigated stock markets. As already noted, from the investment perspective, a low correlation between asset returns is only desirable in falling markets, whereas in the up markets, a high correlation is desirable. The investors that internationally diversify their portfolios in stock markets for which $\hat{\rho}^{+}(c)-\hat{\rho}^{-}(c)$ is positive and increases with exceedance level would therefore be better off than those who invest in stock markets for which $\hat{\rho}^{+}(c)-\hat{\rho}^{-}(c)$ is negative and becomes even more negative when the exceedance level increases.

Let us now take into account that stock markets can be very volatile, especially in crisis periods. For investors with shorter investment horizons (shorter than 20 trading days) it is therefore important to know whether the time interval of returns impacts the estimates of exceedance correlation and correlation asymmetry test. For this purpose, the exceedance correlations and test of asymmetric correlation of [9] were recomputed for daily returns.

As the results in Table 4 show, for a singleton exceedance level, $c=\{0\}$ , there are now 5 out of 21 investigated stock market pairs for which the lower-tail correlation (i.e. for returns below the mean return for each the returns series) is higher than the upper-tail correlation: ATX-ISEQ, CAC40-DAX, CAC40-DAX, DAX-IBEX35, and FTSE100-ISEQ. However, for a set of exceedance levels $c=\{0,0.5,1,1.5\}$ we find that as the exceedance level is increased, the upper-tail correlation normally exceeds the lower-tail correlation, and at the exceedance level $c=\{1.5\}$ the upper-tail correlation is higher than lower-tail correlation for all investigated stock market pairs. For a singleton exceedance level, as well as for the set of exceedance levels, the null hypothesis of symmetric correlation, however, still cannot be rejected for any of investigated stock markets pair.

Comparing the results of Tables 3 and 4, evidently the results of exceedance correlation estimates are very much dependent on the time interval of returns. Whereas for the 20-day moving average returns, the upper-tail correlations are mostly smaller than lower-tail correlation, for daily returns the upper-tail correlations are higher than the lower-tail correlations for most of investigated stock indices pairs (or all indices pairs at exceedance level $c=1.5$). Another implication for investors in the investigated stock markets is therefore to investigate exceedance correlations and asymmetry of correlation for those returns intervals (daily, weekly, monthly, etc.) that correspond to their investment horizons.

Table 4

Results of the correlation asymmetry test for daily returns of stock indices

| Stock market indices | $c = \{0\}$ | | $c = \{0, 0.5, 1, 1.5\}$ | | | | |
|---|---|---|---|---|---|---|---|
| | p-value | $\hat{\rho}^{+}(c) - \hat{\rho}^{-}(c)$ | p-value | $\hat{\rho}^{+}(c) - \hat{\rho}^{-}(c)$; $c = 0$ | $\hat{\rho}^{+}(c) - \hat{\rho}^{-}(c)$; $c = 0.5$ | $\hat{\rho}^{+}(c) - \hat{\rho}^{-}(c)$; $c = 1$ | $\hat{\rho}^{+}(c) - \hat{\rho}^{-}(c)$; $c = 1.5$ |
| ATX-CAC40 | 0.6306 | 0.0619 | 0.7039 | 0.0619 | 0.1106 | 0.2168 | 0.3106 |
| ATX-DAX | 0.5106 | 0.0766 | 0.4122 | 0.0766 | 0.1904 | 0.2569 | 0.3491 |
| ATX-FTSE100 | 0.7974 | 0.0344 | 0.6866 | 0.0344 | 0.1151 | 0.2375 | 0.3208 |
| ATX-FTSEMIB | 0.5358 | 0.0767 | 0.8903 | 0.0767 | 0.1276 | 0.2025 | 0.1723 |
| ATX-ISEQ | 0.6364 | -0.0635 | 0.0779 | -0.0635 | 0.0362 | 0.0071 | 0.2445 |
| ATX-IBEX35 | 0.8477 | 0.0250 | 0.5700 | 0.0250 | 0.0530 | 0.2076 | 0.3122 |
| CAC40-DAX | 0.9919 | -0.0014 | 0.9999 | -0.0014 | 0.0085 | 0.0190 | 0.0262 |
| CAC40-FTSE100 | 0.9118 | 0.0177 | 0.9991 | 0.0177 | 0.0344 | 0.0629 | 0.0709 |
| CAC40-ISEQ | 0.9822 | -0.0033 | 0.5144 | -0.0033 | 0.0449 | -0.0349 | 0.0979 |
| CAC40-FTSEMIB | 0.8868 | 0.0187 | 0.9777 | 0.0187 | 0.0483 | 0.0885 | 0.1420 |
| CAC40-IBEX35 | 0.9201 | 0.0132 | 0.8917 | 0.0132 | 0.0340 | 0.0706 | 0.1795 |
| DAX-FTSE100 | 0.9816 | 0.0034 | 0.9946 | 0.0034 | 0.0280 | 0.0625 | 0.0898 |
| DAX-ISEQ | 0.8467 | 0.0254 | 0.0586 | 0.0254 | 0.0887 | -0.1046 | 0.1005 |
| DAX-FTSEMIB | 0.8390 | 0.0241 | 0.9919 | 0.0241 | 0.0536 | 0.0860 | 0.1002 |
| DAX-IBEX35 | 0.9860 | -0.0020 | 0.8721 | -0.0020 | 0.0265 | 0.1177 | 0.1235 |
| FTSE100-FTSEMIB | 0.7304 | 0.0459 | 0.9030 | 0.0459 | 0.0841 | 0.1473 | 0.2024 |
| FTSE100-ISEQ | 0.9542 | -0.0089 | 0.1047 | -0.0089 | 0.0342 | -0.1126 | 0.1168 |
| FTSE100-IBEX35 | 0.8241 | 0.0291 | 0.8704 | 0.0291 | 0.0466 | 0.1154 | 0.1898 |
| ISEQ-FTSEMIB | 0.8115 | 0.0307 | 0.4956 | 0.0307 | 0.0862 | 0.0727 | 0.2560 |
| ISEQ-IBEX35 | 0.9592 | 0.0064 | 0.4934 | 0.0064 | 0.1024 | 0.1175 | 0.2512 |
| FTSEMIB-IBEX35 | 0.8245 | 0.0288 | 0.9574 | 0.0288 | 0.0642 | 0.1369 | 0.2098 |

*Notes: Two sets of exceedance levels were used to perform the test. The first is the singleton $c = \{0\}$ and the second is $c = \{0, 0.5, 1, 1.5\}$. $\hat{\rho}^{+}(c) - \hat{\rho}^{-}(c)$ represents the difference in correlation for the exceedance level $c$. P-value is a significance level of rejecting the null hypothesis of symmetric correlation. The lower the p-value, the more asymmetric are the lower- and the upper-tail correlations.*

## Conclusions

This paper provides the most recent evidence of the asymmetry of correlation between Eurozone stock market returns for the period between December 3, 2003 and January 27, 2012, applying the test of [9]. The asymmetry of correlation was investigated pair-wise, by estimating the exceedance correlations between two stock market returns at a time. In order to check whether the results are sensitive to time interval of stock market returns, computations were performed for 20-day moving average returns and for daily returns.

We found that when 20-day moving average returns are used, the Eurozone stock markets' returns dynamics are more (pair-wise) correlated in falling markets than in up markets, whereas the opposite holds true for daily returns. This, however,did not impact the results of the test of correlation asymmetry for the data of our sample.

The results of the paper have important implications for investors in the investigated stock markets. From an investment perspective, low correlation between asset returns is only desirable in falling markets, whereas in up markets it is undesirable, and thus investors prefer assets that are highly correlated. The investors that internationally diversify their portfolios in stock markets for which the difference between correlation in up and down markets is positive, and increases with exceedance level, would therefore be better off than those who invest in stock market for which the difference in the upper- and lower-tail correlation is negative and becomes even more negative when the exceedance level increases. Investors in the investigated stock markets should investigate the exceedance correlations and the asymmetry of correlation for those returns intervals (daily, weekly, monthly, etc.) that correspond to their investment horizons.

**References**

[1] Andjelic, G. and Djakovic, V. (2012) 'Financial Market Co-Movement between Transition Economies: A Case Study of Serbia, Hungary, Croatia, and Slovenia', *Acta Polytehnica Hungarica*, 9(3), pp. 115-134

[2] Ang. A. and Chen, J. (2002) 'Asymmetric Correlations of Equity Portfolios', *Journal of Financial Economics*, 63(3), pp. 443-94

[3] Berben, R-P. and Jansen W. J. (2005) 'Bond Market and Stock Market Integration in Europe', *Working Paper, No. 60*, DNB – De Nederlandsche Bank

[4] Boyer, B. H., Gibson, M. S. and Loretan, M. (1999) 'Pitfalls in Tests for Changes in Correlations', *International Finance Discussion Paper, No. 597*. Board of the Federal Reserve System, Washington DC

[5] Cappiello, L., Engle, F. R. and Sheppard, K. (2003) 'Asymmetric Dynamics in the Correlation of Global Equity and Bond Returns', *Working Paper No. 204*, European Central Bank

[6] Clare, A. and Lekkos, I. (2000) 'Decomposing the Relationship between International Bond Markets', *Working Paper No. 123*, Bank of England

[7] Forbes, K. and Rigobon, R. (2002) 'No Contagion Only Interdependence: Measuring Stock Market Comovements', *Journal of Finance*, 57(5), pp. 2223-61

[8] Gang, J. and Xiang, L. (2011) 'Risk Perception and Equity Returns: Evidence from the SPX and VIX', *Bulletin of Economic Research*, Online first, DOI: 10.1111/j.1467-8586.2011.00409.x

[9] Hong, Y., Tu, J. and Zhou, G. (2007) 'Asymmetries in Stock Returns: Statistical Tests and Economic Evaluation', *Review of Financial Studies*, 20(5), pp. 1547-81

[10] Kang, B. U., In, F., Kim, G. and Kim, T. S. (2010) 'A Longer Look at the Asymmetric Dependence between Hedge Funds and the Equity Market', *Journal of Financial and Quantitative Analysis*, 45(3), pp. 763-89

[11] Longin, F. and Solnik, B. (2001) 'Extreme Correlation of International Equity Markets', *Journal of Finance*, 56(2), pp. 649-76

[12] Markowitz, H. (1952) 'Portfolio Selection', *Journal of Finance,* 7(1), pp. 77-91

[13] Yang, J., Zhou, Y. and Leung, W. K. (2010). 'Asymmetric Correlation and Volatility Dynamics among Stock, Bond, and Securitized Real Estate Markets', *Journal of Real Estate Finance and Economics*. Online first. DOI: 10.1007/s11146-010-9265-0

# Genetically Evolved Agents for Stock Price Prediction

**Milan Jakel**

Research Institute of the IT4Innovations Centre of Excellence, Silesian University in Opava, Bezrucovo namesti 13, 746 01 Opava, Czech Republic, milan@jakel.cz

**Petr Sosík**

Research Institute of the IT4Innovations Centre of Excellence, Silesian University in Opava, Bezrucovo namesti 13, 746 01 Opava, Czech Republic, petr.sosik@fpf.slu.cz

*Abstract: Our intention is to evolve agents genetically to maximize their stock price prediction ability. A newly designed stock price prediction algorithm benchmark is used as a fitness function. A portfolio of seven US blue-chip stocks has been set for experimental purposes. We use daily time series of stock prices from 2000 to 2011 divided into two segments, in-sample for genetic algorithm evolution and out-of-sample for evaluation. Agents act as prediction algorithms based on Japanese candlestick patterns expressed as logical formulas and encoded by a tree encoding. Evaluation by the benchmark shows this is a promising way to develop successful stock prices prediction algorithm.*

*Keywords: agent; genetic; multi-agent; genetic algorithm; time series; financial; stock; stock market; prediction; forecast*

## 1 Introduction

The original methods for financial time series prediction are based on mathematical statistics [3, 16]. When predicting a stock market we create a prediction of a particular time series [3]. Each stock traded on a stock exchange is characterized by a time series. A record of such a time series includes four price values: open price, high price, low price and close price [8, 20]. For short, we call it an OHLC time series. See an example of an OHLC time series in Table 1.

Table 1
An example of several records of daily OHLC stock time series

| Date | Open | High | Low | Close |
|---|---|---|---|---|
| 2011/12/19 | $ 17,34 | $ 17,57 | $ 17,22 | $ 17,36 |
| 2011/12/20 | $ 17,36 | $ 17,38 | $ 17,22 | $ 17,22 |
| 2011/12/21 | $ 17,22 | $ 17,30 | $ 17,17 | $ 17,21 |

The multi-agent paradigm [10, 21] offers another possible approach to stock time series prediction. An agent acts as a prediction algorithm. The stock exchange is the environment in which such agents operate. The environment provides an input for the agent. The input is an OHLC time series of a particular stock. The agent affects the environment by producing stock exchange orders for buying and selling stocks. The scheme of an agent and a multi-agent environment is depicted in Figure 1.

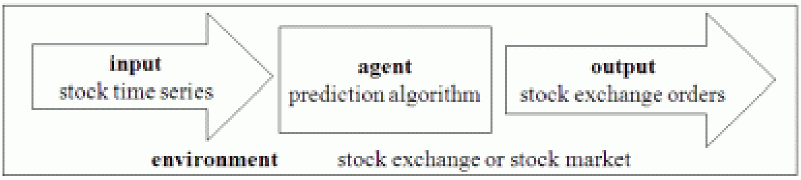



Figure 1
Scheme of an agent, its inputs, outputs, and the environment

There are four types of **stock exchange orders** [8, 20]:

- BUY to buy stocks,
- SELL to sell stocks previously bought using a BUY order,
- SHORT to borrow and sell stocks,
- COVER to buy and return stocks previously borrowed and sold using a SHORT order.

For clarification, we consider MARKET stock exchange orders only [20].

Agents can profit from an **uptrend** of stock prices by realizing a **long** position. A long position is opened by placing a BUY order. It is closed by placing a SELL order. On the other hand, agents can profit from a **downtrend** of stock prices by realizing a **short** position. A short position is opened by placing a SHORT order and it is closed by placing a COVER stock exchange order.

Agents in our experiment **autonomously decide** when to open a long position (long trade) or when to open a short position (short trade) or when to do nothing. For this purpose, there is a **prediction algorithm** as a crucial part of the internal structure of the agent. In our experiment, all positions (trades) are closed (terminated) at the end of the trading day, and trading costs (broker commissions) are not included.

Many recent papers deal with genetic evolution of prediction algorithms based on indicators of technical analysis [8, 20]. In this paper, agents do not use genetically evolved indicators of technical analysis but Japanese candlestick patterns [17, 19] as their prediction algorithms. We generalize this approach and draw inspiration from principles of data mining and knowledge discovery [11].

In the process of time series prediction, it is necessary to set an **error function**. And when doing a genetic evolution of agents, it is necessary to set a **fitness function**. In most of the recent papers, there are widely-known mean square error (MSE) and root mean square error (RMSE) error functions [3] used both as fitness functions and as evaluation functions. In this paper, we use the newly-designed stock price prediction algorithms benchmark which is based on the simulation of trades the prediction algorithm has generated. The benchmark is able to compare both numerical and categorical prediction algorithms.

For simplicity, let us consider the term 'agent' and the term 'prediction algorithm of agent' as synonyms and let us later in this paper solely use the term 'agent'.

# 2 Literature Review

## 2.1 Recent Papers

Nowadays, stock market prediction is frequently based on techniques and tools of artificial intelligence. Construction methods of stock price prediction algorithms by genetic algorithms are presented in [1], [4], [6], [13], and [14]. Soft computing methods, like neural networks and fuzzy systems, for time series prediction are described, e.g. in [15].

In [1], the authors use genetic algorithm (GA) with tree encoding. Entry rules they search for are based on mathematical statistics, while entry rules in this paper are based on candlestick patterns.

Chen [4] uses GA to generate entry rules based on indicators of technical analysis [8, 20], for instance, relative strength index (RSI). Chen's fitness function includes a standard deviation of returns (i.e. profits and losses), while our fitness function includes a variance coefficient of the set of profits and losses. Chen uses a generational genetic algorithm and a fixed generations termination condition, as we do in our experiment.

Genetic programming and genetic network programming are the basic applications and extensions of genetic algorithms. In [6], they use genetic network programming to develop a stock trading model based on indicators of technical analysis. Their multi-agent network system is trained by SARSA learning algorithm.

In [13], Japanese candlestick patterns are modelled using fuzzy linguistic variables. This model is further used in [14] and extended with a genetic algorithm for the selection of the fittest candlestick patterns.

In [9], we proposed a design of a stock price prediction algorithm based on case based reasoning (CBR). The problem of measuring the distance between the original case (stored in a knowledge base) and the new case, which is the main problem to solve during CBR system design, inspired the author to create the stock price prediction algorithms benchmark described in the next section.

## 2.2 Stock Price Prediction Algorithms Benchmark

The main idea of the benchmark is to evaluate a stock price prediction algorithm. It is done by quantifying two description statistics values of the set of profit and losses. Such a set is resulting from transactions executed (simulated) according to entry signals generated by the prediction algorithm. We use the benchmark in this paper in two ways:

- to construct a fitness function for the GA;
- to benchmark the fittest agents resulting from the GA.

The price at which the position is opened is called the **entry price** (denoted as *Entry*) [20]. The price at which the position is terminated is called the **exit price** (denoted as *Exit*) [20]. The benchmark calculations are specified below.

Let $t$ be the number of entry signals generated by a given agent operating on a given time series. Let $P(1), P(2), \ldots, P(t)$ be a set of coefficients of profit or loss resulting from the corresponding transactions.

For a long trade $r \in \langle 1, t \rangle$ we calculate the coefficient of profit or loss as follows:

$$P(r)_{long} = Exit \,/\, Entry \tag{1}$$

For a short trade $r \in \langle 1, t \rangle$ we calculate the coefficient of profit or loss as follows:

$$P(r)_{short} = Entry \,/\, Exit \tag{2}$$

$P(r)$ coefficients are relative numbers. In the benchmark, we calculate the average and the variance of the $P(r)$ relative numbers set. The geometric mean is suitable to calculate an average of such set of relative numbers.

The geometric mean $M$ of a set of $t$ values $P(1), P(2), \ldots, P(t)$ is defined as follows:

$$M = \left( P(1)\,.\,P(2)\,.\,\ldots\,.P(t) \right)^{\frac{1}{t}} = \sqrt[t]{P(1)\,.\,P(2)\,.\,\ldots\,.P(t)} \tag{3}$$

Let *P(1), P(2), ..., P(t)* be a set of *t* values and let *M* be the geometric mean of the *P(1), P(2), ..., P(t)* set. The variation coefficient *C* of the *P(1), P(2), ..., P(t)* set is defined as follows:

$$C = \frac{\sqrt{\sigma^2}}{M} \tag{4}$$

where $\sigma^2$ is the variance defined as follows:

$$\sigma^2 = \frac{1}{t}\sum_{i=1}^{t}(P(i) - M)^2 \tag{5}$$

# 3 Multi-Agent System Specification

We assume a population of *n* agents *A(1), A(2), ..., A(n)*. We assume *m* generations of agents. Let us denote the initial generation as *A(1, 0), A(2, 0), ..., A(n, 0)* and the last generation as *A(1, m), A(2, m), ..., A(n, m)*. Each agent represents a prediction algorithm based on a candlestick pattern encoded by a tree encoding.

We assume a set of *k* OHLC time series *S(1), S(2), ..., S(k)*. The environment of the multi-agent system [10, 21] is represented by the *S(1), S(2), ..., S(k)* set.

## 3.1 Agents, Candlestick Pattern Logical Formulas

Japanese candlestick patterns [17, 19] can be expressed in a form of logical formulas. In Figure 2, there are the Bullish Engulfing pattern and the Bearish Engulfing pattern depicted in a candlestick chart.

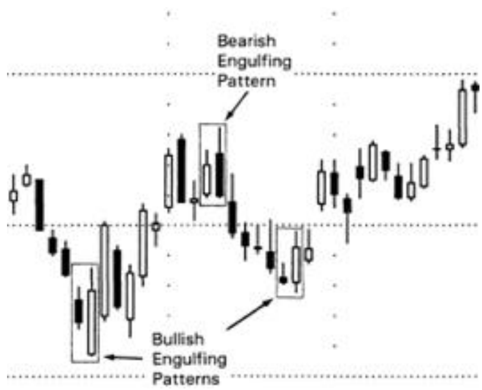



Figure 2
Bullish Engulfing pattern and Bearish Engulfing pattern. Source: [19]

An occurrence of the Bullish (or Bearish) Engulfing pattern in a candlestick chart should be interpreted as a long (or short) position entry signal [19].

The Bullish Engulfing pattern is expressed in a form of logical formula as follows:

$$( ( Open[1] > Close[1] ) \text{ AND } ( Open[0] < Close[0] ) ) \text{ AND } ( ( Open[1] < Close[0] ) \text{ AND } ( Open[0] < Close[1] ) ) \tag{6}$$

The Bearish Engulfing pattern is expressed in a form of logical formula as follows:

$$( ( Open[1] < Close[1] ) \text{ AND } ( Open[0] > Close[0] ) ) \text{ AND } ( ( Open[1] > Close[0] ) \text{ AND } ( Open[0] > Close[1] ) ) \tag{7}$$

Candlestick pattern formulas (6) and (7) are composed of four variables *Open[], High[], Low[], Close[]* with the **shift parameter** in square brackets. Value *[0]* means no shift. Usually, when there is no shift, i.e. *[0]*, the shift parameter is omitted. Usually, *Open[]* is indicated as *O[]* for short, *H[]* stands for *High[]*, *L[]* stands for *Low[]*, and *C[]* stands for *Close[]*. Value *[1]* means the first previous record of the time series, value *[2]* means the second one, etc.

Furthermore, candlestick pattern formulas (6) and (7) are composed of numerical operators (+, -), comparison operators (<, >, >=, =<), logical operators (*AND, OR*), and parentheses *(, )*.

Each agent *A(c, d)* is represented by a formula composed like (6) and (7). There are two different types of agents: **long agents** and **short agents**. Long agents generate long entry signals only and short agents generate short entry signals only. The candlestick pattern formulas are evaluated for each record of given time series. *TRUE* evaluation of the formula poses an entry signal. *FALSE* evaluation of the formula poses no signal.

## 3.2 Environment, Stock Price Time Series

For the purposes of our experiment, a portfolio of seven stocks listed in the US stock index Dow Jones Industrial Average (DJIA) has been compiled. All stocks in the portfolio are traded at NYSE and NASDAQ stock exchanges. The portfolio is listed in Table 2.

Table 2

Portfolio of seven global stocks used in our experiment. Source: [22]

| Symbol | Company name | Sector |
|---|---|---|
| AA | Alcoa Inc. | basic materials |
| BA | Boeing Co. | aerospace |
| CAT | Caterpillar Inc. | industrial goods |
| DIS | Walt Disney Co. | entertainment |
| GE | General Electric Company | machinery |

| | | |
|---|---|---|
| IBM | International Business Machines Corp. | computer systems |
| KO | The Coca-Cola Company | consumer goods |

We use an OHLC time series with **daily timeframe** [3]. One record of a daily time series represents one stock exchange trading day [8, 20]. Before being used in our experiment, the given time series were split adjusted and dividend adjusted [8].

### 3.3 Entry Signals and Corresponding Transactions

Let *S(f, i)* be the *i*-th record of time series *S(f)*. Agent *A(c, d)*, where $c \in <1, n>$ and $d \in <0, m>$, working on time series *S(f)*, where $f \in <1, k>$, generates entry signals as follows:

- for every record *i* of time series *S(f, i)* is the logical formula *A(c, d)* evaluated,
- if the evaluation of *A(c, d)* on *S(f, i)* is *TRUE* and *A(c, d)* is a long agent, then an entry signal for a long trade is generated,
- if the evaluation of *A(c, d)* on *S(f, i)* is *TRUE* and *A(c, d)* is a short agent, then an entry signal for a short trade is generated,
- if the evaluation of *A(c, d)* on *S(f, i)* is *FALSE*, then no entry signal is generated.

Entry signals are generated on daily stock time series. This method of generating entry signals is mentioned in [20] and it is called **swing trading** or **intraday trading**. Once we have an entry signal, we enter the market (open a position) at the beginning of next trading day, i.e. 'at open' [8, 20]. We terminate the position at the end of the same trading day, i.e. 'at close'.

# 4 Genetic Algorithm Specification

A genetic algorithm (GA) is an evolutionary search heuristic resulting in a fittest solution of an encoded problem. The advantage of GA is in its parallelism. Information on GA proposed in this chapter is based on and related to references [2], [5], [7], [12], and [18].

First of all, an encoding mechanism must be designed to represent each agent as a genome. Typically used encodings in GA are binary encoding, value encoding, permutation encoding, and tree encoding. In our experiment, the **tree encoding** is used. See below the specification.

To evaluate the quality of particular agents, a fitness function is needed in GA. A fitness function is used by a selection operator to select quality agents for reproduction. In our experiment, a **multi-objective** fitness function **based on** stock price prediction algorithms **benchmark** is used. See below the specification.

The initialization method sets the way how the initial generation of agents *A(1, 0), A(2, 0), …, A(n, 0)* must be initialized at the beginning of the GA evolution. In our experiment, the initial agents are generated by a random function in a form similar to candlestick pattern formulas given in expressions (6) and (7). Such an initialization method implements **randomly generated** candlestick pattern logical formulas.

The selection operator determines how agents are selected from the population for the crossover operator. There are two basic methods to select agents to be parents for crossover: roulette wheel selection and tournament selection. In our experiment, we use the **tournament selection**; i.e. a set of two or more randomly selected agents is compared and the fittest agent is selected for crossover.

The crossover operator exchanges parts of genomes among two or more parent agents to create a new offspring. In our experiment, the **subtree crossover** is used. Subtree crossover produces an offspring by replacing a subtree of one parent agent by a subtree of other parent agent. There is a set of rules to keep the order to secure the validity of the newly created candlestick pattern logical formulas. The principle of the subtree crossover is depicted in Figure 4.

The mutation operator randomly changes agents resulting from the crossover operator. In our experiment, the **subtree mutation** is used. The subtree mutation replaces the parental subtree with a randomly generated subtree. There is a set of rules to keep the order to secure the validity of newly created candlestick pattern logical formulas. Mutation introduces more randomness into the population. The principle of the subtree mutation is depicted in figure 5.

The replacement method specifies how the population is updated by removing parents and adding offsprings. In GA, there are several commonly used replacement methods: uniform replacement, crowding replacement, tournament replacement, and parental replacement. In our experiment, the **parental replacement** is used, i.e. every parent agent is replaced by its offspring agent.

The termination condition determines when to stop the evolution and obtain the solution that is represented by the fittest agent from the last population. There are several termination conditions usually used in GA: fitness convergence, diversity convergence, fitness target, and fixed generations. In our experiment, we use the **fixed generations** termination condition. The evolution is stopped when the *m*-th generation is completed.

## 4.1 Tree Encoding, Tree Crossover, Tree Mutation

In our experiment, for the purpose of encoding candlestick pattern logical formulas into GA genomes, we use the tree encoding. The principle of the tree encoding is illustrated in Figure 3. There is Bullish Engulfing pattern logical formula (6) encoded into the tree genome.

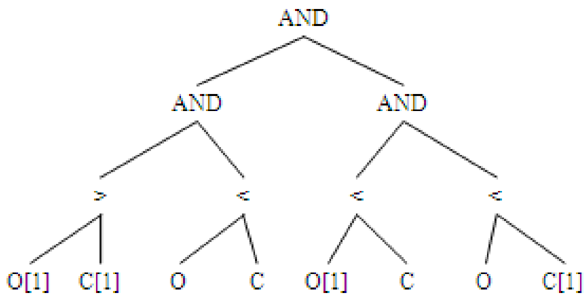



Figure 3

Bullish Engulfing pattern logical formula encoded into the tree genome

In Figure 4, there is an example of a subtree crossover operator. And in Figure 5, there is an example of a subtree mutation principle.

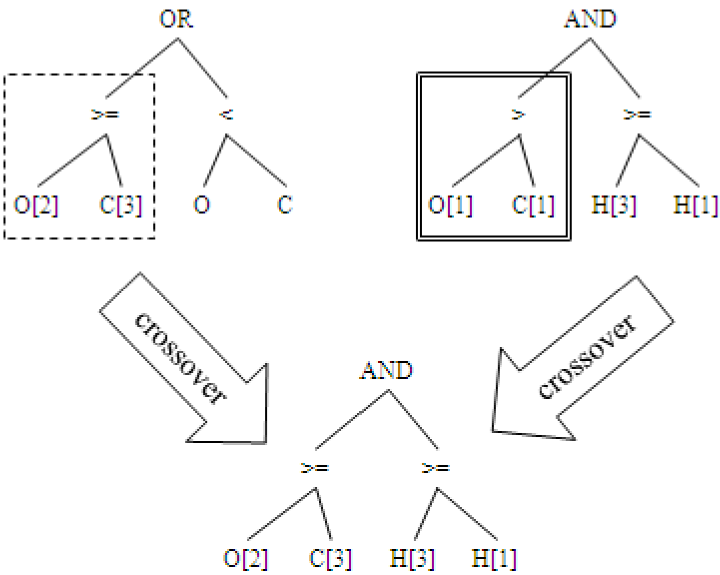



Figure 4

An example of the subtree crossover. The offspring is created by replacing a subtree (marked by double line region) of one parent agent by a subtree (marked by dashed line region) of other parent agent.

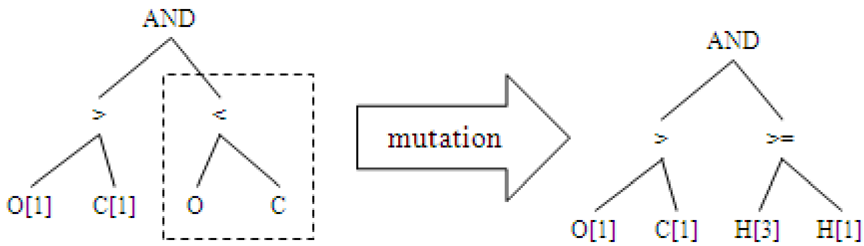



Figure 5
An example of a subtree mutation. The parental subtree intended for mutation is marked by dashed line region. The arrow points from the parental agent to the offspring agent.

## 4.2 Fitness Function

We use a multi-objective fitness function in our experiment. There are three objectives:

- to maximize the average of profits and losses,
- to minimize the variance of profits and losses,
- to maximize the number of trades (i.e. the number of generated entry signals).

Let *P(r)*, where $r \in <1, t>$, be a set of *t* coefficients of profit or loss of a given prediction algorithm (or agent) operating on a given time series. The first objective is based on the benchmark calculation of geometric mean (3) of the *P(r)* set. The second objective is based on benchmark calculation of variance coefficient (4), (5) of the *P(r)* set.

To calculate the value of the fitness function of agent *A(c, d)* operating on time series *S(f)* we firstly need to determine:

- the possible maximum average *MaxM* (geometric mean) of the set of coefficients of profit or loss *P(r)* that can be reached on a given time series *S(f)*,
- the possible minimum variation coefficient *MinC* of the set of coefficients of profit or loss *P(r)* that can be reached on a given time series *S(f)*,
- the possible maximum number of trades *MaxNum* that can be done on a given time series *S(f)*.

Let us assume the agent *A(c, d)* operating on time series *S(f)*. The agent has generated a set of entry signals. The set of profit or loss coefficients *P(r)* has been calculated according to expressions (1) and (2). Let *M* be the geometric mean of *P(r)* set calculated according to expression (3). Let *C* be the variation coefficient of *P(r)* set calculated according to expressions (4) and (5). And let *Num* be the number of elements of *P(r)* set, i.e. the number of entry signals of *A(c, d)* on time series *S(f)*. The fitness function is defined as follows:

$$Fitness = ( M / MaxM ) + ( MinC / C ) + ( Num / MaxNum ) \quad (8)$$

## 4.3 Summary of GA Parameters

In this chapter, we provide a list of parameters of the genetic algorithm (GA) used in our experiment, see Table 3.

Table 3
Summary of genetic algorithm parameters used in our experiment

| **Parameter** | **Value** |
|---|---|
| initialization method | random function |
| population size | 50 agents |
| encoding | tree encoding |
| maximum allowed tree depth | 4 levels |
| fitness function | multi-objective, see expression (8) |
| replacement method | parental replacement |
| selection operator | tournament selection |
| tournament size | 2 agents |
| crossover operator | tree crossover |
| crossover production | 45 agents (90% of population size) |
| mutation operator | subtree mutation |
| mutation production | 5 agents (10% of population size) |
| termination condition | fixed generations |
| number of generations | 25 generations |

According to the results of our preliminary experiments we have set the parameters of the GA to achieve the best convergence. We have set the population size to 50 and total number of generations to 25 as for the terminal condition. Such a GA is called **generational genetic algorithm**.

When using a tree encoding, it should be convenient to set a maximum for the number of levels the tree encoding can use. When the value of the maximum is above the optimum, the GA is harder to converge. If allowing a large **maximum tree depth**, then also allowing candlestick pattern formulas created by the GA to become much more complicated, and this is probably resulting into an **overfitting** [3] of given stock price time series. We have set the maximum allowed tree depth to 4 levels.

The GA was best to converge when we set the percentage of offspings created by the crossover operator to 90 percent and the less 10 percent was created by the mutation operator.

Time series *S(f)* used in our experiment represent a continuous period of daily stock price data from $3^{rd}$ Jan 2000 to $30^{th}$ Dec 2011. The given time series are split into two periods. The **in-sample** period of given time series is used for the evolution of agents, i.e. for the **training** of the multi-agent system. The in-sample period runs from $3^{rd}$ Jan 2000 to $31^{st}$ Dec 2008. The **out-of-sample** period is used for the **evaluation** of the multi-agent system. The out-of-sample period runs from $1^{st}$ Jan 2009 to $30^{th}$ Dec 2011.

# 5 Experimental Results

In our experiment, the genetic algorithm (GA) was launched 14 times. For each of the seven given time series we created 25 populations of long agents and 25 populations of short agents. The fittest long agent and the fittest short agent for a particular time series are the solutions found by the GA search heuristics.

Table 4
Agents with the best fitness

| Type of agent | Agent (logical formula of the prediction algorithm) |
|---|---|
| AA long | (H[1] >= C[4]) AND (O - L[1] <= H[4] - H[2]) |
| AA short | (O - L[2] > H[2] - O[2]) OR (C[1] - H[3] >= H[2] - H[4]) |
| BA long | (L[2] - H[3] > H[2] - C[5]) OR (L[2] - L[3] > H[1] - C[5]) |
| BA short | (C[3] - C[4] > C[4] - L[1]) AND (C[2] - C[4] >= H[1] - H[2]) |
| CAT long | (H[2] >= H[4]) AND ( (L[3] - L[4] > O[4] - L[5]) OR (H[3] < H[5]) ) |
| CAT short | (H - C[4] < O[3] - O[1]) AND (H[4] > H[2]) |
| DIS long | (C[5] - C[4] > O - L[1]) OR (H[1] - L[5] <= L[5] - L[2]) |
| DIS short | (L >= L[4]) AND (C[4] - O[4] >= O[2] - C[3]) |
| GE long | (H[4] <= O[2]) AND (L[4] - L[1] <= O[2] - H[4]) |
| GE short | (H - C[1] <= L[3] - L[1]) OR (C[4] >= H[2]) |
| IBM long | (L[3] - C[1] < C[1] - H[2]) OR (L - C[2] >= O[4] - L[3]) |
| IBM short | (H[2] - L[1] >= H - O[4]) AND (L[4] - H[5] >= O - H[3]) |
| KO long | (O[3] - H[4] >= C[5] - H[1]) AND (L[2] >= O[3]) |
| KO short | (L[5] >= O[3]) AND (H - L[3] > C - H[4]) |

Table 4 list the agents (logical formulas of prediction algorithms based on candlestick patterns) which have gained the highest fitness function values on the in-sample periods of a given time series. These agents are evaluated both on in-sample and on out-of-sample periods of a given time series using the benchmark according to expressions (3), (4), and (5). The results of the evaluation are shown in Table 5.

Table 5
Benchmark results of the fittest agents both on in-sample (INS) and on out-of-sample (OOS) periods

| Type of agent | Geometric mean | | Variation coefficient | |
|---|---|---|---|---|
| | INS | OOS | INS | OOS |
| AA long | 1.00**266** | 1.00**143** | 1.984 | 2.646 |
| AA short | 1.00**195** | 1.00**210** | 2.149 | 2.645 |
| BA long | 1.00**132** | 1.00**018** | 2.024 | 1.884 |
| BA short | 1.00**125** | 1.00**050** | 1.737 | 1.500 |
| CAT long | 1.00**060** | 1.00**242** | 1.734 | 1.957 |
| CAT short | 1.00**021** | 1.00**296** | 1.997 | 2.240 |
| DIS long | 1.00**158** | 1.00**099** | 1.968 | 1.605 |
| DIS short | 1.00**113** | 1.00**008** | 1.783 | 1.318 |
| GE long | 1.00**045** | 0.99858 | 1.694 | 1.611 |
| GE short | 1.00**103** | 1.00**462** | 1.989 | 2.293 |
| IBM long | 1.00**070** | 1.00**286** | 1.509 | 1.111 |
| IBM short | 1.00**110** | 0.99974 | 1.714 | 1.246 |
| KO long | 1.00**148** | 1.00**068** | 1.274 | 0.897 |
| KO short | 1.00**073** | 1.00**031** | 1.511 | 1.062 |

The INS values of geometric mean and variation coefficient benchmark the quality of the fittest agents resulted from the GA. The same INS time series that is used for the training of the multi-agent system is also used for the calculation of the benchmark.

The OOS values of the geometric mean and the variation coefficient benchmark the same agents but now operating in an **unknown environment**. The unknown environment for the agents is the OOS period of given time series.

By comparing the INS and OOS values in Table 5 we can assess the **quality of** proposed stock price **prediction algorithms** (i.e. agents) found by the GA. Some of the agents work better in the evaluation than in the training, that is, 'AA short', 'CAT long', 'CAT short', 'GE short', and 'IBM long'.

The OOS geometric mean of the 'AA short' agent is greater than the INS one. That means the average profit per trade reached in the evaluation period is greater than the one reached at the end of the GA evolution. The value of 1.00210 at 'AA short' OOS results means the average profit of the set of trades done on the AA daily time series is 0.21% per trade.

## Conclusions

We apply genetic algorithms (GA) to construct candlestick pattern logical formulas. Then we use the constructed patterns as stock price prediction algorithms. The quality of prediction algorithms is measured by a newly designed stock price prediction algorithms benchmark.

Both in-sample and out-of-sample benchmark results confirm that some of genetically evolved agents (i.e. prediction algorithms based on candlestick pattern logical formulas) are doing well at the stock price time series prediction.

The fittest short agent for the GE stock price time series resulting from the GA gained the highest benchmark evaluation as for the out-of-sample test. The value of 1.00462 means the average profit of trades done by 'GE short' from January 2009 to December 2011 is 0.46% per trade.

The GA we have used was easy to converge, although the GA parameters were not necessarily optimal. Our experiment has demonstrated the feasibility of using GA for stock price prediction algorithms construction.

**Acknowledgement**

The author wishes to thank Prof. Jozef Kelemen from the Institute of Computer Science, Silesian University in Opava for consultations on agents and multi-agent systems, Assoc. Prof. Petr Sosík from the Institute of Computer Science, Silesian University in Opava for his supervision and reading several versions of the manucript, and Prof. Dušan Marček from the Institute of Computer Science, Silesian University in Opava for consultations on financial time series forecasting.

This work was supported by the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070). Research was partially funded by the Silesian University in Opava under the Student Funding Scheme, project no SGS/7/2011.

**References**

[1] Allen, F., Karjalainen, R.: Using Genetic Algorithms to Find Technical Trading Rules. *Journal of Financial Economics* **51** (1999) 245-271

[2] Bauer, R. J.: *Genetic Algorithms and Investment Strategies*. Wiley, New York, 1994

[3] Box, G. E. P., Jenkins, G. M., Reinsel, G. C.: *Time Series Analysis: Forecasting and Control*. Wiley, New York, 2008

[4] Chen, J.-S.: Trading Strategy Generation Using Genetic Algorithms. *Asian Journal of Information Technology* **4** (2005) 310-322

[5] Chen, S.-H.: *Genetic Algorithms and Genetic Programming in Computational Finance*. Springer, Berlin, 2002

[6] Chen, Y., Mabu, S., Shimada, K., Hirisawa, K.: A Genetic Network Programming with Learning Approach for Enhanced Stock Trading Model. *Expert Systems with Applications* **36** (2009) 12537-12546

[7] Deboeck, G.: *Trading on the Edge*: *Neural, Genetic, and Fuzzy Systems for Chaotic Financial Markets*. Wiley, New York, 1994

[8] Graham, B., Zweig, J.: *The Intelligent Investor: The Definitive Book on Value Investing*. Harper Business Essentials, New York, 2003

[9] Jakel, M.: Stock Investment Recommendation Making Supported by Case Based Reasoning. In: Kelemen, J., Nehéz, M.: *Proceedings of 6th International Workshop on Knowledge Management, Bratislava 2011*. Vysoká škola managementu, Bratislava, 2011

[10] Kelemen, J.: Agents from Functional-Computational Perspective. *Acta Polytechnica Hungarica* **3** (2006) 37-54

[11] Kelemen, J., et al.: *Knowledge in Context: Few faces of the knowledge society*. Iura Edition, Bratislava, 2010

[12] Koza, J. R.: *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, 1992

[13] Lee, C.-H., Liu, A.: A Financial Decision Supporting System Based on Fuzzy Candlestick Patterns. In: *Proceedings of 9th Joint Conference on Information Sciences*. Atlantis Press, Taiwan, 2006, pp. 1289-1292

[14] Lee, C.-H., Liaw, Y.-C., Hsu, L.: Investment Decision Making by Using Fuzzy Candlestick Pattern and Genetic Algorithm. In: *Proceedings of 2011 IEEE International Conference on Fuzzy Systems*. IEEE, Taiwan, 2011, pp. 2696-2701

[15] Marček, M., Pančíková, L., Marček, D.: *Ekonometria a soft computing* (in Slovak). EDIS, Žilina, 2008

[16] Marček, M.: *Viacnásobná štatistická analýza dát a modely časových radov v ekonómii* (in Slovak). Silesian University, Opava, 2009

[17] Marshall, B. R., Young, M. R., Rose, L. C.: Candlestick Technical Trading Strategies: Can They Create Value for Investors? *Journal of Banking & Finance* **30** (2006) 2303-2323

[18] Mitchell, M.: *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, 1998

[19] Nison, S.: *Japanese Candlestick Charting Techniques: A Contemporary Guide to the Ancient Investment Techniques of the Far East*. 2nd edition, New York Institute of Finance, New York, 1991

[20] Williams, L.: *Long-Term Secrets to Short-Term Trading*. Wiley, New York, 1999

[21] Wooldridge, M.: *An Introduction to MultiAgent Systems*. 2nd edition. Wiley, New York, 2009

[22] Yahoo! Finance [Online] Available: http://finance.yahoo.com/ [2012/01/20]

# The IT Aspect: Towards a More Efficient System-Control and Audit of EU Funds for the Cohesion Policy

**Judit Fortvingler**

Faculty of Economic and Social Sciences
Budapest University of Technology and Economics
Magyar tudósok körútja 2, H-1117 Budapest, Hungary
E-mail: fortvingler@finance.bme.hu

*Abstract: Assuming that not more audits but better coordinated ones are preferred, the key objective of the article is to explore how deeper exploitation of the current IT background could improve the efficiency of audits and controls of EU funds for Cohesion policy. After presenting the internal control and external audit functions, various scenarios of IT utilization which form different levels of IT convergence are delineated. Finally, the author concludes that audit and control activities could be further developed, particularly by a more extensive use of the IT background. From a professional point of view, direct access to core national databases would provide auditors with additional information on systems at Member States level. The research method applied was interviews with professionals of the European Court of Auditors, as well as with the developer of the Unified Monitoring Information System and different authorities in Hungary. In addition, documents were consulted which are available at the Historical Archives of the European Union and the European Court of Auditors, through the Postgraduate Research Grant Programme, which offered a unique opportunity.*

*Keywords: Unified Monitoring Information System; multi-level assurance system; Cohesion policy*

## 1 Introduction

The EU funds assigned to the Cohesion policy in support of growth and jobs represent over one-third of the total budget. The allocation of these funds to final beneficiaries is carried out under a complex system, where the implementation and the control activity are shared between the European Commission and

Member States ('shared management')[1]. The Member States bear primary responsibility for creating an effective management and control system, while the European Commission plays a supervisory role. Witnessing a growth in audit costs for the programming period 2007-2013, the European Court of Auditors (hereinafter the Court) assessed the estimated error rate to exceed five percent and identified serious deficiencies in shared management areas (ECA, 2012a). One may say that an error rate of about five percent is a good and acceptable value. However, an estimated error rate surpassing two percent, which is generally used by the European Court of Auditors as the ‘materiality threshold’, implies a qualified opinion on the implementation of the EU budget, that stakeholders expect not to have.

To curb the increasing costs and push the high error rate down, they suggested improvements in cooperation and coordination between the different actors of the existing control and audit systems (European Commission DG Regional Policy, 2010; European Parliament, 2011).

Even though existing literature that focuses on cooperation is already extensive, it mainly examines two aspects. On the one hand, the cooperation between the Court and supreme audit institutions (SAI's) has been dealt with for years (Castells, 2005). The pilot project on coordinated audit and the activity of the Contact Committee of the Supreme Audit Institutions present remarkable examples of their commitment to achieve a higher degree of collaboration. On the other hand, the Commission’s internal control framework has been reviewed by scholars numerous times. It is a common view that the Commission has made tremendous efforts to construct its multi-level assurance system in order to accomplish smoother financial management. Nevertheless, some information stored in the computerised system has not yet been fully exploited.

The objective of this article is to address this deficiency by investigating how the IT aspect can add value to the work of auditors and controllers. This leads to the key research question: How can the existing IT background enhance the efficiency of audits and controls for Cohesion policy?

The article is structured as follows. First, the author describes the audit and control system of shared management and gives insight into the efforts made in the past to create a system where reliance on others' work is a key factor. Secondly, the most significant challenges of the current system are revealed. Finally, the potential role of IT in expediting the process of convergence, which in turn defends the EU’s financial interest, is examined. The conclusion summarizes the research findings and their implications.

[1] Although the agricultural expenditure formulates the other pillar of EU funds under shared management, this article concentrates on the poliy area of Cohesion.

The empirical background for the answer to the research question includes regulatory frameworks, both at EU and national level, and interviews with professionals of the European Court of Auditors, the State Audit Office of Hungary, the Directorate General for Audit of European Funds (audit authority), the National Development Agency (managing authority), and the developer of the Unified Monitoring Information System. In addition, academic publications and internal documents of the Court, the latter available through the Postgraduate Research Grant Programme, served as a solid basis for the research.

# 2 The Multi-Level Assurance System for Cohesion Policy

## 2.1 The Audit and Internal Control Functions

The Commission holds the overall responsibility for the implementation of the EU budget in accordance with the relevant regulations. For funds of the Cohesion policy under shared management, the management and control activities are performed in cooperation with Member States. Hence, each Member State has to put in place an adequate management and a multi-level control system, both of which guarantee sound financial management of EU funds while ensuring regularity and eligibility of the expenditures made from these funds. At the top of the internal control system, the Commission itself, in addition to the audit activity of the Directorates-General, has its own Internal Audit Service.

Why is system quality such a key factor for auditors? As an external guardian, the European Court of Auditors is in charge of scrutinizing public spending and safeguarding the financial interests of EU citizens. The Court adopts the so-called system-based approach, which means that audit engagement starts with a thorough analysis of the auditee's internal control system, to collect evidence that proves it is functioning effectively. If system assessment shows that it is operating well, the extent of direct testing can be reduced. This is fundamental to understanding why the quality of any system has crucial importance for auditors.

Currently, neither the Court's nor the Commission's auditors have direct access to national IT databases; they receive core data on request from SAI's or other authorities.

Table 1
The internal control and external audit systems

| | **Internal control** | **External audit** |
|---|---|---|
| **at national level (Member States)** | Implementing authorities | Supreme Audit Institutions |
| **at European Union level** | Directorates-General<br>Internal Audit Service | European Court of Auditors |

Prior to analysing how the existing IT background could enhance control and audit activities, one must know the elements of the complex control system at the national level. As implied in the provisions of the current regulatory framework, each Member State has to assign certain authorities with different mandates: a managing authority, a certifying authority, an audit authority, and optionally, intermediate bodies.

The *managing authority's* main task is to implement the EU budget at the national level, complying with the principle of sound financial management. To fulfill its duty, it certifies that the expenditure declared by beneficiaries satisfies the conditions of the approval decision and the EU and national rules. In addition, the authority submits reports on the implementation to the Commission and evaluates operational programmes. With respect to IT, the managing authority ensures that an adequate computerised system records and stores core data for planning, verification, audit, and evaluation purposes, which provides the certifying authority with all the necessary information to verify the expenditure to the Commission. The managing authority's control activity consists of, on the one hand, administrative verifications for invoices submitted by beneficiaries, and on the other hand, on-the-spot verifications of projects. To expedite the flow of information, documents of verifications have to be available for other actors of the control and audit systems. In Hungary, the National Development Agency, under the surveillance of the Government Commissioner for Development, has been designated as the managing authority.

The *intermediate bodies* are optional elements of the control system. They perform on-the-spot checks, the frequency of which varies depending on the volume of the funds the beneficiary is entitled to receive. The managing authority supervises the activities of intermediate bodies by investigating the selection procedures of, in general, at least 5 percent of the projects. In Hungary, there are several intermediate bodies that the managing authority can delegate some of its tasks to[2].

Within the whole project cycle, the managing authority and the intermediate bodies are primary users of the IT system, which provides them with core data for different purposes (e.g. monitoring, risk-based sampling). Unlike auditors, their principal duties do not include data analysis for system assessment.

The *certifying authority* is responsible for submitting certified statements of expenditure and applications for payment to the Commission. Its control activity includes both on-the-spot and administrative checks at the organizations taking part in the implementation. The authority ensures that the statement of expenditure and the underlying transactions are accurate and admissible, stemming from reliable accounting systems. Taking the example of Hungary, the State Treasury,

2 The tasks and a list of designated intermediate bodies are laid down in a government decree.

under the governance of the Minister of National Economy, is assigned to act as a certifying authority.

Finally, the heart of the *audit authority's* activity is to verify that the management and the control system are functioning effectively. To fulfill its duty, the authority performs system audits and project audits of a randomly drawn statistical sample. The sample to be audited each year has to be drawn from the expenditures submitted to the Commission in the preceding year. During on-the-spot audits, the original documentation held by beneficiaries is examined for comparison with the expenditures declared and subsequently recorded in the IT system. Moreover, compliance with the selection criteria and approval decision in the implementation phase is evaluated.

The authority annually issues an opinion about whether the management and control system is functioning effectively, in order to provide a reasonable assurance that statements of expenditure submitted to the Commission are accurate and that the underlying transactions are legal and regular. The audit authority must have all the information needed to form an opinion. That is why its activity is perceivable at each level of the control chain. Finally, as the main cog in the machinery of the control chain, the Commission supervises the operation of the audit authority. It may conclude that it can rely on the opinion issued by the audit authority; thus, the Commission will perform on-the-spot audits only if there is evidence of deficiencies in the national system.

In Hungary, the Directorate General for Audit of European Funds, under the governance of the Minister of National Economy, has been given the role of audit authority.

The elements of the multi-level assurance system for EU funds under shared management are presented below.

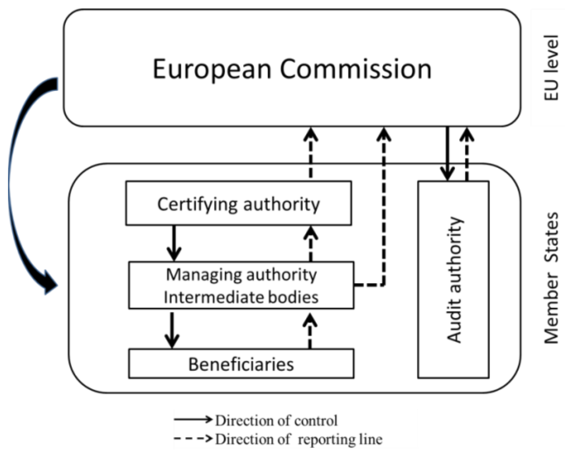



Figure 1

The multi-level assurance system

Both the certifying authority and the audit authority have direct access to the national IT database. During the interviews, a modest communication gap was observed between what the developer of the Unified Monitoring Information System (UMIS), the IT background operating in Hungary, has been told the UMIS should be capable of and what its users can actually expect from it. Consequently, there must be room for a better use of IT resources by enhanced communication.

## 2.2 The Challenges of the Current Control and Audit Systems

The Cohesion policy has been spotlighted over the past few years due to deficiencies in the control system and the high error rate, which is still over five percent. Not surprisingly, this area is a risky business by nature. The most hazardous characteristics of the systems are (Weber, 2010):

- the large number of beneficiaries;
- the numerous authorities at national level;
- the pressure to absorb the EU funds;
- the complex regulatory framework.

Mendez et al. (2011) argue that the Commission's administrative reform has generated an audit and control explosion in the field of Cohesion policy since the mid-2000s. Besides the internal organizational change, the drive to achieve a positive Statement of Assurance[3] and the intention to improve the internal control framework all contributed, as a secondary effect, to the intensification of audit and control activity. A study on the implications of the legislation for cost effectiveness of structural funds affirmed that the programming period of 2007-2013 has experienced a massive increase in the audit effort (European Parliament, 2011). Some scholars share the idea that the quantity of audits is sufficient; therefore, not more audits but better coordinated ones are desirable. To resolve the problem of increasing audit costs, a stronger cooperation and coordination between auditors at different levels has been suggested. With respect to internal control, the so-called 'single audit model' (Cipriani, 2010) has become widely accepted, a model which favours the idea that different building blocks of the system place assurance on the work of previous controls performed by lower layers, which diminishes the danger of duplication. Regarding external audit functions, there is increased cooperation between the Court and SAI's. Though the Treaty of Amsterdam declares that the Court and the SAI's 'shall cooperate in a spirit of trust while maintaining their independence, these institutions witness implementation problems caused by different mandates and dissimilar relationships with their national parliaments. A pilot project on coordinated audit, with the participation of the Court and a few SAIs, proved that this divergence

[3] The Court's opinion on the reliability of the EU accounts, and on the legality and regularity of the underlying transactions

represents a real challenge for the auditors and has an impact on future cooperation.

To conclude, one can hardly expect that with enhanced cooperation between audit and control functions these bodies can tackle the problematic issues of increasing costs and material error rate in the short term. What other possibilities are there to exploit the full potential of the current system? IT is certainly one.

# 3 The IT Aspect

The general principles of the management and control systems, established according to the provisions of Commission Regulation (EC) No. 1083/2006, include the stipulation that Member States shall arrange for 'reliable accounting, monitoring and financial reporting systems in computerised form'. In addition, the Regulation delegates to the managing authority the duty of operating such a system, which records and stores all the data on implementation necessary for the financial management of funds, monitoring, verification, and audit activities. Furthermore, the certifying authority should have accounting records of expenditure confirmed to the European Commission in computerised form.

It is of prime importance to emphasise that regulatory framework at EU level defines neither the detailed characteristics nor the correct structure of such an IT system, but rather it stipulates features it must be capable of (e.g. recording and reporting financial transactions, irregularities, and financial corrections imposed by Member States).

Turning to the national context, the UMIS has been developed to store and synchronize all the core data for policy areas financed by the European Regional Development Fund, the European Social Fund, and the Cohesion Fund. The complexity of UMIS forms a basis for monitoring and ex-post audit activities, as it covers the whole project cycle: from planning until evaluation:

- electronic submission of applications, automatic input of electronic applications;
- on-line information for applicants (status of application/ project, contract modifications, submission of missing underlying documents);
- electronic submission of payment claims (input of invoices);
- electronic submission of project reports (input of indicators into UMIS monitoring module);
- data input from web based functions is stored in a separate web database and automatic data exchange occurs every 10 minutes, which synchronizes core data among systems;
- public information on the managing authority's website (statistics, reports, report generators).

Data exchange occurs with public databases. For instance, the UMIS provides the national monitoring system of the Hungarian State Treasury with project and payment data, and the account management system of the Treasury with electronic payment requests. Additionally, the National Tax and Customs Administration of Hungary makes certain information available to the UMIS, as beneficiaries of EU funds are not allowed to accumulate tax or other public charge obligations. Otherwise, the reimbursement is suspended.

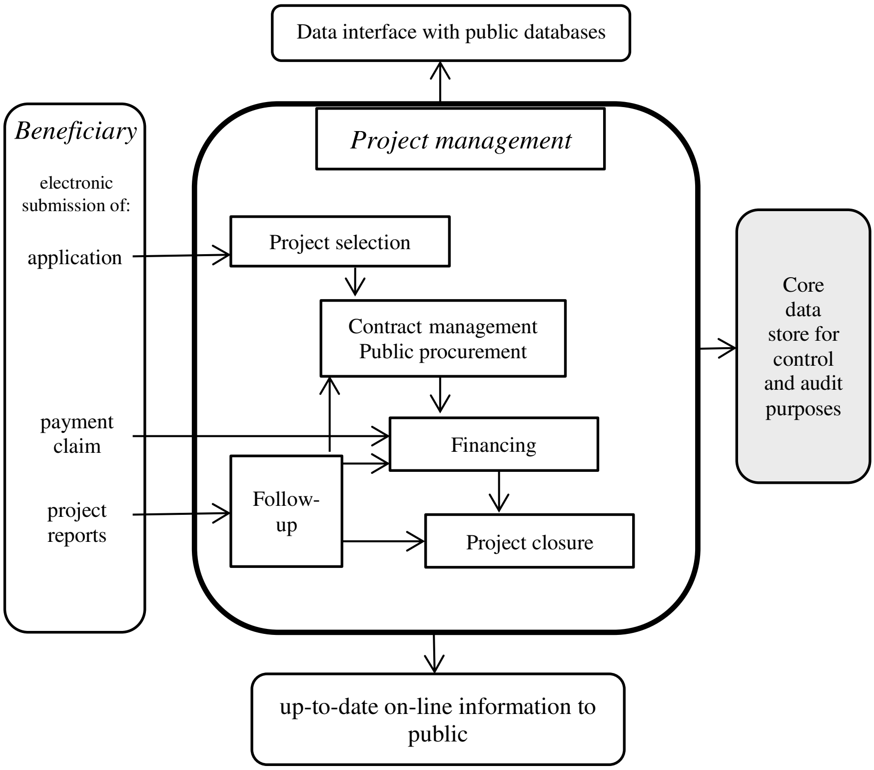



Figure 2
The complexity of UMIS

The increasing cost of controls and audits has been a cause of concern for a number of years. This article puts forward the idea of how the existing non-human resources (e.g. IT) could be better exploited without requiring additional financial resources. How could the IT systems add value to audit and control functions to make audits more efficient? How could the UMIS contribute to transparency in the utilization of EU funds and support the work of both external and internal auditors at the EU and national levels?

In fact, the degree of contribution greatly depends on the level of IT system convergence and rights of access, and thus several scenarios exist. Level 1 represents the current situation, where each Member State has its own IT system, and auditors at EU level have no direct access to national IT systems. Consequently, data service, required by the Court or the Commission for audit purposes, is time-consuming, as, in the long information chain, it sometimes goes back too slowly to the developer of the UMIS. The interviews with professionals of the Court revealed a perceptible time lag in data transfer, which has an influence on the timing of audit visits to Member States. On the other hand, data transferred to EU auditors appears to be rather stocktaking in manner; it includes funds which have been allocated to certain projects / programmes at a given time. Owing to the lack of access to core data, no further conclusions could be drawn from analysing changes in historical data recorded in the IT system, with respect to the overall functioning of the management and control systems.

Level 2 suggests a more developed and coordinated system, built on a stronger exploitation of the UMIS. In this scenario, the national IT systems remain unchanged but EU auditors have direct access to those systems. From the IT aspect, it could easily be solved and it is a cost-effective way of utilizing the existing sources.

In parallel, what kind of benefits could be achieved for auditors of funds of Cohesion policy? First, direct access by the Court and the Commission to core data stored in the UMIS would result in time savings due to the elimination of time-consuming data inquiries from the supreme audit institutions and audit authorities at the national level. This is rather a technical point of view. More importantly, from the professional aspect, direct audit evidence could be gained for system assessments with respect to management and control systems operating in Member States. Unlike Level 1, either the database could be analysed or the control activity of the managing authorities or intermediate bodies could be examined. Hence, it would be possible to bring together additional information for system assessment, which greatly influences the extent of substantive testing.

The direct access to core database addresses the issue of transparency. Cipriani (2010) emphasises that financial correction often manifested in substitution of ineligible cost, and 'Member States tend to over-declare national expenditure in order to create a buffer of eligible items'. Although the regulatory framework permits such substitutions, there is considerable debate as towhether this practice is desirable and conforms to the original aim of Cohesion policy. In this field, the Court suggests some issues for consideration. First, ineligible expenditure might be systemic by nature, and if not addressed appropriately, it can be substituted by another ineligible one. Secondly, if the replacing expenditure has originally been financed by national funds, 'cohesion spending is turned into *ex post* support for the budgets of Member States' (ECA, 2012b). The Commission's view on substitution of ineligible expenditure is that Member States should have the right to make such changes to optimize the utilization of EU funds if deficiencies appear at national level. In case of direct access to core database, auditors would have insights into the practice of substitution of ineligible expenditure by new expenditure, which would be an additional source of information when determining the outcomes of system assessment. At first sight, one can expect that difficulties may derive from the fact that auditors have to be familiarised with national IT systems. However, this obstacle can definitely be overcome; auditors frequently face the challenge of getting to know different IT systems and solutions in their day-to-day activities (e.g. computer-assisted audit techniques – CAATs).

To sum up, a slight change (direct access to core data) at the IT level, though supported by professional arguments, opens up a more political point of view.

Finally, level 3 represents the pinnacle of IT convergence. At this stage, Member States use a unified IT system for recording and storing data of EU funds. The advantages of the previous level could be enumerated here as well. In addition, time savings could be reached deriving from not having to know the diversity of IT systems across the EU. From the aspect of auditing, this alternative is the most 'convenient' way of collecting data for system assessment and for sampling. To a large extent, the unified IT system at EU level is inevitably a politically sensitive question; it would require political consensus at the highest level, embedded in EU legislation, and not modus vivendi.

Table 2

Scenarios of management and monitoring IT systems convergence

| **Characteristic of IT system** | **Degree of convergence** | **Pros** | **Cons** |
|---|---|---|---|
| **National IT system for recording and storing core data** | ***Level 1*** No access to national IT systems by EU auditors | - Right of access does not occur | - Time-lag in data transfer (difficulties in planning of audit visits)<br>- Lack of deep analysis of core data (narrowing information for conclusion) |
| | ***Level 2*** Direct access to national IT systems by EU auditors | - Time savings, cost-effectiveness<br>- Additional audit evidence for system assessment<br>- Core data base analysis<br>- Higher level of transparency | - Tackling right of access<br>- Variety of IT system put in place by Member States |
| **Unified IT system at EU level** | ***Level 3*** Direct access to unified IT system by EU auditors | - Time savings<br>- Easier data collection for system assessment and sampling | - Political sensitivity<br>- Additional regulation for the unified system |

Irrespective of the level of convergence and hence the extended access rights of data, even in the current circumstances, stronger cooperation and communication are advised so that the IT background could satisfy auditors' needs. This goes far beyond the written provisions of the regulatory framework, and it requires a contiguous interaction between different fields, e.g. IT, auditing, and project management.

## Conclusions

The Cohesion policy, while complex in its implementation, is definitely one of the most exciting and in the meantime one of the most risky businesses of the European Union. While audit and internal control efforts have experienced a tremendous increase, the error rate (the proportion of ineligible items) is still over 5 percent. However, stakeholders expect the control and audit explosion to bear fruit and result in smoother financial management over all the European Union.

Scholars primarily investigate the development of the internal control system of EU funds and the enhancing cooperation between the Court and the SAI's when searching for the remedy for the high error rate. Less attention has been paid to the potential of the existing IT resources to support controllers and auditors. It was detectable during interviews prepared at the Hungarian and even the EU level that the IT aspect, as a source of evidence for audit purposes, has not yet been fully exploited.

This author finds that a higher convergence between IT systems at the EU level would result in a more efficient audit system. As a first step, the right of access to core data stored in national databases opens up the possibility for auditors to draw conclusions on system assessment by the analysis of historical data. The increasing transparency over EU funds due to access to core data would shift Member States to a more efficient implementation of the budget, which could result in a diminishing error rate. This is not to say that an even stronger cooperation between SAIs and the development of the internal control system is not worthwhile. It should be emphasised that the IT aspect is only one of the building blocks of a complex system that could help secure EU citizens' interests in achieving the original goal of the Cohesion policy.

After having examined how the current IT background could contribute to a more developed and effective audit system of EU funds in general, for further research, it is beyond dispute worthwhile exploring how the functionality and interfaces of such an integrated IT system could be built up in such a way as to reflect the expectations of different players of the whole EU budget system.

## Acknowledgement

The article is based on the author's presentation made at the European Court of Auditors in Luxembourg on October 26, 2011. This work was supported by the Postgraduate Research Grant Programme created by the European University Institute (Historical Archives of the European Union) and the European Court of Auditors.

**References**


[1] Castells, A.: "External Audit Institutions: the European Court of Auditors and its Relationship with National Audit Institutions of the Member States", in: *Public Expenditure Control in Europe*, 2005, pp. 127-147

[2] Cipriani, G.: *"The EU Budget: Responsibility without Accountability?"*, Brussels, Centre for European Policy Studies, 2010

[3] Cipriani, G.: "The Responsibility for Implementing the Community budget", Brussels, Centre for European Policy Studies Working Document, No. 247/June 2006

[4] European Commission: "Synthesis of the Commission's Management Achievements in 2010", Communication to the European Parliament, the Council and the Court of Auditors, Brussels, 1.6.2011

[5] European Commission DG Regional Policy: "2010 Annual Activity Report", Brussels, 31st March 2011, http://ec.europa.eu

[6] European Court of Auditors: "Annual Report on the Implementation of the Budget", 2012a, http://eca.europa.eu

[7] European Court of Auditors: "Structural funds: Did the Commission Successfully Deal with Deficiencies Identified in the Member States' Management and Control Systems?", Special report, No. 3, 2012b, http://eca.europa.eu

[8] European Parliament, Directorate-General for Internal Policies: "What are the Implications of the Current Legislation for Cost Effectiveness and Quality Control in Structural Fund Spending? What Role for Performance Auditing?", Brussels, 2011

[9] Levy, R.: "Managing Value-for-Money Audit in the European Union: the Challenge of Diversity", Journal of Common Market Studies, 1996, Vol. 34, pp. 509-529

[10] Mendez, C. and Bachtler, J.: "Administrative Reform and Unintended Consequences: an Assessment of the EU Cohesion Policy 'Audit Explosion'", Journal of European Public Policy, August 2011, pp. 746-765

[11] Mendez, C. and Bachtler, J.: "Who Governs EU Cohesion Policy? Deconstructing the Reforms of the Structural Funds", Journal of Common Market Studies, 2007, Volume 45, Number 3, pp. 535-564

[12] Treaty on the Functioning of the European Union, available on http://eur-lex.europa.eu

[13] Weber, M.: "European Court of Auditors' Approach to the Audit of Cohesion Policy", Seminar on ECA DAS Methodology and Audit Missions in Member States, Budapest, 18 November 2010

# Benefits of Smart Grid Solutions in Open Electricity Markets

**Nenad Katic**

University of Novi Sad, Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia,
e-mail: nenad.katic@telventdms.com

*Abstract: Smart Grid Solutions include different technologies to improve the efficiency of electricity distribution network operations (e.g. distribution automation systems, smart metering, home automation and demand response solutions). The benefits arising from the implementation of distribution automation systems are investigated and evaluated in this article (e.g., the reduction of power losses and the reduction in outage costs and network development costs), taking into consideration the environment of an open electricity market. The investment costs and operation and maintenance costs for the deployment of such systems are analysed as well. Finally, a comparison of the benefits and costs is provided, and relevant conclusions to achieve profitability of the distribution automation systems are considered.*

*Keywords: Cost Benefit Analysis; Distribution Automation System; Open Electricity Market; Power Distribution; Smart Grid Solutions*

## 1 Introduction

In an open electricity market (a deregulated electricity industry sector), Power Distribution Companies (PDC) become “service providers”, as regulated Utilities, providing the “transport” of electric energy from the transmission side to end consumers (customers). On the other side, the sale of electric energy is provided by competitive “energy providers” or directly negotiated between large customers and producers. In such a new environment, PDCs have limited revenue and are firmly controlled by regulators and technical regulations on the quality and reliability of the electric energy supply.

The PDC’s service, as a monopoly, is a regulated business and the revenue comes from the “the charge for use of the system”. All system users (power consumers, customers) connected to the power distribution system pay for the service of energy transport to connection (consuming) point. The revenue of each PDC is controlled (regulated) by the regulator, applying one of the common methods in the deregulated electricity industry [1] – [4]:

a) *Rate of Return*: Annual revenue of PDC is established with reasonable operating costs and a certain return on asset approved by the regulator:

$$R_t = \sum_{i=1}^{n} p_i . q_i = \sum_{j=1}^{m} Cost_j + r.V_{assets} \quad (1)$$

$p_i$ – unit price of delivered service $i$ ,

$q_i$ – estimated number of service units $i$, to be delivered in a considered regulatory period,

$Cost_j$ – operating costs of class $j$,

$r$ –rate of return approved by regulator,

$V_{assets}$ – value of assets approved by regulator.

The disadvantage of this method is that there is not enough incentive for the regulated company to improve the efficiency of operation; rather the company focuses on approving assets with the regulator and increasing the rate of return.

b) *Price cap*: incentive method to improve the efficiency of operation, because it is impacted with "efficiency factor" set by regulator:

$$P_{i,t} = P_{i,t-1} . \left( 1 + \frac{RPI - X}{100} \right) \pm Z \quad (2)$$

$P_{i,t}$ – unit price of delivered service $i$ in the next year/ period $t$,

$P_{i,t-1}$ – unit price of delivered service $i$ in previous year/ period $t-1$

RPI – Retail Price Index in %,

$X$ – efficiency factor in %,

Z – corrective factor for external deviations.

If the efficiency factor is higher, then the PDC has a strong incentive to improve the efficiency of operation and produce profit, which stays in company since the regulatory period is normally several years.

c) *Revenue cap*: incentive and more sensitive method assuming changes of transported energy and number of customers:

$$R_t = \left( R_{t-1} + F_{client,t} . \Delta Client_t + F_{energy,t} . \Delta Energy_t \right) . \left( 1 + \frac{RPI - X}{100} \right) \pm Z \quad (3)$$

$R_t$ – maximal revenue in the next year/ period $t$,

$R_{t-1}$ – revenue in the previous year/ period $t-1$,

$F_{client,t}$ , $F_{energy,t}$ – correction factors assuming changes in the number of customers and delivered energy in period *t*,

$\Delta Client_t$ , $\Delta Energy_t$ – estimated changes in the number of customers or delivered energy in period *t*.

d) *Benchmarking*: the comparison of performance indices between similar regulated companies in a given region.

In the next step, the approved revenue is recalculated on "network usage tariffs" based on one of the common methods used in the deregulated electricity industry:

a) *Postage stamp method*: in this simple method, all transportation costs (regulated revenue) are distributed on service units (delivered energy, demand power, connection power, etc.) for all network users, regardless of the point of connection. The same network tariff is paid, only multiplied by the quantity of delivery.

b) *MW mile method*: the tariff depends on the distance between the points of generation and consumption, but not on the real impact of the transaction on the system (load flow).

c) *Modulus of use*: in this case, load flow calculation results are accounted, and the real impact of each transaction on the system.

The regulatory methods commonly applied in PDCs are *Rate of Return,* as a simpler method for young electricity markets (e.g. Serbia [5]), and *Revenue Cap* as a more complex method for developed markets (e.g. Portugal, Norway [6], etc.), whereas tariffs in distribution are normally calculated with the simple *postage stamp method*.

In any case, the PDC is responsible for covering power losses in the distribution network, and only a certain level of losses is accounted in approved costs. If the PDC operates with higher losses than approved, the additional costs can be paid only from PDC's profit; but if it operates with lower losses, the PDC's profits will increase. Regarding the reliability of supply, it is penalized either by incentive revenue cap methods or by direct compensation to customers under certain conditions. If the PDC operates with higher efficiency and better reliability indices, additional profit will be generated; but with worse reliability, profit will be reduced.

Under the pressures of regulated and limited revenue, the PDCs in the deregulated electricity industry sector have to invest in advanced technologies, known as *Smart Grid Solutions*, in order to improve the quality and reliability of supply and to reduce power losses in the distribution power lines. Smart Grid Solutions should improve the efficiency and reduce the costs of network operation. The real challenge is the quantification of the benefits provided by Smart Grid Solutions and the comparison of these benefits with related costs, in order to estimate the real profitability of such investments, which is the main subject of this article.

## 2 Smart Grid Technologies

*Smart Grid Solutions* (SGS) encompass different technologies targeted to improve the efficiency of the distribution network operation:

- Distribution Automation Systems (DAS),
- Advanced Metering Infrastructure (AMI) and remote meter reading and data management,
- Smart power equipment with improved operation characteristics,
- Distributed Generation management, Energy Resources management (e.g. battery storages, electrical vehicles) and Demand Response systems, which can affect distribution network operation.

*Distribution Automation Systems (DAS)* implementation benefits and costs are analysed in this article. DAS typically consist of the following modules:

- Centralized Control System Software, including basically three systems: SCADA (Supervisory Control and Data Acquisition System), DMS (Distribution Management System) and OMS (Outage Management System), integrated in the solution.
- Centralized Control System Hardware, including typically Main Data Center servers, Disaster Recovery Data Canter servers and User workstations.
- A communication system to support high speed transfer of data over wide areas.
- Field execution system – remotely controlled equipment in field (load breaking switches and regulators with motor drives) with a communication subsystem (remote terminal units and radio or line communication modems).

The central intelligence of DAS is the *Distribution Management System* (DMS), a (mathematical) modeling of the power network, a set of calculation engines and algorithms to provide network operation optimization, analysis, reporting and long-term planning. The DMS calculation results applied in field by communication and field devices can provide important targets to the PDCs:

- Reduction of power losses, optimizing power flow and voltages in power network,
- Improved reliability of supply, reducing outage times with advanced Fault Management technologies,
- Improved power quality, with optimal Voltage/Var regulation,
- Overall improvement of customer services,
- Significant revenue increases.

Advanced DMS provides an efficient user interface for power network control, monitoring and visualization (schematic and geographical), comprehensive technical database, real-time and simulation (study) environments, historical data access and storages, web technologies, as well as a wide set of power functions:

- *Fundamental power functions* for the calculation of power flow and the network operation state, which are typically "resident" functions automatically triggered by the software system,
- *Network operation functions* for switching scheduling and execution, management of faults and emergency situations, typically used by operators in the control room for the real-time network operation,
- *Network optimization functions* for the optimization of power flow and voltages, typically used by engineers in support of the control center, for short-term operation planning,
- *Network analysis and Reporting functions* for the analysis of short-circuits, relay protection settings, contingency problems, equipment rating, and harmonic pollution, and to provide reports about outages, power losses, etc., typically used by engineers in simulation mode,
- *Network Development functions* for (medium-term) maintenance planning and long-term analysis of the network design, construction and development, typically used by engineers in investment and planning departments.

# 3 DAS Benefits

Benefits, achieved in one PDC after DAS implementation are analysed and evaluated in this section. The benefits are analysed for one example of a live distribution network (DN) presented in Figure 1. The DN is supplied by one High-Voltage (HV) Substation, 110/35 kV, then by four 35 kV power lines supplying four Distribution Primary Substations (DPS) at 35/10 kV, and by 40 outgoing 10 kV feeders. The medium voltage (MV) 10 kV distribution network has 350 kilometers of underground and overhead power lines, 360 Distribution Secondary Substations (DSS) at 10/0,4 kV with a total of 90 MVA transformer power. The DN supplies 25,000 consumers (customers). The peak power of the DN is approximately 50 MW, with injected electrical energy annually into the DS (EEIA) of approximately 200 GWh, on the 110 kV (purchase) side.

The PDC is a regulated company and revenue is approved by a regulator, using, for example, the *revenue cap* method. The PDC earns revenue by charging consumers for the use of the distribution system, applying tariffs calculated with the *postage stamp method.* Customers pay a specific tariff according to consumption type, based on the quantity of consumed kWh, regardless of the connection point location. According to the definition of the *revenue cap* method

(3), tariffs are calculated for one regulatory period (normally 2 to 3 years) and highly impacted by *efficiency factor X*, applied to motivate the PDCs to improve efficiency of operation. If the PDC operates more efficiently and decreases costs during one regulatory period, all profit will belong to the PDC. Assuming that the PDC initially complies with *efficiency factor X*, any decrease of operation cost will be a *profit* to the PDC in one regulatory period. On the other hand, if the PDC is initially over *efficiency factor X*, incurring business losses, any decrease in operation cost will be a benefit to the PDC in one regulatory period. The target of this analysis is to evaluate the DAS benefits (profits) in relative units compared with the PDC's revenue or amount of EEIA.

Electric energy injected into the DN annually (EEIA) in this example is 200 GWh. Power losses in the considered DN are approximately 10 %, and therefore energy sold (billed) to customers is 180 GWh. Customers pay to the PDC a network charge for use of the distribution system and electric energy consumed to energy providers. The PDC's annual revenue (PRA) is based on the network charge, which in this example is approximately **$5,400,000** annually (an average of 3US¢ per kWh delivered). With PRA, the PDC has to purchase electric energy for power losses (20 GWh or $600,000 annually with $30.00/MWh), cover all operation costs and generate a certain profit for shareholders. Any reduction of power losses or operation costs would directly increase the PRA and profit.

DAS benefits can be differentiated as:

a) Reduction of power losses;
b) Reduction of operation costs (reduced maintenance, outage time and non-supplied energy);
c) Reduction of network construction costs (improved utilization of facilities and postponement of investment);

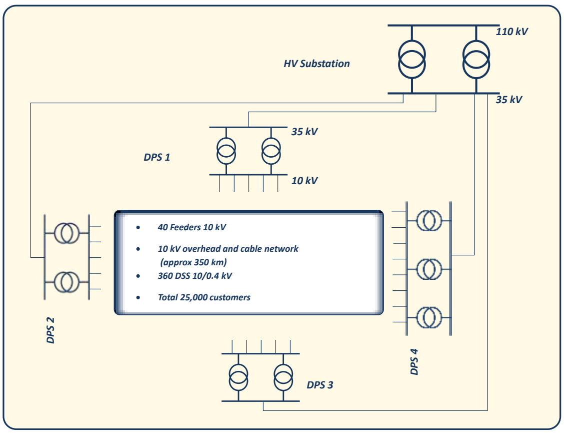



Figure 1
The example of a live Distribution Network

## 3.1 Reduction of Power Losses

Typical breakdowns or power losses in the distribution network are presented in Fig. 2.

*Optimal Feeder Reconfiguration:* Power flow calculation [7]-[9], based on the real-time state estimation of the power distribution network and load profiling [10]-[12], results in power balance of network operation state, for all transformers, feeders and branches. The distribution network is operated without loops, meaning that each feeder has a certain number (or none) of “normally open switches” (NOS), providing eventual restoration of supply in the case of incidents, but normally open in the regular operation state. In large distribution networks with hundreds of feeders, there could be hundreds of NOS, providing thousands of combinations of NOS locations to open all loops. Changing the location of a NOS and keeping the network without loops is actually changing of the power flow and operating performances (loading, power losses, voltages). The target is to find the “optimal” combination of NOS locations to achieve an optimal power flow according to certain objective (minimum power losses, best voltages, reduced overloads, etc.). The typical DMS function “Optimal Feeder Reconfiguration” provides such a solution [13]-[14]. If the results are applied in the network operation, a 20% reduction of MV network power losses may be achieved, with insignificant costs of changing NOS locations [15]. Since MV network power losses are approximately 25% of total power losses, the reduction of total power losses would be 5%. In terms of energy, total energy losses would be reduced 3 to 4%, whereas the reduction in EEIA would be in the range of 0.3 to 0.4 %.

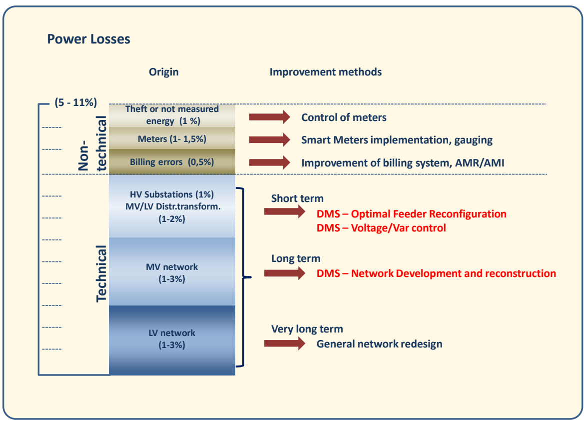



Figure 2
Breakdown of power losses

*Voltage and Var Control:* DMS solutions can provide centralized (optimal) control of voltages and reactive power flow in the distribution network [16]-[19]. Based on the real-time network operation state, the DMS function "Voltage/Var Control" will calculate the "optimal" setting of the voltage regulation devices and the switching of capacitor banks applied in the distribution network, according to certain objectives (minimal voltage deviation, minimal power losses, minimal or maximal power demand, etc.). If the "minimal power losses" objective is applied, test results give an 8 to 10% reduction in MV network power losses, whereas the reduction of total power losses would be approximately 2%, and the reduction of EEIA in the range 0.1 to 0.3%.

Finally, the DMS functions "Optimal Feeder Reconfiguration" and "Voltage/Var Control" have to be coordinated in the field since both have an influence on the network operation state, improving voltages and reducing losses. The cumulated effect of both functions normally leads to a 25% reduction in MV network power losses, giving an approximately 6% reduction in total power losses and a 0.5% reduction in EEIA.

In the DN example, the PDC will annually purchase 6% less energy for power losses (1.2 GWh), which is a savings of $36,000 annually, or 0.67% of PRA.

## 3.2 Reduction of Network Operation Costs

Nowadays, DAS is applied in a small number of utilities. Incidents (faults) and outages are resolved with a crew in the field and with a manual operation of switches, as presented in Figure 3.

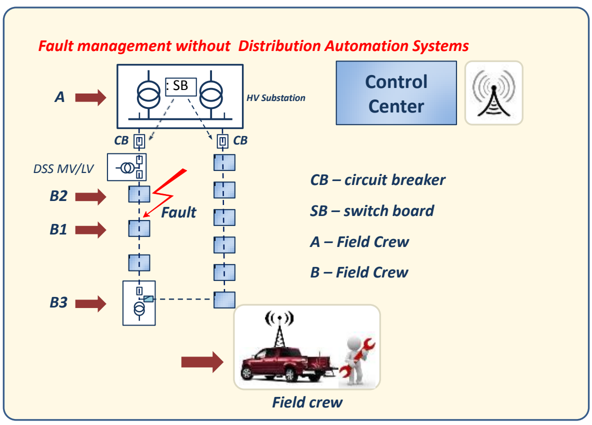



Figure 3
Fault Management without automation

In the case of a fault on the MV feeder, one crew (A) is sent to a substation to check protection and operate feeder circuit breakers, while the second crew (B) is sent to check the fault point along the MV feeder and to manually operate the MV switches. After several operations of the circuit breaker and MV switches, the fault will be localized (B1 , B2, …) and supply restored to non-faulty parts of the MV feeder (B3, …). The whole procedure will take several hours, sometimes much longer, while customers along the feeder will suffer a long outage. Additionally, there is a high risk of wrong switching and non-optimal decisions because the operator has limited information about the fault and power loadings in the network and relies only on his experience, without technical tools to support his decisions.

In the example of the DN, the average MV feeder loading is 1 MW, and non-supplied energy is 2 MWh per average fault (2 h). An average failure rate for MV lines is in the range of 0.2 (cables) to 0.5 (overhead lines) faults/km, annually; therefore in the DN (360 km), there are approximately 150 faults/year. Energy non-supplied (ENS) is approximately 300 MWh annually.

*PDC direct outage costs:* In the example of the DN, the PDC annual costs related to resolving faults in the MV network are as follows:

- Revenue lost $9,000 (ENS=300 MWh x 3US¢/kWh),
- One Breaker failure (300 switching on fault) ≈ $5,000,
- 4 load switches failures (800 switching) ≈ $16,000,
- Field crew costs (800 x $20.00 /switching) ≈ $16,000,

Total PDC direct outage costs ≈ $46,000.

*Customers outage costs:* Customers sustain damages due to power outages, the cost of which is estimated at 5 times (for households) up to 100 times (for sensitive industrial customers) the price of the delivered kWh. In the example of the DN, the customers are 80% residential and 20% industrial, and therefore the customers outage costs will be on average 20 times the price of delivered kWh. If customers are paying $0.08 to $0.10/kWh for delivered kWh (network charge and energy price), the total damage to customers in the example of the DN would be $480,000 annually (300 MWh x$1.60/kWh). Customers are not paid by the PDC directly for such damage, but this amount presents the level of real (social) damage.

*Penalties paid by the PDC:* The liberalization of electrical sector introduced penalties to be paid to customers in the case of outages. For example, according to regulations in Holland, utilities pay compensation of EUR 30.00 (US$40.00) to every customer suffering an outage longer than 4 hours. In the example of the DN, 10% to 15% of outages are longer than 4 hours, and therefore PDC should pay compensation for 20 outages every year. Since each feeder has 625 customers on average, the PDC should pay approximately $500,000 annually in compensation to customers. This figure is close to the real customer damage calculated in the previous chapter and demonstrates that individual compensation is estimated realistically.

*Implementation of DAS:* DAS would significantly improve the reliability of network operation. In the DAS project, the PDC would implement the SCADA/DMS/OMS system with remote control of each HV Substation and with MV load-switches and fault detectors at critical locations of the distribution network (Fig. 4). In this case, fault management would be significantly improved:

- HV Substations are remotely controlled and there is no need for crew A to go to the HV Substation any more,
- Critical locations in the MV distribution network are remotely controlled and equipped with fault detectors, so fault location is automated without reclosing the circuit breaker and many operations of load switches,
- If necessary, one crew (A) will go along the feeder to provide manual switching of non-remotely controlled switches and to help with supply restoration,
- Dispatchers will have DMS software for the estimation of power flow and feasible supply restoration sequence, with very small risk of wrong decisions or wrong switching steps [20],
- Outage duration will be 5 times shorter, down to 15 to 20 minutes (and if there are more RTUs in the MV network, even shorter); in addition, the ENS will decrease 5 times,
- The PDC's outage costs will be reduced because of a reduced number of switching and equipment failures in the field.

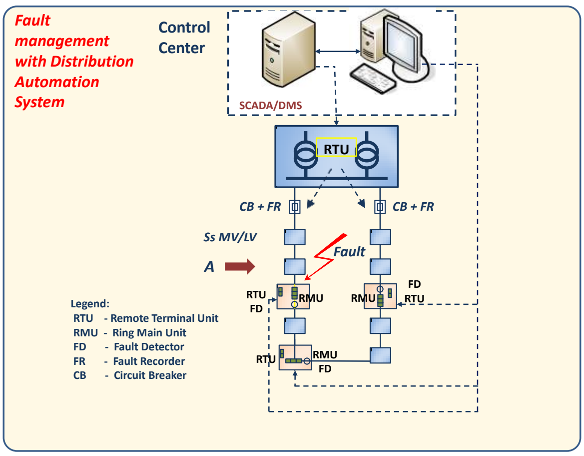



Figure 4
Fault Management after automation

PDC direct outage costs would reduce as follows:

- Revenue lost $1,800 (ENS = 60 MWh x 3US¢/MWh),
- Breaker failures once in 2 years ≈ $2,500,
- One switch failure (300 switching) ≈ $4,000,
- Filed crews costs (300 x $20 /switching) ≈ $6,000.

Total PDC direct outage costs ≈ $14,300

Customer outage costs would be reduced 5 times as well, due to the reduced outage time, as well as the reduced penalties paid to customers (down to $100,000 annually).

In the DN example, savings after the DAS implementation would be in the reduced outage costs:

- PDC direct outage costs would be reduced 70%, providing a savings of $31,700or 0.58% of the PRA.
- Penalties paid by the PDC (if applied) would be reduced 80%, providing a savings of $400,000 or 7.4% of the PRA.

## 3.3 Reduction of Network Construction Costs

DMS tools for network operation optimization and long-term planning analysis enable more efficient utilization of power distribution facilities and the postponement of investments in network constructions up to 20%.

The construction of distribution facilities, due to new customer connections or resolving of power overloads, increases the number of distribution transformers 1% annually on average, as well as the length of power lines by 0.5%. In the example of the DN, 4 new DSS and 2 kilometers of MV power lines should be built every year on average; however, with DMS, investment would be postponed and reduced to 3 DSS and 1.6 km of MV power lines, giving a savings of approximately $20,000annually.

*DMS Large Area Restoration* functionality enables the efficient planning of large HV supply transformer outages. The capacities of adjacent HV Substations and the available resources in the MV network are (more) efficiently utilized, consequently postponing the construction of one new HV Substation. In the example of the DN, one new HV Substation should be built every 10 years ($1,000,000 investment); however investment can be postponed and save $ 30,000 of interest.

The total reduction of network construction costs would save up to $50,000 annually, or 0.92% of the PRA.

### 3.4 The Total DAS Benefits

The total benefits after DAS implementation, as discussed in previous chapters, would provide an annual savings (profit) to the PDC of:

- *2% of the PRA*, if the PDC is not paying compensation to customers because of outages, or up to
- *9% of the PRA*, if the PDC is paying the full compensation to customers because of outages.

# 4 Investment in Distribution Automation

Investment in DAS encompasses four sectors [21]:

- *SCADA/DMS Software,* deployed in Distribution Control Centre (DCC), including the creation of a network model and database, operation network diagrams (schematic and geographic), delivery, training, testing and commissioning of the system.
- *SCADA/DMS Hardware,* deployed in DCC with (redundant) servers, communication front-end, workstations with several displays, GPS, printers, web servers, Large Screen Display, delivery, training, testing and commissioning of the system.
- *Remote terminal units (RTU),* implemented in HV substations and field locations (some of MV/LV substations or pole mounted switches), equipment for signal conversion (transducers, relays), ancillary equipment (power supply, relays, cabling), with assembling, delivery, training, testing and commissioning of the system.
- *Communication system* (radio, fiber optic, telephone leased lines), including system design, licenses for radio frequencies, delivery of equipment, delivery, training, testing and commissioning of the system.

The principle of distribution (MV network) automation is presented in Figure 5 (minimum automation to achieve a significant improvement in fault management); however detailed engineering analysis must be realized in every specific project.

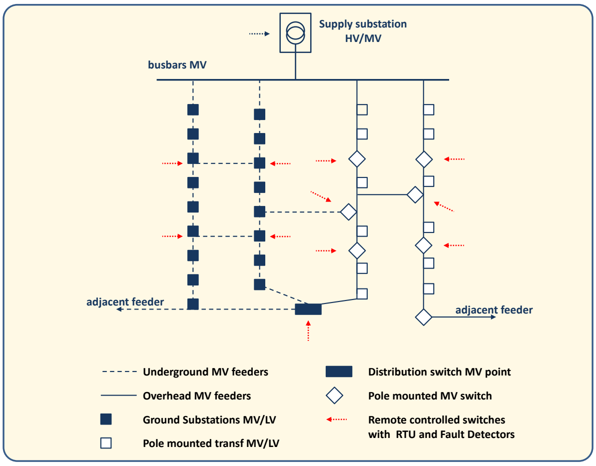



Figure 5
Principle of distribution network automation

More RTUs in the MV network will bring better results but will rapidly increase investment. Some of the DMS functions may be used for the detailed analysis and design of the optimal level of automation, keeping investment inside a limited budget. In the DN example, the following results were obtained (Figure 6):

- Each of (four) Substations 35/10 kV should be automated (remotely controlled by the SCADA/DMS system),
- Approximately 25 DSS 10/0.4 kV (25% of the total number) should be automated with ring main units (RMU), RTUs and Fault Detectors. The DSSs are located on 12 underground feeders, which averages 2.5 DSS per underground feeder,
- Approximately 70 pole mounted switches (PMS) should be automated with RTUs, motor drive, local control box and Fault Detectors. PMS are located on 28 overhead (10 kV) feeders, which averages out to 2.5 PMS per overhead feeder,
- All three of the existing distribution switching yards should be automated with RMU and RTU.

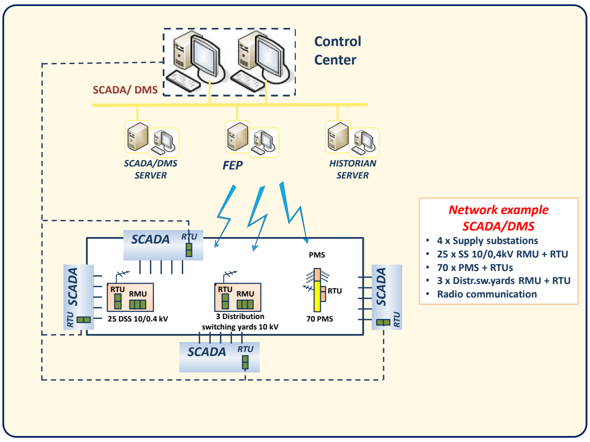



Figure 6
Optimal automation of DN

The calculation of the investment costs is presented in the Table 1.

In reality, every utility has already invested in certain automation equipment that can be included in DAS (e.g. SCADA, HV substations RTUs, and communication systems); consequently, DAS investment will actually be lower. On the other hand, there is always a certain amount of old power equipment which should be replaced in order to be included in the automation system, which will increase the investment (e.g., old manually operated air-breaking switches in the DSS should be replaced with advanced RMUs with SF6 gas and motor drives). Therefore, for the simplicity of this analysis, it is assumed that reduced investment due to existing automation equipment, equals investment in the new power equipment. The total investment in DAS in the network example is approximately *$1 million.*

Table 1
DAS investments in the example of DN

| DAS description | Unit price ($) | No of units | Total price ($) |
|---|---|---|---|
| SCADA/DMS Software | | set | 200,000 |
| Hardware in DDC | | set | 100,000 |
| Substations 35/10 kV, RTU, cabling and interfaces | 50,000 | 4 | 200,000 |
| Ground substation 10/0,4 kV, RTU, cabling and interfaces | 5,000 | 25 | 125,000 |
| Pole mounted RTU and interfaces | 3,000 | 70 | 210,000 |
| Distribution switching yard, RTU and interfaces | 10,000 | 3 | 30,000 |
| Communication system: FO (interfaces) or radio system (equipment, interfaces, antennas) | 1,500 | 100 | 150,000 |

| **Total DAS investment** | | | **1,015,000** |
|---|---|---|---|
| The new power equipment: RMU in ground substations | 8,000 | 28 | 224,000 |
| The new power equipment: PMS load breakers | 4,000 | 70 | 280,000 |
| Total DAS investment with replacement of power equipment | | | 1,619,000 |

The average DAS investment costs can be presented:

- ***$20.000/MW*** of the network peak load (the peak load in the example of DN is 50 MW, with a necessary investment in DAS of $1 million),
- ***$40/customer (meter)*** (2 kW/customer is the participation of individual customer load in the network peak load),
- ***18% of the PRA*** (the PDC has an annual revenue of $5,400,000in the example of the DN, with necessary investment in DAS of $1 million)

# 5 Cost Benefit Analysis

The cost/benefit analysis is made on an example of a large PDC. If a large PDC has 1000 MW peak load, then EEIA would be approximately 4000 GWh/year, with 3600 GWh on the consumption side. The annual revenue of such a company (PRA) would be approximately $108 million (M$). If such company decides to invest in deployment of distribution automation system, total investment costs would be close to 20 M$, with a project execution time of 2 years. Annual operation & maintenance costs (O&M) are typically from 2% to 3% of the investment cost and, adding 5% for depreciation, the total annual cost of system operation would be close to 8% of the investment cost, or approximately 1.6 M$/year. The PDC pays a certain compensation for power outages to customers, which has significant impact on the level of the DAS benefits, as discussed in chapter III of this article. Annual DAS benefits are 2% when compensation is not paid and up to 9% of the PRA when the full compensation is paid to customers. The sensitivity of DAS benefits due to the level of compensation paid to customers will be considered.

The considered period of system operation is 10 years, which is close to the life time of the SW, HW, RTU and communication equipment. In the first step, for simplicity of analysis, actualization will be neglected. In the first year of operation, DAS will start providing savings, but the full benefits will be reached in the second year of operation.

In the cost/benefit analysis, the three following economic factors are calculated:

$$C/B = {C_t}/{B_t} \tag{4}$$

*C/B* – Cost/benefit factor,

$C_t$ – Total cost of investments and operation in a 10-year period of consideration,

$B_t$ – Total benefits in 10 years of operation.

$$P = {B_t}/{C_t} \tag{5}$$

*P* – Profitability of the project, as a ratio of total benefits and costs in 10 years of operation

$$R = {C_i}/{B_t} \cdot T \tag{6}$$

*R* – Payback period, the period of time necessary for return of investment.

*Ci* –Investment cost,

*T* – Period of consideration, in this case 10 years.

The Cost/Benefit analysis results (without actualization) are presented in Table 2 and in Figure 7.

The total cost in 10 years of operation would be 36 M$ (an investment of 20 M$ and O&M costs of 16 M$), whereas the total benefit would be in the range from 19 M$ (no compensation paid to customers for outages) to 86 M$ (the full compensation paid to customers). The DAS project will become profitable when benefits reach 4% of the PRA, which will happen when a certain (small) level of outage compensation is paid to customers by the PDC. If outage compensation is paid to the full extent, then the DAS project will have the high profitability of 2.39 and a payback period of only 2.33 years.

Table 2

Cost/Benefit analysis results - without actualization

| | Total (M$) | C/B | P | R (years) |
|---|---|---|---|---|
| Investment | 20.00 | | | |
| O&M | 16.00 | | | |
| Benefit 2%PRA | 19.00 | 1.89 | 0.53 | 10.53 |
| Benefit 3%PRA | 29.00 | 1.24 | 0.81 | 6.90 |
| Benefit 4%PRA | 38.00 | 0.95 | 1.06 | 5.26 |
| Benefit 5%PRA | 48.00 | 0.75 | 1.33 | 4.17 |
| Benefit 6%PRA | 58.00 | 0.62 | 1.61 | 3.45 |
| Benefit 7%PRA | 67.00 | 0.54 | 1.86 | 2.99 |
| Benefit 8%PRA | 77.00 | 0.47 | 2.14 | 2.60 |
| Benefit 9%PRA | 86.00 | 0.42 | 2.39 | 2.33 |

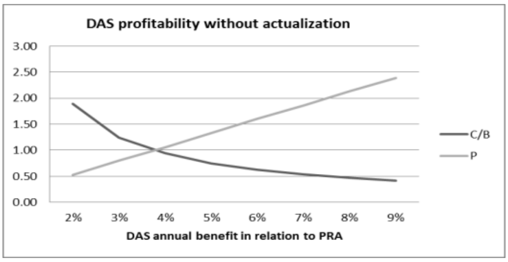



Figure 7

Profitability of DAS project without actualization

However, if actualization is not neglected, then annual benefits and costs have to be derived (recalculated) on the initial year using the actualization rate. The initial year is the first year of the system operation:

$$C_{t,a} = C_i + \sum_{i=1}^{10} \frac{0{,}08 \cdot C_i}{(1+p)^i} \tag{7}$$

$C_{t,a}$ – Total cost of investments and operation in a 10-year period of consideration, derived in the initial year.

$p$ – actualization rate, normally 6%

$$B_{t,a} = \sum_{i=1}^{10} \frac{B_i}{(1+p)^i} \tag{8}$$

$B_{t,a}$ – Total benefits in 10 years of operation, derived on the initial year.

$B_i$ – annual benefit in the year "$i$"

The results of the Cost/Benefit analysis with an actualization rate of 6% are presented in Table 3 and in Figure 8.

Table 3

Cost/Benefit analysis results - with actualization 6%

| | Total (M$) | C/B | P | R (years) |
|---|---|---|---|---|
| Investment | 20.00 | | | |
| O&M | 11.78 | | | |
| Benefit 2%PRA | 13.78 | 2.31 | 0.43 | 14.52 |
| Benefit 3%PRA | 21.14 | 1.50 | 0.67 | 9.46 |
| Benefit 4%PRA | 27.55 | 1.15 | 0.87 | 7.26 |
| Benefit 5%PRA | 34.91 | 0.91 | 1.10 | 5.73 |
| Benefit 6%PRA | 42.27 | 0.75 | 1.33 | 4.73 |
| Benefit 7%PRA | 48.69 | 0.65 | 1.53 | 4.11 |
| Benefit 8%PRA | 56.05 | 0.57 | 1.76 | 3.57 |
| Benefit 9%PRA | 62.47 | 0.51 | 1.97 | 3.20 |

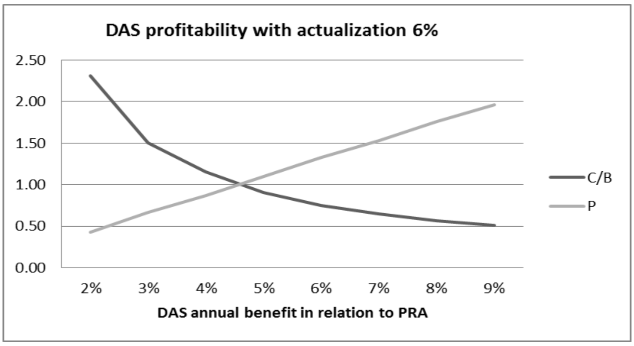



Figure 8
Profitability of DAS project with actualization

The total cost in 10 years of operation would be 31.78 M$ (an investment of 20 M$ and O&M costs of 11.78 M$), whereas the total benefit would be in the range from 13.78 M$ (no compensation paid to customers for outages) to 62.47 M$ (the full compensation paid to customers). The DAS project will become profitable when benefits reach 4.5 % of the PRA, which will happen when a medium level of outage compensation is paid to customers by the PDC. If outage compensation is paid to the full extent, then the DAS project will have the high profitability of approximately 2 and a payback period of 3.2 years.

According to the cost/benefit analysis presented in this article, it can be concluded that DAS projects are profitable investments. Profitability depends mostly on the level of compensation which the PDC is paying to customers for non-planned outages in power network operation:

- If the PDC is not paying any compensation to customers for outages, which is unlikely, the DAS projects would hardly be profitable, but accounting other non-financial benefits, could be close to profitability,
- If the PDC is paying a certain amount, but not the full compensation, to customers for outages, the DAS projects will be profitable in the life time of the system operation, with profitability between 1 and 1.5 and a payback period between 4 and 5 years.
- If the PDC is paying appropriate (almost full) compensation to customers for outages, the DAS projects will be highly profitable, with profitability close to 2 and a payback period of approximately 3 years.

However, since the cost/benefit analysis presented in this article considers only the DAS benefits which can be precisely financially evaluated, but not all other benefits which cannot be precisely financially calculated (e.g. customer satisfaction, improvements in data management and reporting, improved simulation analysis, better setting of equipment, etc.), profitability should be even higher than calculated.

## Conclusions

Distribution Automation Systems (DAS) are advanced Smart Grid Solutions which can significantly improve the efficiency of operation of electricity distribution networks. In deregulated and open electricity markets, the Power Distribution Company (PDC) is considered as a service provider, the revenue of which comes from charges for use of the distribution network. With limited revenue, PDCs must invest in advanced technologies, known as Smart Grid Solutions, to improve the quality and reliability of supply and to reduce power losses. However, the quantification of the benefits provided by Smart Grid Solutions and the comparison with related costs is the real challenge nowadays, necessary in order to estimate the real profitability of such investments. The benefits achieved via applying DAS are analysed and evaluated in this article, using an example of a real distribution network. The DAS benefits calculated in this way are compared with the PDC's annual revenue (PRA). If the PDC is not paying any compensation to customers because of incidents, benefits are 2% of the PRA, but if the PDC is paying compensation benefits, they are up to 9% of the PRA.

Cost/Benefit analysis, made on the example of a large Power Distribution Company, reveals a good profitability from DAS implementation in a 10-year lifetime and system operation. However, since analysis cannot consider all possible benefits, because some of them cannot be precisely financially evaluated, the real profitability of the DAS implementation should be even higher. The general conclusion is that distribution automation projects are profitable investments.

## References

[1] L. Philipson, H. Lee Willis, Understanding Electric Utilities and De-Regulation (Marcel Dekker Inc., 1999)

[2] S. Stoft, Power System Economics (Wiley, 2002)

[3] D. Kirschen, G. Strbac, Power System Economics (John Wiley & Sons, 2004)

[4] N. Katic, V. Borozan, S. Halilcevic, Power System Economics (EU project Cost Effective and Environmental Friendly Energy Systems – CEFES TEMPUS Programme CD_JEP-18126-2003, 2007)

[5] Energy Agency of Republic of Serbia, Access to and Use of System Charging Methodology for Electricity Distribution (http://www.aers.rs/Files/Metodologije/Preciscen/Eng/2008-12-22_MetodologijaDistribucijaEEPrecisceniTekst.pdf , 2010)

[6] H. T. Ylvisaker, The future Network Company – a Norewegian perspective, Distributech Europe Conference, November 2000, Vienna, Austria

[7] D. Sirmohammadi, H. W.Hong, A. Semlyen, G. X. Luo, A Compensation-based Power Flow Method for Weakly Meshed Distribution and Transmission Networks, IEEE Trans. on Power System, Vol. 3, No. 2, May 1988, pp. 753-762

[8] S. Cheng, D. Shirmohammadi, A Three-Phase Power Flow Method for Real-Time Distribution System Analysis; IEEE Trans. on Power System, Vol. 10, No. 2, May 1995, pp. 671-679

[9] G. X. Luo, A. Semlyen, Efficient Load Flow for Large Weekly Meshed Distribution and Transmission Network; IEEE Trans. on Power System, Vol. 5, No. 4, November 1990, pp. 1309-1316

[10] I. Roytelman; S. M. Shahidehpour, State Estimation for Electric Power Distribution Systems in Quasireal-time Conditions, IEEE Trans. on Power Delivery, Vol. 8, No. 4, October 1993, pp. 2009-2015

[11] M. E. Baran, A. W. Kelley, State Estimation for Real-Time Monitoring of Distribution Systems in Quasi Conditions; IEEE Trans. on Power System, Vol. 9, No. 3, August 1993, pp. 1601-1609

[12] C. N. Lu, J. H. Teng, W.-H. E. Liu, Distribution System State Estimation; IEEE Trans. on Power System, Vol. 10, No. 1, February 1995, pp. 229-240

[13] I. Roytelman, V. Melnik; S. S. H. Lee, R. L. Lugtu, Multi-Objective Feeder Reconfiguration by Distribution Management System; IEEE Trans. on Power System, Vol. 11, No. 2, May 1996, pp. 661-667

[14] M. E. Baran, F. F. Wu, Network Reconfiguration in Distribution Systems for Loss Reduction and Load Balancing, IEEE Trans. On Power Delivery, Vol. 4, No. 2, April 1989

[15] Giorgio di Lembo, Paola Petroni, Christian Noce, Reduction of Power Losses and $CO_2$ Emission: Accurate Network Data to Obtain Good Performances of DMS Systems, CIRED 20$^{th}$ Conference Prague, June 8-11, 2009, Prague, Czech Republic

[16] I. Roytelman, B. K. Wee, R. L. Lugtu, Volt/Var Control for Modern Distribution Management System, IEEE Trans. on Power System, Vol. 10, No. 3, August 1995, pp. 1454-1460

[17] I. Roytelman, B. K. Wee, R. L. Lugtu, T. M. Kulas, T. Brossart, Pilot Project to Estimate the Centralized Volt/Var Control Effectiveness, IEEE Trans. on Power System, Vol. 13, No. 3, August 1998

[18] V. Strezoski, N. Katic, D. Janjic, Voltage Control Integrated in Distribution Management Systems, Electric Power Systems Research, No. 60, 2001, pp. 85-97

[19] A. Hamouda, K. Zehar, Improvement of the Power Transmission of Distribution Feeders by Fixed Capacitor Banks, Acta Polytechnica Hungarica, Vol. 4, No. 2, 2007, pp. 47-62

[20] D. S. Popovic, R. M. Ciric, A Multi-Objective Algorithm for Distribution Networks Restoration, IEEE Trans. on Power Delivery, Vol. 16, No. 3, July 1999, pp. 1134-1141

[21] N. Katic, V. Strezoski, D. Popovic, Business Benefits of DMS Software Application in Competitive Distribution, CIRED 17$^{th}$ Conference Barcelona, Session 6, 2003, Barcelona, Spain

# An Approach for e-Commerce On-Demand Service-oriented Product line Development

**Azubuike Ezenwoke[1], Sanjay Misra[2], Mathew Adigun[3]**

[1]Department of Computer and Information Sciences, Covenant University, Ota, Nigeria, azu.ezenwoke@covenantuniversity.edu.ng

[2]Department of Computer Engineering Atilim University, Ankara, Turkey, smisra@atilim.edu.tr

[3]Department of Computer Science, University of Zululand, Kwadlangezwa, Kwa-Zulu Natal, Republic of South Africa, madigun@pan.uzulu.ac.za

*Abstract: The growth of Small, Medium and Micro Enterprises (SMMEs) is important to the economic development of Africa. This growth can be greatly enhanced by leveraging IT in business activities since e-commerce is a vital tool to allow participation in globalization. Many SMMEs cannot afford to own e-commerce facilities and to reduce cost. An SMME can pay for just the e-commerce facility they use without owning the services or infrastructure. Due to the dynamic nature of the business domain, delivering such on-demand functionalities involves high flexibility in adapting to new client requirements; therefore, a systematic approach to software component reuse must be adopted to reduce cost and the time to market for new products. This work explores the reuse capabilities of a hybridization of Service Oriented Architecture (SOA) and Software Product Line (SPL).*

*Keywords: E-commerce; Service Oriented Architecture; Software Product line; and Small Medium and Micro Enterprises*

## 1 Introduction

The emergence and growth of the World Wide Web (WWW) has greatly impacted businesses in various industry sectors, changing the way business is done. An aspect of business that has been largely influenced by the WWW is the sales and purchases of goods and services. This is referred to as electronic commerce (e-commerce). In the emerging global economy, e-commerce is an important tool for business development, with the potential to increase the market reach of businesses and to improve business efficiency, and, therefore, it is a veritable catalyst for economic growth [1]. In Africa, there exists some evidence of e-commerce adoption in large organizations, but there seems to be insignificant use of e-commerce tools among Small, Medium and Micro Enterprises (SMMEs).

Some of the reasons for the low adoption in Africa are a lack of Information and Communication Technology (ICT) expertise and an inability to afford the total cost of ownership of e-commerce infrastructures and services [2]. To accelerate economic growth, adequate investment must be made to support the growth of SMMEs and to leverage ICT in business operations. SMMEs should be able to afford ICT to facilitate e-commerce by paying for e-commerce services on-demand (EoD).

On-Demand Computing is a paradigm that facilitates the availability of computing resources to users on request. Utility computing is a type of on-demand computing that enables resource provisioning through payment models, such as subscription or pay-as-you-use. The term "Utility" is derived from real world provision of utilities, such as electricity, water and gas, where consumers do not own major infrastructures but pay for the resources used. In enterprise computing, software or hardware resources are accessible, as services, over a network, at the user's request. This mode of software delivery would reduce the acquisition and operations cost of ICT infrastructures and services, such as Customer Relationship Management, on-line Payment Processing, Report Generation and Analysis, Order Management Systems, etc.

The business environment is characterized by high frequency of change. Changes could result from mergers and acquisitions, new market opportunities, new customer demands, government policies, and/or technological advancements. To maintain competitiveness, SMMEs must adapt to these changes. For an ICT-enabled enterprise, adapting to these changes means acquiring newer ICT solutions as quickly as possible, which poses a challenge to e-commerce solution developers. Traditional methods of building software from scratch cannot meet the rate of demand; hence, in response, developers must adopt newer approaches to achieve quick, flexible and cost effective application development and deployment. This quest gives rise to the concept of Software Reuse. Reuse is widely used in engineering disciplines where systems are designed by putting together existing components. In software engineering, reuse can take the form of design for reuse or design with reuse. Software reuse is the process of constructing software systems from pre-built software units instead of building software systems from scratch [3]. A crucial research question explored in this paper is: *How can we build e-commerce applications that can be customized to meet users' specific and ever-changing needs?* From the reusability models presented in [4], we explored the reuse capability of the hybridization of Service Oriented Computing (SOC) and Software Product Line (SPL).

# 2 Service-oriented Product Line

SPL and SOC enable the systematic reusability of software/service components and offer the runtime flexibility required for cost and time effective, high-quality application development. The fusion of SOC and SPLE can be regarded as service oriented product line (SOPL) [5]. Consequently, an SOPL can be defined as a family of SOA-based applications. Below we present a brief overview of SOC and SPL.

## 2.1 Service-oriented Computing

Service-Orientation is a paradigm for developing and deploying an application quickly and cost effectively. This paradigm utilizes services as building blocks to enable the flexible composition of applications, and it is realizable through the Service Oriented Architecture (SOA) [6]. SOA is an architectural model for building systems based on the interaction of services. SOA applications are developed on the paradigm of Component Based Development (CBD). The CBD approach enables the development of software applications by assembling existing components. These components can be acquired by leveraging legacy systems, as COTS from a third party vendor, or by developing components with reuse in mind. This facilitates shorter time to market, reduced cost, and increased reuse [7]. In SOA, software components are encapsulated as Services. A Service is a software component that enables access to one or more capabilities with prescribed interfaces [8]. Services can be composed to provide higher and more complex functionalities for distributed applications [9, 10, 11].

To drive business competitiveness, enterprises prefer applications tailored to their business needs. Developers are therefore required to produce variants of software to meet such customization. Therefore, building one-of-a-kind software applications for each variant-request involves an increase in cost and longer time to market. To enable effective reuse of software components, the development process must consider the variations of these similar software products. Realizing these types of variations is non-trivial and usually results in redundant component development. Reusability in this context must therefore be strategic, leveraging existing assets, and this is achievable with the Software Product line (SPL) paradigm [12].

## 2.2 Software Product Line (SPL)

**SPL** is a set of software-intensive systems that share a common, managed set of features that satisfy the specific needs of a particular market segment or mission and that are developed from a common set of core assets in a prescribed way [13]. SPL is used for developing core artifacts that enable systematic reuse, thereby reducing time to market. However, in many SPL approaches, the reusable software components are statically configured at design-time, which makes it difficult to make any changes to the application if the context changes [14].

The key to the SPL paradigm is the modeling and management of variability. Software variability refers to the ability of a software system to be efficiently customized or to be configured for use in different contexts. Variability is achieved through variation points [15]. Variation points are places in the design or implementation that signify some variance in functionality. A variation point is resolved by selecting one or more variances from the associated set of variances possible for that variation point. However, handling variability is non-trivial. In relation to SPL, variability describes the difference between the features of the systems that belong to the product line [16]. A feature is a property or functionality of a system that is relevant to some stakeholder. The commonalities or variability among products in the product line can be represented as features. SPL features are usually captured in domain analysis using feature models. Feature models are often used to model the variabilities in a SPL.

Feature modeling, first proposed by Kang *et al.* in [16] as part of the Feature-Oriented Domain analysis (FODA), is the activity of identifying commonalities and variabilities of the products of the product line in terms of features and organizing them into a feature model. With this model, commonalities are modeled as common features, while variabilities among the product line members are modeled as variable features, from which product specific features can be selected for a particular product instance.

A variable feature can be categorized as:

a) **Alternative feature**- indicates a set of features from which only one must be present in a product.

b) **OR feature**- represents a set of features from which at least one must be present in a product.

c) **Optional feature**- means features that may or may not be present in a product.

The commonality and variability information captured in the feature model serves as a basis for developing product line assets [17].

To achieve greater flexibility, an application should be composed at runtime, and this is achieved with Service-Orientation. The fusion of SOA and SPL, as earlier mentioned, is regarded as SOPL. Consequently, an SOPL can be defined as a family of SOA-based applications. SOA and SPL are relevant to the development of e-commerce because with SOA, e-commerce solutions can be composed of reusable services, and SPL can guide the development of the SOA-based e-commerce product family. The use of Service-Oriented techniques in e-commerce development requires additional activities and artifacts that are not realizable by traditional design and development approaches such as Object-Oriented Analysis and Development (OOAD), CBD, Enterprise Architecture Development (EA), or Business Process Modeling (BPM) [18]. The need for Service-Oriented modeling techniques was identified in [6] where the author stated that current OOAD and CBD approaches do not address the fundamental elements of SOA, which are services, their flows, and the components that realize them [18]. Service Oriented Analysis and Design (SOAD) techniques would require a combination of service design methodology with other software development techniques such as EA, OOAD and BPM [19].

However, an emerging SOAD approach as that proposed in [18] can only be used to model and develop single SOA systems. This approach is made up of three key steps, which include the identification, specification and realization of services, components and flows. To create a suite of applications with each build varying slightly from the other, the service design and modeling approach must be enhanced to develop a family of SO systems. This would facilitate the evolution of variants of the e-commerce application quickly. Our approach for developing a family of Service-Oriented e-commerce solutions is a hybrid of the SOA and SPL techniques. SPL would enable the exploitation of the commonalities of e-commerce applications through the systematic reuse of shared core assets, while managing variations for a particular customer.

# 3 Review of Related Work

IBM proposed a method for the development of Service-Oriented (SO) systems called Service-Oriented Modeling and Architecture, or SOMA [18]. SOMA is a 3-phase method for developing SO systems. These three phases are the Identification, Specialization, and Realization of services, flows and components, with each phase having various steps. The steps for Identification in the SOMA method include the analysis and decomposition of the domain of the emerging SO system, the analysis of the business goals, and the exploration of legacy systems to identify reusable parts. Our approach adapted the first two service identification approaches of the SOMA method. We augmented the service identification process with the feature-Oriented Analysis and Design (FODA) technique [16], in order to capture the commonalities and variability in the features.

In line with the concept of service flows mentioned in [18], we conceptualized an enterprise's business process as the flow of services in achieving business goals. In our approach, reusability is achieved by reusing orchestrated workflows of services. Reuse becomes achievable via specifying how the services would participate in the orchestration at runtime, by resolving the variation points in the business process.

Up to now, most work in the area of SPL has focused on software, but in [15], a product line architecture was proposed to design, deploy and maintain a family of web applications. The family members were composed from reusable component with an in-built variability-handling mechanism. The variability is resolved using an interactive wizard, and the system retrieves the available feature by reflection. However, in [20] it was established that a difference exists in the development of service-oriented systems compared to traditional web applications, and thus a slightly modified lightweight Product line approach with specific variability information was suggested. This approach influenced the work in [21], where a product line architecture for web service-based visual composition of web application was presented. They created domain-specific lightweight product line architecture and used a visual-based tool to configure web services and flows in the development of a specific web application. [22] proposed a feature model-based multiple view SOA variability model. The model was described in UML and SoaML and enables the systematic modeling of variability, independent of platform. In the approach proposed in our work, a visual-based tool was created that would be used by the application engineer to resolve the variation points in the variable business process. The feature modeling approach used was based on the FODA technique. The features to be included in the business process are to be specified, and through reflection, the variable business process is able to select the required variants.

In the same direction of work, Ye *et al.* [23] developed an approach for Service-Oriented Product Line Architecture (SOPLA) for business process families. The core asset of the SOPLA was a variable business process which captured the commonalities and variability in related business processes. Customization is achieved by configuring the varying elements of the business process for a particular target customer. However, it was not specified how to identify the service elements that make up the business process. Also, in [5] an approach was proposed for Service-Oriented Product Line. The approach has four activities, which are component identification, service identification, variability analysis and architectural specification. The identification activities use feature models and business process models as inputs but do not consider how to develop the services and components for the service oriented product line. The Approach in this paper used the concept of variable business process in [23]. We defined a SOPL architecture as a business processes containing all the items of the product line.

# 4 Our Approach

The motivation for this approach was adapted from the SOA reference architecture presented in [5, 18]. We considered a business-driven, top-down development approach. SOA defines an architectural style that provides a set of patterns and guidelines for creating loosely coupled, Business-ready services that can be recomposed in response to new business threats or opportunities [18]. Adopting SOA in developing business applications requires that the business functionalities be encapsulated as services. As stated earlier in this paper, an enterprise seeking to adopt e-commerce solutions may have a slightly different process for achieving its business objectives. SPL provides the paradigm to systematically reuse existing services, flows and components. SPL was also used to develop all the assets that will be shared by e-commerce products in the EoD product line.

To develop the approach presented in this paper, a survey of literature was undertaken, and the insights acquired were used to evolve our approach. A case study of EoD product line development was used to validate the plausibility of the approach. A visualization of the approach taken in this research is shown below.

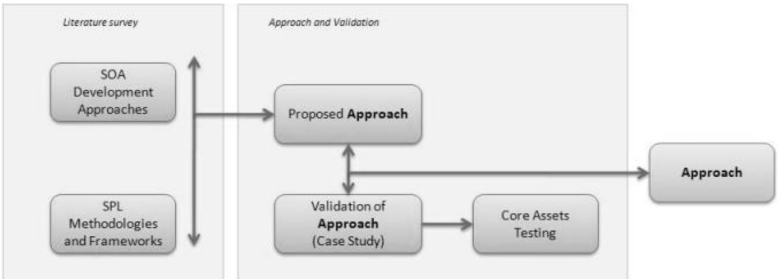



Figure 1
Visualization of Research Approach

Firstly, we embarked on a literature survey to analyze the state of the art in SPL and SOA practices and their relevance and application in the e-commerce domain, and we established that the paradigms can be integrated. Adopting the product line approach enabled us to develop the reusable core assets (product-line architecture, variant-rich business process, web services, etc.) required to realize customized EoD from the users' requirement. Also, a study of existing e-commerce solutions and a survey report of a group of SMMEs were carried out to identify the e-commerce requirements and features.

## 4.1 The Approach Process Architecture

The process architecture, presented in Figure 2, provides insights into the activities involved in our approach. It is an adaptation of the SPL and SOA development life cycles. The SPL process is divided into two: Domain Engineering and Application Engineering [24]. The Domain Engineering phase is where the core assets are developed, and during Application Engineering, products peculiar to the users' requirements are developed from the core assets realized during Domain Engineering.

### 4.1.1 Domain Engineering (DE)

In this phase, the construction of the core assets to be shared among products of the EoD product line is achieved. The business requirements and the features that the EoD will require to achieve business objectives are explored and this domain knowledge acquired is used to construct the domain artifacts (core assets). These core assets include generic services and flows (business processes) and a reference architecture that shows the interaction of these architectural elements and guides the construction of product line members. The Domain Engineering sub-process consists of the following stages: Domain Scoping, Domain Analysis, Domain Design and Domain Implementation. The activities in these stages are described in detail below.

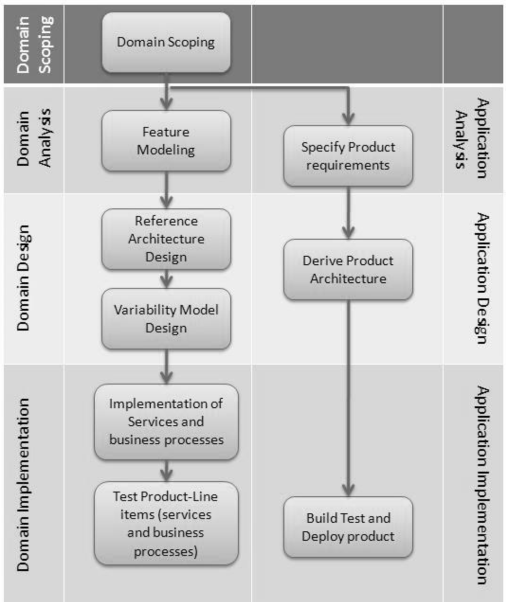



Figure 2
Process Architecture of Our Approach

***Domain Scoping***

During domain scoping, the boundary of the EoD Product Line is determined. The boundary defines the characteristics of members of the Product Line, and hence the derivable variants of the Product Line. These characteristics are compiled in a product map.

***Domain Analysis (Modeling)***

This activity includes the collection and organization of information on systems in the e-commerce domain that share similar capabilities and contain slight differences. The information is used to produce a domain model or feature model.

***Domain Design***

In this sub-process, the reference architecture and Business Processes of the EoD product line are defined. The reference architecture is an abstract structure for all the products in the product line. The variability model for the product line is also specified to resolve variations in the features of products in the product line.

***Domain Implementation***

During this sub-process, a detailed design and implementation of the reusable service elements and the Business Processes are achieved.

### 4.1.2 Application Engineering

In the application engineering phase, a new product instance is derived from the core assets based on the feature model defined during DE. This new product takes all similarities from the product line and the variability is managed using a visual tool, we created to resolve variation points in the product line. The application engineering phase also has the following sub-processes, application analysis, application design, and application implementation. During application analysis, the requirements specification for the product instance is developed from reusing the requirements from the domain engineering phase. In application design, the architecture of the product instance is derived from the reference architecture. During application implementation, the product instance is realized by configuring the reusable service elements and business processes that were created during DE.

# 5 Case Study Implementation

In order to validate the plausibility of our approach, a case study of the development of product line (WebStore) was undertaken. Nongoma is a rural community in Kwa-Zulu Natal, in South Africa. In Nongoma, there exists the Nongoma Tourism, Arts and Crafts center, which consists of local artisans who create and sell arts and crafts artifacts. A survey has been carried out, and below is a summarized report.

## 5.1 Preliminary Nongoma Questionnaire Analysis

Fifteen group members of Nongoma Arts and craft Cooperative were interviewed on the 7th of June, 2005 (35 Attempted Questionnaires). Below is the summary of results:

- 70% of the members are involved in arts and crafts activities
- 10% involved in Drawing, Music and Video production
- 10% involved in agriculture activities

- 5% involved in welding and building activities
- 5% involved in catering and decorations activities.

**Member's Overview**

- Groups can consist of group members ranging from 3 to 25
- 75% of members are females.
- 90% of the members do not speak nor understand English.
- Members educational lever ranges from primary to secondary school (standard 2- 10)
- Members are dispersed in various segments.
- 96% of members have neither been exposed to computers nor any IT infrastructure.
- Members have a positive outlook when confronted with the idea of introducing computers into the business.

**Envisioned Problems**

1. Marketing and advertising
2. Funding
    a. Equipment
    b. Capital
3. Managerial and entrepreneurial skills
4. IT skills

This report formed the basis of the requirements of an e-commerce on-demand solution, called the WebStore. The requirements for the WebStore e-commerce service are:

1. A web presence
2. Online booking for customers
3. An SMS Notification or email Notification for every Order completed by a customer
4. One Monthly Advert of special offers for Business Promotion
5. A basic Accounting and Reporting Service which entitles the SMME to one end-of-month standard report

## 5.2 Development Process

### 5.2.1 Domain Scoping

In domain scoping, the boundary of the e-commerce product line was determined. This boundary defines the characteristics of members of the product line, and hence the derivable product variants of the product line. These characteristics are compiled in a product roadmap. The product roadmap captures the domain scope which is used in domain modeling to produce a domain model.

The Application-Requirements Matrix approach presented in [25] was used to capture the commonality and variability among e-commerce products in the product line. The application-requirement matrix gives an overview of the commonality and variability for a given set of software product line application requirements. The product roadmap defines common and variable features at a higher level of abstraction [25]. The column headings of the Application-Requirements Matrix are the e-commerce products and the row headings are the requirements. The requirement of the SOPL from the report of the survey of SMME conducted in Nongoma, Kwa-Zulu Natal, is shown in Table 1.

Table 1
Application-Requirement matrix obtained from Domain Scoping

| SN | Requirements | Product A | Product B |
|---|---|---|---|
| 1 | E-commerce platform Management | Mandatory | Mandatory |
| 2 | e-commerce Store Front feature | Mandatory | Mandatory |
| 3 | Inventory Management | Optional | Optional |
| 4 | Customer Relationship Management | Optional | Optional |
| 5 | Order Management | Mandatory | Mandatory |

### 5.2.2 Domain Modelling

All the requirements captured in the product roadmap, with further analysis are used to produce a feature model. Feature Model shows a high-level unambiguous view of the common and varying characteristics of e-commerce products within the product line. Feature models decompose concepts into their mandatory (required) and optional features to produce a set of configurable requirements. These common and variable features captured in the feature model serve as a basis for developing product line assets [17]. The feature model of the Webstore product line is presented in Figure 3.

### 5.2.3 Domain Design

The feature model created during domain modeling activity is the input of the domain design activity. In this activity, the domain model created was used to produce the Product Line Architecture (PLA).

### *The Product Line Architecture*

The PLA is a structure that incorporates all the features of all the products in the product line, showing the locations where variations occur and how the variation points will be resolved. The Architecture of the product line is presented in Fig. 4.

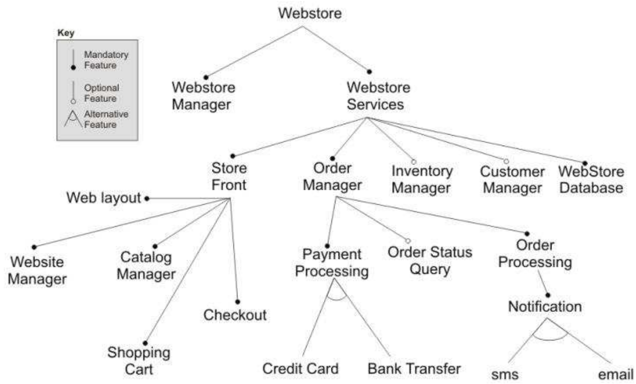



Figure 3

The FODA feature Model of the Webstore Product Line

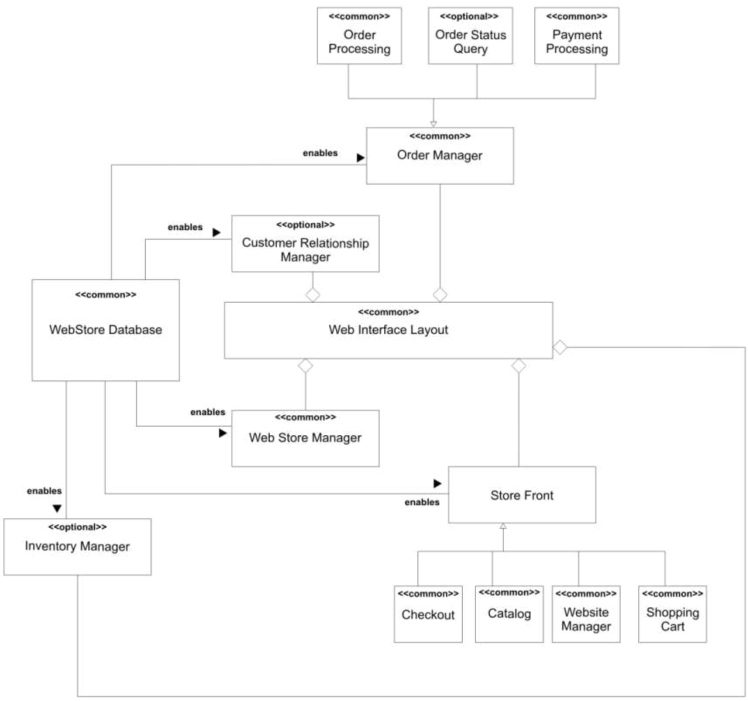



Figure 4

The Architecture Diagram for the Webstore Product Line

### List of the Web Store Business Process

The feature model and the PLA were used to identify the list of business sub-processes (components) that comprise the e-commerce products of the Webstore product line. The identified processes include: purchasing, order manager, inventory manager, and customer relationship manager. The identified processes were further analyzed to identify the various activities required to fulfill each process. Below is list of the processes and the activities/sub-processes they entail:

1) Purchasing
   a) Browse catalog
   b) Add item to shopping
   c) checkout
2) Order Manager
   a) Create new order
   b) Process payment
   c) Notify supplier
   d) Notify customer
   e) Pack and Ship product
3) Inventory Manager
4) Customer Relationship Manager

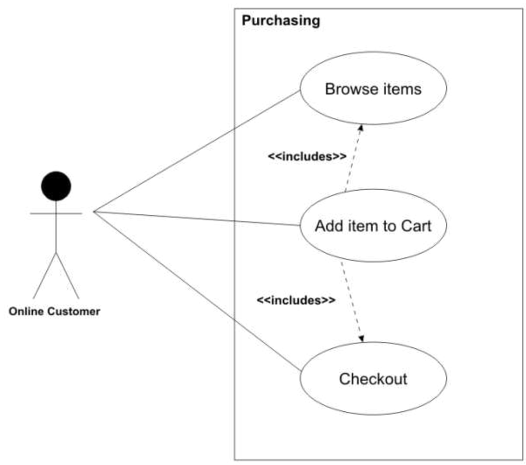



Figure 5
Use case Diagram for Purchasing Process

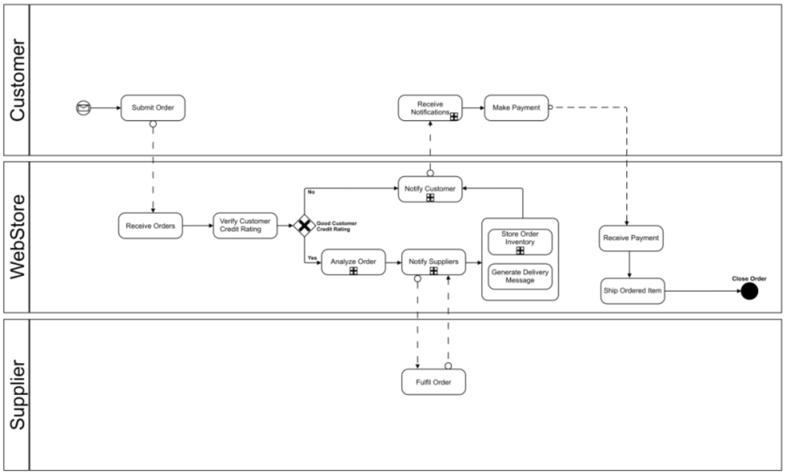



Figure 6

High-level BPMN Model of the Order Manager Process

### 5.2.4 Implementation Details

The implementations of web services to realize the business process were based on Java Platform Enterprise Edition (JEE) using the Netbeans 6.7.1 Java Integrated Development Environment. The web services were implemented as Enterprise Java Bean wrapped with Web Services Description Language. The EJB accessed the MySQL database for data required to deliver relevant functionality using the MySQL J-Connector. The business processes were implemented using the visual tool provided in Netbeans 6.7.1, where web services are orchestrated to provide a composite functionality using Business Process Execution Language (BPEL).

A composite application was created using the CASA editor in Netbeans 6.7.1. The composite application was deployed on the BPEL engine of the glassfish application server. The composite application was tested using the in-built composite application test case functionality in Netbeans 6.7.1. The storefront component, website management component and the WebStore Manager Component were implemented using Hypertext Preprocessor and Hypertext Markup Langauage and hosted on apache web server. The Storefront component communicated with the business process component using Java EE Servlet. The web layout template was implemented as a cascading style sheet file using Macromedia Flash and Macromedia Dream Weaver tools. The Webstore database was implemented in MySQL, which exploits the JDBC technology to connect to the EJBs.

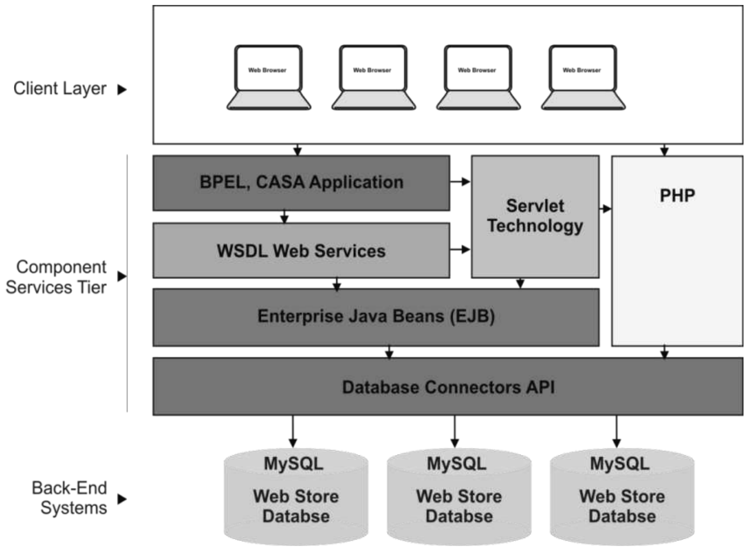



Figure 7
Deployment Technology Stack

Figure 7 shows a view of the run-time deployment architecture of the domain components. It is a 3-tier architecture showing the configuration of the content components as deployed on the Glassfish application server and Apache server in the middle layer. The data layer consisting MySQL database makes up the third layer while request for services are made through the client layer.

### 5.2.5 Application Engineering

In the Application engineering phase, the creation of specific products in the product line was achieved through the reuse of artifacts created during DE and exploiting the product line variability. In application engineering the parameterizable artifacts were configured with concrete parameters and composed to deliver the required functionalities. In the particular instance of our case study two e-commerce products variants were considered. These are the product-A and product-B.

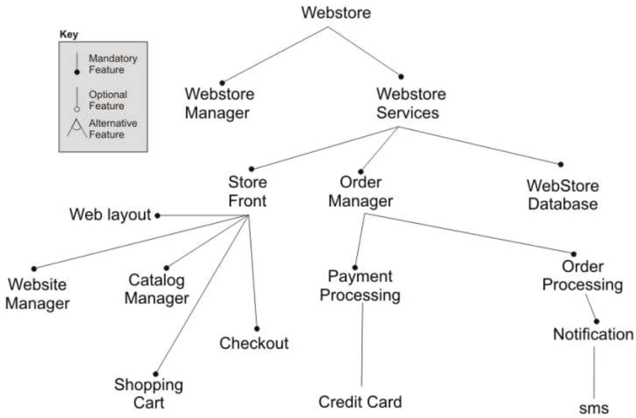



Figure 8
Feature Model of Product Instance-A

## Conclusion and Future Work

To participate in the on-going global economy, access to software applications to e-empower SMME operations in developing economies is vital. Providing these technologies will enable local ES development organizations to contribute directly and indirectly to economic development [26], specifically to employment generation, wealth creation and an increase in national GDP. This contribution can be sustained and possibly increased by productive input factors that can drive the quick development and delivery of quality software products. This work demonstrated the application of SOPL to develop solutions for on-demand computing contexts, using the GUISET platform as a case study. This is important to ES developers, who have the burden of evolving solutions in line with changing user requirements. However, in this work, some required services such as credit card processing and credit card verification were known and bound to the application in advance. The use of a service broker to search and bind to required services based on QoS attributes can be explored. Quality of Service (QoS) is a major characteristic of on-demand services, but in this work, QoS-awareness was not considered in the design of the product line components. Therefore, more work would have to be carried out to explore the use of components to offer services with differentiated QoS. Also, the points of variability in the developed product line are minimal and predictable compared to a real life case study product line. It is recommended that a more elaborate case study development would have to be embarked upon to demonstrate SOPL in on-demand contexts. It is desirable that external services be considered in the modelling of product line reference architecture [27], but we assumed in our work that all the product line assets were developed under the control of the developer organization. Further work would have to be done to incorporate third party and external services as part of the variable product line reference model, and their functionalities invoked through the use of APIs. Finally, in the future, specific tools would have to be developed to model and manage the variability in SOPL, and automate products derivations easily and rapidly.

**Acknowledgments**

A research project, Grid-based Utility Infrastructure for SMME Enabling Technologies (GUISET), is being undertaken at the Centre for Mobile e-Services for Development. The goal of the project is to transform SMME into a global business by providing affordable access to technology on a pay-as-you-go basis. This work was done while the first author was on temporary research fellowship at the Centre at the University of Zululand. The centre is funded by THRIP, Telkom, NRF, Huawei, and Alcatel.

**References**

[1] Terzi, N.: The Impact of E-Commerce on International Trade and Employment. *Procedia Social and Behavioral Sciences* (2011) pp. 745-753

[2] Consoli, D.: Literature Analysis on Determinant Factors and the Impact of ICT in SMEs. *Procedia - Social and Behavioral Sciences* (2012) 93-97

[3] Krueger, C. W.: Software Reuse. *ACM Computing Surveys* (1992) pp. 131-183

[4] Sommerville, I.: *Software Engineering, 7th Edition.* Addison Wesley, (2004)

[5] Medeiros, F. M., Almeida, E. S., Meira, S. R.: Towards an Approach for Service-oriented Product Line Architectures. Proceedings of the *13th International Software Product Line Conference (SPLC).* San Francisco (2009) pp. 1-7

[6] Papazoglou, M. P., Traverso, P., Dustdar, S., Leymann, F.: Service-oriented Computing: State of the Art and Research Challenges. *Computer, 40* (11) (2007) pp. 38-45

[7] Lau, K.-K., Wang, Z.: Software Component Models. *IEEE Transactions on Software Engineering, 33* (10) (2007) pp. 709-724

[8] http://www.oasis-open.org, Accessed: 28.02.2012

[9] Karunamurthy, R., Khendek, F., Glitho, H. R.: A Novel Architecture for Web Service Composition. *Journal of Network and Computer Applications* (2012) pp. 787-802

[10] Bosin, A., Dessì, N., Pes, B.: Extending the SOA Paradigm to E-Science Environments. *Future Generation Computer Systems* (2011) pp. 20-31

[11] Li, X., Fan, Y., Madnick, S., Sheng, Z. Q.: A Pattern-based Approach to Protocol Mediation for Web Services Composition. *Information and Software Technology* (2010) pp. 304-323

[12] Andersson, H., Herzog, E., Ölvander, J.: (In Press) Experience from Model and Software Reuse in Aircraft Simulator Product Line Engineering. *Information and Software Technology* (2012) http://dx.doi.org/10.1016/j.infsof.2012.06.014

[13] Bass, L., Kazman, R.: *Software Architecture in Practice.* M. A: Addison-Wesley (2003)

[14] Lee, J., Kotonya, G.: Combining Service Orientation with Product Line Engineering. *IEEE Software, 27* (3) (2010) pp. 35-41

[15] Balzerani, L., Di Ruscio, D., Pierantonio, A., De Angelis, G.: A Product line Architecture for Web Applications. *Proceedings of the 2005 ACM symposium on Applied computing*. Santa Fe, New Mexico: ACM New York (2005) pp. 1689-1693

[16] Kang, K. C., Cohen, S. G., Hess, J. A., Novak, W., Peterson, A.: *Feature-Oriented Domain Analysis (FODA) Feasibility Study.* Software Engineering Institute: Carnegie Mellon University (1990)

[17] Lee, J., Muthig, D., Naab, M.: A Feature-oriented Approach for Developing Reusable Product Line Assets of Service-based Systems. *The Journal of Sytems and Software, 83* (7) (2010) pp. 1123-1136

[18] Arsanjani, A., Allam, A.: Service-Oriented Modeling and Architecture for Realization of an SOA. Proceedings of *IEEE International Conference on Services Computing*. IEEE Computer Society Washington, DC, USA (2006) p. 521

[19] Zimmermann, O., Krogdahl, P., Gee, C.: *Elements of Service-Oriented Analysis and Design.* from IBM developerWorks. Available at: http://www.ibm.com/developerworks/library/ws-soad1/ (2004) Accessed:21.07.2010

[20] Capilla, R., Yasemin Topaloglu, N.: Product Lines for Supporting the Composition and Evolution of Service-oriented Applications. *Proceedings of the 1st International Workshop on Principles of Software Evolution*. IEEE Computer Society Washington (2005) pp. 53-56

[21] Karam, M., Dascalu, S., Safa, H., Santina, R., Koteich, Z.: A Product Line Architecture for Web Service-based Visual Composition of Web Applications. *Journal of Systems and Software*, 81(6) (2008) pp. 855-867

[22] Abu-Matar, M., Gomaa, H.: Variability Modeling for Service-oriented Product Line Architectures. Proceedings of the *15th Software Product Line Conference (SPLC)*. Munich (2011) pp. 110-119

[23] Ye, E., Moon, M., Kim, Y., Yeom, K.: An Approach to Designing Service-Oriented Product line Architecture for Business Process Families. *The 9th International conference on Advanced Communication Technology*. Gangwon-Do (2007) pp. 999-1002

[24] O'Leary, P., Almeida, E. S., Richardson, I.: The Pro-PD Process Model for Product Derivation within Software Product Lines. *Information and Software Technology* (2012) pp. 1014-1028

[25] Pohl, K., Böckle, G., Linden, F. V.: *Software Product Line Engineering: Foundations, Principles, and Techniques.* Berlin Heidelberg: Springer-Verlag (2005)

[26] Jaakkola, H.: Towards a Globalized Software Industry. *Acta Polytechnica Hungarica*, 6(5) (2009) pp. 69-84

[27] Khoshnevis, S.: An Approach to Variability Management in Service-oriented Product Lines. *Proceedings of 34th International Conference on Software Engineering (ICSE)* (2012) pp. 1483-1486

# Interactive Value Production through Living Labs

**Imre Hronszky, Katalin Kovács**

Budapest University of Technology and Economics
Műegyetem rkp. 3-9, H-1111 Budapest, Hungary, hronszky@eik.bme.hu

ValDeal Innovations Zrt., Gyár u. 2, H-2040 Budaörs, Hungary, kovacs@chic.hu

*Abstract: Interactive, co-creative relations and collaboration of consumers, users and producers are quickly developing recently. Living Labs [LL] integrate users in the development process of new technologies as co-creators themselves. They have a special bridging role between market pull and technology push innovation and they realise some sort of concurrent innovation. LLs are an interactive search for new products/services in real life milieus together with users/consumers, without the mediation of marketing experts. They receive a different role. Our article highlights LLs first as providers of a collaborative working environment for users. This paper emphasises that LLs have a strong methodology and describes and assesses the "LL Harmonization cube". Further, it puts emphasis on the difference between a prospective notion of a LL and the recent reality of the LL practice. Recent LLs mostly work at the end phase of the innovation 'chain'. The article secondly outlines what advantages LLs can bring for SMEs. The main added value of LLs for SMEs is that they provide for innovation services by integrating SMEs as users in a collaborative working environment that would otherwise not be available for them.*

*Keywords: Open innovation; Living Lab; Harmonization cube; SME involvement*

# 1 On Open Innovation, the Supply Side

With some approach to history, in the majority of the 20th Century the dynamic of innovation was 'linear', in-house, 'closed' innovation. Today, innovation is a networked phenomenon where the main issues happen in a globally connected world at the most varied intersections, for example at interactions among disciplines, interactions with the suppliers, interactions with consumers, etc. The innovation dynamic is thus full of feedback and often unavoidable repeated restarting of circles before a dynamic of iterative cycles is reached. Innovation is getting more and more open.

Linear innovation integrates the need and the demand sides very specifically. Linear innovation realizes a technological push, developed in the R&D

department of the firm, answering some broad demand only. You can successfully realize it when different factors are in coincidence, among them a stable broad demand that can be satisfied by mass products, this is still partly true with drug production, further when the decisive expertise can be gathered and isolated in the lab of the firm, when consumers are passive in the process, etc. Proprietary relations, command and control organizational mode within firms, mass production, and stability of supply on the market were integrated into a working dynamic in the earlier period and were tought to be the only possible model of innovation dynamics. But nowadays, 'linear', 'closed' innovation could not be successfully, efficiently realized by even the biggest firms. This is so because it would be either inefficient, or losing time or impossible to financing. The Ford T phenomenon may be the icon of closed production system with some product that could successfully count on some huge, stable, undemanding consumer layer. Henry Ford did not aim at the innovation of his product. Two generations later, until the early 90s, IBM can be seen as icon of closed innovation. That system integrated innovation vertically into the production and distribution value chain.

Relying on experts inside the huge firm, the R&D department "within the walls" was the model for managing innovation in the second half of the last century. That was complemented with some stable co-operation forms, such as co-operation with some university labs. First customization efforts were supported by this structure. For a number of reasons that model definitely ceased to work appropriately at the end of the millennium. Among the numerous factors leading to its inappropriateness, first of all we mention the changes on the demand side, around the emerging need for individualization of the commodities for not only thin layers of customers. When the individualized needs of customers became to be taken into account, as far as possible, the relation between the knowledge and orientation in the R&D departments and the concentration on users imagined by them, became more and more inappropriate. It is possible to say that a special sort of democratization started with this. But this inappropriateness is still partly preserved in practice, notwithstanding the huge development of marketing. [We have no place here to emphasize how immense new roles are being acquired by internet-based marketing and marketing research and integrated with long tail distribution based vendor activity, and how it revolutionizes the business. All this is most important. But we try in this short article to put the emphasis on the full-fledged interactive relation between the producer and consumers in the LLs.]

This customization tendency was coincided with the massification of educated layers, who could in principle provide that expertise needed by a firm that was unable to find in its R&D department. In addition, the potential of knowledge spin-offs multiplied quickly. Some possibility for a very strong supply of experts from outside and expertise from users emerged by the end of the millennium. Expertise was rather rare earlier and transactions costs to utilize expertise from the outside seemed or proved to be too costly. While integrating the R&D department into the vertical value chain had been justifiable both by transaction costs and human relations considerations from the early 20s in the 20th Century, and while it had been more efficient to integrate R&D internally rather than looking for expertise outside, under these conditions, ironically, at the end of the 20th Century it started to become more efficient to follow the reverse path and to give up this vertical integration for a much more 'horizontal', open one. Changing vertical integration of [a part of] R&D for radically changing the search for expertise on the globalizing market and involving customers as a kind of producer as well, based on their special user expertise, started to become a good alternative. The internet, as an enabling medium, has an especially important role in this transformation.

This new possibility is termed 'open innovation'. In the globalizing environment every agent moving on the innovation scene has to try, in its own interest, to rely on, explore and exploit the disproportionally bigger knowledge base outside any firm. Henry Chesbrough termed a part of this new phenomenon 'open innovation' in 2003 and elegantly provided important knowledge of the changes, in which, in issues of R&D&I, the walls of the firms are getting more and more interpenetrable. The flow is really two directional. While we have numerous examples of when firms earlier tried to obtain and at least to keep on the shelf a lot of patents without any intention to realize them, but in this way hindering rivals from getting access to them at least, and counted stopped alternatives in their innovation funnel as unavoidable costs of innovation, many firms, by now, not only regularly outsource their R&D&I tasks but also make marketable their unrealized innovative ideas and development alternatives, looking in that way for some additional revenue. With regularly in-sourcing and outsourcing innovation tasks putting them in the globalizing marketplace, a new type of managing innovation emerges. Competition is supplemented by co-operation in this case on a market base, and becomes different from pure rivalry. Market-based worldwide co-operation has become essential in the competition for innovation. This is further complicated by open source activities financed by big firms in which research results are made open, even for competitors or together realized by competitors, in the hope of using the results for different targets. In addition, for example, some sorts in the open source movement and wikinomics [commons based peer production] are non-market based forms of co-operation. Actually, all these things are 'open' innovations.

Chesbrough introduced the fortunate term 'open innovation' in his path-breaking book in 2003 [9] but concentrated on the market based interactions. He concentrated on the specificities of the globalizing R&D&I market in which especially outsourcing the tasks, buying and selling ideas, prototypes or semi-final products started to become decisive constructed advantage in the global competition race. His main message was that, first of all, an appropriate business model is essential to realizing this possible advantage.

# 2 On Open Innovation, the Demand Side

Chesbrough originally concentrated on product and production development and conceptualized the open market as B2B interaction, the division of labour among firms. The reasons for both limitations made by him at that time are already over, for analysis as well. Chesbrough himself has by now turned his attention from product innovation to service innovation [10]. He identifies service innovation as the escape route from the global 'commodity trap', the complex effect of the tendency that results from the interaction of diversification of products, of the moving of the production to countries where the wage is low and shortening of time products have on the market. Open service innovation is offered as a solution.

By responding to the constraints and opportunities of the changing innovation dynamic, many firms become networked in their innovation dynamic too. Gigantic firms such as the leading firms in aerospace industry move toward the realization of the vision that their strength is the ability to develop an ecology of a very far reaching division of labour with very different suppliers, and their basic core competence is the ability to manage the collaboration. What about the consumers?

Mass customization was a pioneering step 30-40 years ago in most developed countries. The current challenge is formulated, based on some practical cases already, by the late Prahalad and his followers as 'personalizing consumer experience'. [3] In a world of the personalized consumer experience, the role of the firms is profoundly changing. In the developing ecology, producers and consumers mutually provide an enabling environment to each other.

Concerning the involvement of consumers/users, one has to consider that product development could be, up to a limit, imagined for 'virtual users', as it was imagined simply within the firm walls for a long time in the history of industrial society, and only gradually was market research included as correction tool when the product, less the service needs became more individualized. But service innovation may work much less than product innovation without interaction with real users, without taking into account their individualizing needs. The notorious need for 'a hole in the wall' [Peter Drucker] can, of course, not only be satisfied by choosing from a much richer offering then by becoming proprietor of some drilling machine but the consumer may know much more about her contextual need situation than any producer. In a lot of cases, only s/he may deepest know the target and some essential conditions to reach it at her local place. One way to realize an interactive relationship is through utilizing the immense new possibilities of interactive marketing; another is realizing an LL relation as coproduction, interactive value production in the production process with the producers in a real-life situation, in its fullest developed form dominated by the user.

It is a commonplace in management literature in the last twenty years that ’the customer is the king’. One way to learn about the customer is to get as much information as possible, or rather, to get the appropriate information about their behaviour through observation and experimentation with them. For our intentions in this article the important issue is that the customer is made an object of, in its developed form, an interactive, participative observation so that s/he herself gives the needed answers. We all know the great development over the last ten years, a strong revolution in marketing and marketing research both in terms of the social sciences and economics disciplines utilized and the changing technological base, the systematic exploitation of the Internet.

There are some limits here nevertheless, impossible to transcend. On the one side, experimenting with agents unavoidably preserves some reification even in the most open experimentation. So, not only at least some elements of the behaviour of the user unavoidably remain covered for the experimenting agents, but even the behaviour of the consumer as a whole may be distorted. On the other side, the mediation of the received information has its shortages, too. First of all, gathering information needs mediation by the marketing researchers. They have to translate for themselves first, and then to the producers, what they received as input from the customers. The difficulty can be indicated by referring to the ‘stickiness’ factor, including the tacit dimension, of the main bulk of knowledge users have. And of course, any observation situation preserves some sort of artificiality. Putting customers in an artificial milieu may have its unavoidable limiting effects on the possible information gathering. The question may be raised what could be perhaps a better basis for learning about the customer that can overcome, substitute or add to the information gathering process developed by market researchers. The ‘artificiality’ of situations when innovators initialize involvement of possible users to develop a new product in a co-evolutionary way, as is rather typical in software development, is quite a different issue from such experimentation when they interact with the developers-innovators only through mediation. We are interested in this short article in LLs. In them there is some trial to transcend this limit. Speculation on what the inherent limits of realizing LL relation between the producers and users are cannot be the object of this article, unfortunately.

The question of providing the most enabling place for customers to realize themselves is of the highest importance, for the consumer is really starting to become the decisive factor in the innovation race, and innovation capability is envisioned as the decisive strategic factor in the global economic competition. Trusted by the European Commission to assess the European tasks in the global competition race, in 2006, the Aho report [7] called for the EU to start to concentrate the entire societal-economic dynamic around innovation. In this, the demanding consumer is one of the essential elements in the production-consumption circle. There is a growing general bad feeling in most developed countries over the last decade that the limitless availability of basic scientific knowledge and the nearly irresistible process of outsourcing of labour based on law wage into countries where labour is cheap, may finally result in losing the leading position of the recently most developed countries in global economic growth. One could add that the changing relation between the recently still leading countries and the BRICS, especially China and India, in the innovation race too, is about to become another very important factor in innovation, one worth being concerned about in the traditionally leading countries.

The typical suggestion for a solution to the challenge is traditional. It is to improve on the R&D supply capabilities in product and production. There are other suggestions. Chesbrough, in 2011 [10], shifted attention to improving service innovation. Amar Bhidé [1] tries to moderate the starting panic and suggests as a 'consolation' turning more attention to weak places in the innovation chain. He emphasizes 'mid-level innovation', and taking advantage of innovating for the 'venturesome consumer'. He agrees that the race in outsourcing cheap labour or that basic science is within reach for everybody are not issues to reverse, but rather he assesses that this is less important for the global innovation race than is generally recognized in the literature. In his assessment, the main issue concerning global economic growth remains the competition in the markets of the most developed countries. A decisive turn to the 'venturesome consumer' is a decisive driver to preserve the position, he suggests. 'Venturesome consumers' may be ready to develop and identify new needs which are costly and require intensive R&D&I on the supply side. All this means critical important feedback as a driver in the 'innovation chain'; that is, concentrating on the further integration of the steadily emerging needs of the 'venturesome customers' inside the 'innovation chain', as soon as possible as a co-evolutionary dynamic of producers and consumers. With this we have a suggestion for solution to find one adequate microeconomic and management answer to the macroeconomic challenge of global economic growth for both the most developed and the developing countries.

Bhidé concentrates on 'venturesome customers' and narrows the focus on customers to rich customers. There is still some widespread misunderstanding that the realization of the growing market for poor consumers does not need their interactive involvement in the production process as users. We have no place here to outline how essential can be the interactive user involvement in this dynamic, as the successful practices, including LL trials for example in South-Africa or India, demonstrate, but just notice that "reverse innovation" is rapidly growing as a main innovation type. And Living Labs are quickly developing in interactive value production with poor consumers too.

Eric von Hippel believes that most of the innovative product and production technology ideas in history have been introduced not by producers and professional developers but by users. [6] Already a look at history makes it rational to think that concerning some unavoidable domestication of the mass products, i.e. accommodation of them to local conditions, users have been constrained to make their steady improvement efforts. Notwithstanding the uniformity overtone in mass-production, innovation in the diffusion phase has been unavoidably important. But Hippel also turns attention to a special group of users, the 'lead-users'. Interaction with them turns on its head the traditional relationship of producer and consumer, and the firm becomes the producer of the products envisioned by 'lead-users'. The history of innovation in the period of mass production was a period when the innovation dynamic leading to mass products was dominated by professional inventors whose activities were mainly subjected to and regulated by the joint effect of the supply side; and mass needs and the 'closed' form of innovation can be seen as the trial by the firm to get under control the innovation capability for producing mass products.

The recent overarching technological revolution strengthens the capability of 'professional inventors' in the labs of the firms but threatens as well with the consequence that it will provide a further growing stream of innovations for 'virtual users', as long as the old producer-consumer relation remains preserved. There is a very quickly growing trend to explore and exploit application possibilities in the recent technological revolution to utilize its potential. But costs and risks of R&D&I are also very quickly growing while the time component along the value 'chain' is getting shorter and shorter. Breakthrough innovation possibilities multiply in accelerating waves and with this innovations are turning to be a high risk [also high benefit] enterprise. With the acceleration of the race, there is much less time and space than in the period of mass production to thoroughly exploit new basic results of the exploration phase, because after some very partial exploitation, the recent dynamic forces companies to turn towards the exploitation of some newer basic results.

With this the decisive challenge, the bottleneck problem may become to 'meet' the 'customer'. This need may partly include that s/he may not want to change her consumption in the directions the supply side suggests. This can partly occur because those offerings may not meet existing or emerging [real] needs of [possible] users. That means that among the quickly growing abundance of new commodities [services included] there is a relatively narrow, perhaps even relatively narrowing range of those commodities which are really needed in the form they are offered. For example, the great majority of software is still not user friendly, as the EC estimates.

A double process emerges quite quickly. First, the importance of professional researchers and inventors inside the firm is balanced by those incomparably larger number of professional researchers and inventors who are outside but ready to cooperate. [This is recognized by Chesbrough in his first book, in 2003.] This process strengthens the supply side but may still threaten via the overproduction of not really needed products that consumers may later reject. The performances of open innovation realized this way are challenged by the need for essential inclusion of [possible] users and consumers in a new, integrated mechanism of research and innovation. Herewith, we have the challenge to realize the integration of the demand orientation of R&D&I in the innovation chain to balance the innovation dynamic, provided we understand both existing and emerging 'demands' as 'demand'.

The following figure, made after Donald Stokes 1997 [4], represents the place of 'use-inspired research' as a result of some integrating 'concurrent' activity. [A 'concurrent' activity solves the problem simultaneously from both sides. Stokes' perspective was to improve 'use-inspired research' by turning researchers consciously in this direction, but he himself did not speak about inclusion of open innovation, either in the form of innovation outside the walls of the firm or inclusion of interactive customers.]

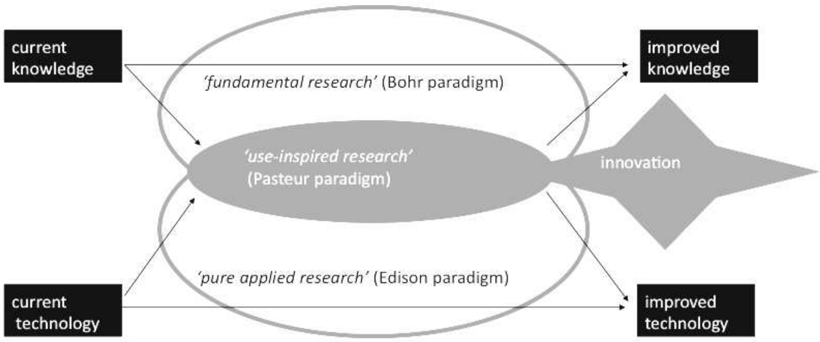



Figure 1
Use-inspired research [12]

A research dynamic involving 'use-inspired research' is the furthest element of inclusion of the 'demand side'. It is the inclusion as active participants of the possible users already in the first element of the innovation 'chain'. This is the realization of a co-evolutionary approach. In the course of the realization of this dynamic, a selection process of possible users is quite natural. It is important to see that, with emerging innovations attributing the usual role of the demand side, representing 'the visible demand of buyers' [12], will be misleading. This is extending the structure of incremental innovations to every innovation. We can place emerging demands on the demand side, still, or with high rationality, on the supply side and identify them as a special sort of push. It is rational to say that in the context of early co-creation with possible users a 'contextual push' based on needs and dreams of possible future users evolves [12]. This is different from the market pull, especially when the market is still very unripe. This 'contextual push' provides a substitute for the still missing demand side and turns to be the market demand with ripe products. The challenge we concentrate on is to realize the 'contextual push' as fully as is meaningful as early as possible.

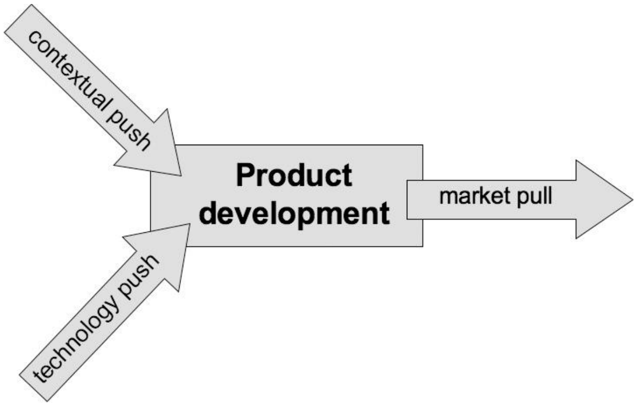



Figure 2
Product development and the relation of technology and 'contextual push' and market pull innovation [12]

There is a growing belief that the chance of successful innovation moves from setting the task of innovating 'for the consumer' through moving to innovating 'with them' to the level of innovating 'by them'. The role of the firms is then to provide for an evolutionary environment. In this environment possible users, either other firms or human users, are enabled to develop their visions on their needs and can also find in firms the instruments to materialise the solutions that satisfy them. In this respect, we have the common denominator with lead user driven innovation, with open source based innovation, or with the famous pioneering, the LEGO Mindstorm issue. But it is important to see that there are rather different types of users, just as rather obstinate types of behaviour. They can be differentiated according to their possible roles in the innovation dynamics, more as pioneers or laggards, or their need for co-creation, etc.

# 3 What is a Living Lab [LL]?

One possibility of the answer is that an LL is an immediate, interactive, co-creative search, by active participation of users/consumers in the innovation process for new products/services in real life milieus. An LL form of co-operation means overcoming the mediation by market researchers by the immediate co-operation with the producers and the overcoming of the testing attitude by realizing a more symmetrical, co-creative relation. As with any other type of basic innovation of innovations, a working cycle of a production process in an LL ends in a mutually enabling change leading to some new 'product' [product or service] and some new societal infrastructure. These are embedded in the reproduction dynamic of a living network.

It is worthwhile to turn to some comparison. In an often cited article, Pieter Ballon et al. [16] speak of a set of testing and experimentation platforms (TEP). They cover with their description the multiplicity of different test and design facilities. Among the different elements of TEP are real life user contexts, imagined by experts, with the purpose of observing the behaviour of those who are put in this environment. The observation aims at getting knowledge of the social and economic changes related to ICT developments. This knowledge is first of all expected to be used for making decisions among emerging technological alternatives by diminishing the uncertainty, in terms of acceptance by the consumers, surrounding their application in the real life milieu. “In addition to technological and innovation support, these facilities are set up for understanding and guiding the social and economic changes related to digital technologies and ICT developments. This refers to experimental settings, often imitating real life user contexts, where ICT developers and users interact and exchange views for optimal technological introduction.” [16]

In the interpretation of Ballon et al. TEP provides the needed correction for some so called system failures. „In order to get a better grip on the innovatory use and how this can contribute to the technological landscape, these users are more and more investigated in direct contact with the technological prototype or service [that is being developed]. The ‘virtual user’ (Flichy, 1995) is replaced by the ‘real’ user in the innovation process.”[16] This is most important but we have here the user still as an object for learning by experts when s/he interacts with developers in an as far as possible real milieu. This type of interaction is realized in a research milieu to learn from the interactive behaviour of the users by their observation. But it may give rise to immediate co-creation and aim at improving the innovation dynamic and turn into an LL relation among the participants.

LLs may lead much more in the direction of co-creation with the users than simply realising interactive test processes. An LL may offer some sort of service for possible users so that they can solve their existing problem by integrating themselves in this service milieu and use its service for their own purposes by transforming it. In this way, first they can produce an individual ‘product’ [in the overarching meaning of ‘product’, i.e. product or service], by way of application of their service but in an interactive process perhaps even led by the customer. Second the process of that ‘customization’ creates the possibility for a special type of generalization too, for the service provider. LL experiences provide them, to express it paradoxically, with non-standardisable “prototypes” of which working in real life milieus can be most rationally expected. LLs may provide here not just a qualified collaborative working environment to realise some individualized ‘product’, but also a learning milieu for interactive value creation of users with the service providers to serve for generalization. Better to say, two subclasses may perhaps be differentiated. In one subclass LL type collaboration in the dynamic of ‘concurrent’ innovation provides for over-bridging the pre-commercial gap, or improving any other previous elements of the ‘innovation chain’ for the producer.

In the other subclass, realising a co-creation process first ends up in an individual product such as a changed city milieu like that realised in the Arabianrata project in Finland. [Arabianrata is a district of Helsinki that was reconstructed based on an LL project] Cases in this subclass can then serve as analogies for generalization but can not lead to standardizable prototypes. That is made by 'inductive' reasoning, unavoidably penetrated by expert opinion, provided by co-creation in the whole LL process including the users and further in the network of LLs everywhere where the first LL experience is intended to be extended.

It is important to make a projective concept of the LL, as a fully developed activity, serving as a vision for orientation; and it is also important to recognize the limits of the recent achievements of the existing LLs in reality. LLs are complex systems with a range of different levels in terms of user co-creativity. It is natural that numerous configurations recently classified as LLs are still at a quite low level. Concerning the phases of the 'innovation chain', LLs are recently most important in over-bridging the pre-commercial gap, and these LLs may prove to be most important in the future, too. But some others deserve the LL name for their activity in the last phase of the innovation dynamic only.

LLs as working units spread rather quickly. There were over 200 by 2011, mostly in Europe, and it can be expected that their number may grow to over 300 following the announcement of the fourth wave of ENoLL, the European Network of Living Labs membership enlargement in Budapest, in May 2011. Membership in ENoLL has a very important function when providing possibilities to utilize synergistic effects. Developing a flexible but well-developed set of methodologies is sine qua non for the lasting success of LLs, just as is providing a stable quality of the services that LLs can offer. Methodologies and quality assurance measures work for keeping and raising the quality of any LL, as well as the interoperability of different LLs.

# 4 The Harmonization Cube Methodology and Quality Assurance of LLs

European LLs have been unified by ENoLL into a network with the aim of strengthening collaboration and utilizing complementarities and resources among the members. There is a common methodology already to harmonize and exchange best practices of LLs developed by Mulder in 2007 [15] as the so called harmonization cube. In the Helsinki manifesto, which was communicated during the launch event of the first wave of LLs [November 2006], it is stated that, "This approach should ensure that common methodologies and tools are developed across Europe that support, stimulate and accelerate the innovation process It can be said that harmonization of LL methods and tools is key." [11]

The harmonization cube defines the main interoperability elements from organizational, technical and contextual points of view, by the stages of LL maturity. It details the main elements of the evaluation methodology divided by the maturity stages of the LLs - and the direction for further development- the stages of setup, sustainability and scalability put on the vertical axe, in this order.

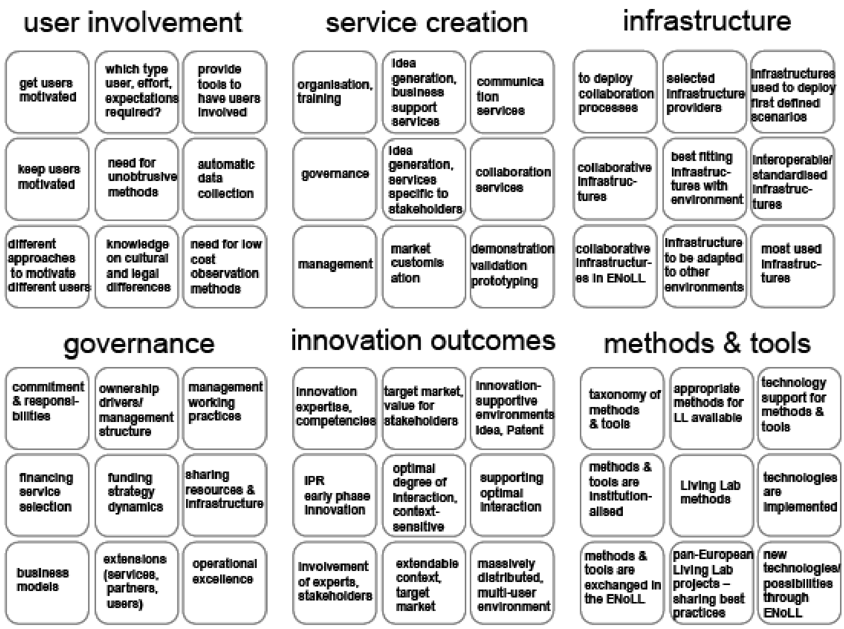



Figure 3

The elements of the harmonization cube [11]

The harmonization cube details the main elements that need to be considered in order to realise a more effective way of operation, interactive value creation, and interactive co-development for new products/services in the user's real life milieu.

The harmonization cube methodology focuses on the main elements that should be analyzed when evaluating LLs. The maturity of each element is measurable on a scale. The main elements of the interoperability cube assess the LL processes in detail, focusing on the main elements [11]. The key features of a LL are: immediate, interactive, co-creative user-involvement, iteratively involving all key actors across the stages. The methodology provides development opportunities for interactive, co-creative user-involvement methods. User involvement has different aspects that should be focused on, according to the maturity level of LLs. Methodical user involvement means firstly the identification of what the participating users are interested in, their behaviour, their roles, the identification of the user's motivations, the user's incentives, and the user's selection methods, e.g. by main segments or by interactions in the LL network. Once user-involvement is successfully realized in the initial phases, then the main focus is gaining the users' interests by managing the community dynamic of the LL management, and this serves for the further development of the LL.

The service creation concept helps to analyze the way of collaboration and communication between the LL stakeholders, idea generation services, training services and management. The aim is to create common and efficient collaboration and communication among the stakeholders and strong partnerships between actors, and to organize the LLs coherently and effectively. A well developed, enabling infrastructure, especially the ICT infrastructure, is a key element in order to realise user involvement and LL management effectively. The issues of ownership, IT infrastructure architecture and its efficiency are measurement targets.

The governance pillar of the methodology consists of realizing the co-operation and interoperability between stakeholders, enhancing the level of openness and creating stability and balance in funding. The methodical analysis of an LL consists of determining and enhancing the adaptation possibilities of worldwide applicable innovation methodologies and their possible related collaboration methods, user involvement tools, standards and best practices. The final results of an LL's activity are measured by modifications in product development processes, lifecycle shortening, perhaps IP agreements, as well as the ability to respond interactively to continuously changing user needs.

## 4.1 SME Involvement – an Additional Pillar to the Harmonization Cube

The harmonization cube has six sides. Building on this concept, a spider diagram was introduced by the CO-LLABS project. To assess its importance, it adds an additional pillar focusing on SME involvement in LLs.[1] SMEs can have both a user as well as a manufacturer role in an LL environment. Already existing LLs may provide the opportunity for SMEs to interactively test their products in advance, early in the development phase of the technology, by providing the service of customized testing according to the SME's unique requirements. As EC INFSO stated: "The underlying motivation is that LLs provide services to SMEs that otherwise would not be available to them. Focus is on how SMEs and their business partners can be involved in LLs in the best way in order to collaborate in open innovation, and on sharing experience among LLs initiatives and beyond as regards SME involvement in co-creation of LLs practices" [5]. The mentioned methodological pillars affect therefore the SMEs' business models. The main changes in business models derive from the necessary adaptation of user feedbacks and therefore change in the development processes in order to achieve a

[1] CO-LLABS [Community-Based LLs to Enhance SMEs Innovation in Europe] project focuses on describing LL best practices. Modes of their operation and methods of user-involvement are analyzed focusing on the SMEs.

more effective and easily variable infrastructure and working method. [Somewhat more can be found on the roles of LLs for SMEs from one author of this article in 'Open innovation for SMEs, 2010 [14]]

## 4.2 The User-Involvement Side of the Interoperability Cube

The key role of an LL, as already mentioned above, is to involve users effectively in the innovation process in order to make it more effective. In order to establish a productive co-creation process with the end users, appropriate user-involvement methods and tools should be chosen and adapted by the managers of LLs. We highlight the "user involvement" side of the harmonization cube in the following.

| | ORGANIZATIONAL ISSUES | CONTEXTUAL ISSUES | TECHNOLOGICAL ISSUES |
|---|---|---|---|
| **SETUP** | get users motivated | which type of user, effort, expectations required | provide tools to have users involved |
| **SUSTAINABILITY** | keep users motivated | need for unobtrusive methods | automatic data collection |
| **SCALABILITY** | different approaches to motivate different users | knowledge on cultural and legal differences | need for low cost observation methods |

Figure 4
The user-involvement side of the harmonization cube [11]

Based on the field of operation and the maturity of the LLs, different methods and tools should be considered and adapted, and the following main elements of the harmonization cube need to be negotiated. According to the user involvement side of the interoperability cube, three issues, organizational, contextual and technological, have to be taken into account when planning and adapting the user-involvement methods and techniques.

The organizational issue of the user-involvement is first of all about the motivation of end users to participate in the innovation process of an LL. Based on the organization's maturity level, whether at setup, sustainability or scalability, the goal of motivation might differ. In the initial phase of an LL, the motivation techniques have the goal of getting users motivated. Getting users motivated means the collection of relevant users, winning their interest in the work of the

LL, and making them enthusiastic about the user-involvement activity. In the more mature phase of a LL, when we have the motivated users actively participating in the innovation process, we should make efforts to sustain their interests and keep them motivated. In the last phase of LL maturity, we should focus on different users in different ways in order to keep them motivated.

The contextual issue of user-involvement is about creating the appropriate methods in order to set up a LL. It focuses on differentiating the users and is based on the groups; it deals with conducting analysis on their expectations in order to enhance their motivation level. In the sustainability phase of an LL, unobtrusive methods should be developed in order to keep the users motivated. In the scalability, the mature phase of an LL, even the users' cultural and legal differences should be considered.

The aim of the technological issues of user involvement is to provide tools for a reliable background for the co-creation activity. The method of user-involvement is mostly based on ICT technologies, providing a reliable background for data collection and the observation of all these processes at low cost.

## 4.3 User Involvement Techniques

There are several possible techniques to integrate users into the development process of products and services. The first method is to gain user opinions about a product or service, with a technique already used by marketing experts used interviews. Interviews can be differentiated by face-to-face, focus groups, informal conversation, a guided approach and even survey interviewing. Regarding LLs, preparing such an interview does not mean interactive user-involvement. It is used to get an early opinion from possible later users about a product or service. The emergence of internet and ICT technologies provides a useful infrastructural background for user-manufacturer interaction on a higher level. Virtual communities like chat room sites or online forums are virtual places where people exchange messages with other persons or a group of people. By commenting on each other's posts regarding a product or service, new ideas and directions of development might occur.

Usability testing is the next phase of user involvement, with the aim of measuring how well people can use the products or services for their intended purposes in an as realistic situation as possible to discover errors and areas of needed improvement.

In the frame of the so called 'COLLABS' project, there was a marketing research, a survey conducted with existing LLs in order to get an overview of their user-involvement methods. According to the survey, the most frequent methods for user involvement in the idea generation phase of products/services are interviews. According to the results, 90% of LLs use different kinds of interviews for this purpose, and even focus-group interviewing is used by 70% of LLs [13].

In the development phase of products/services, the dominant means of user-involvement are the usability and prototype tests. Workshops with customers as well as virtual prototype tests are adopted by 30% to 40% of the LLs [13]. These results show that actual LLs are still far from the prospective concept, the ideal of LLs.

# 5 Some Notes on LLs in Practice

## 5.1 The Aim of the Research

In order to analyse the operation of LLs, and the added value they bring, further user involvement techniques used in practice were researched in an empirical study conducted by Katalin Kovács, focusing on the added value of LLs in the renewable energy business.[2] For the purpose of this marketing type research, working LLs were analysed by executing internet search and, when necessary, semi-structured qualitative interviews based on the main elements of the harmonization cube. This research involved altogether 27 organizations using the "Living Lab" method in renewable energy innovation. We should emphasize that the LLs involved in the analysis target the development of sustainable, "green" buildings. Most of them also deal with the development of energy efficient technologies. The question and analysis categories were focused on the main added value of LLs, their role in renewable energy innovation and their contribution to the realisation of sustainable households. The core component of the LL analysis is about the role of users and the method of user involvement in the development process of technologies.

## 5.2 The Activity and Relevance of LLs in Renewable Energy Innovation

LLs are mostly initiated in the frame of government or EU programs. Their sustainability is established by the collaboration of companies, as the collaboration decreases their individual cost of development. Renewable energy LLs aim to decrease the fossil energy usage of buildings and to enhance energy efficiency.

2 Empirical research summarized in this paper is a part of the project executed in the frame of the fellowship program of the "IFZ - Inter-University Research Centre for Technology, Work and Culture", Graz in the 09.01.2012-30.04.2012. period. The research received funding from the Austrian Agency for International Cooperation in Education and Research, Centre for International Cooperation and Mobility, Austria (http://www.oead.at/). Project ID: ICM-2011-03428. Further results of the research executed in the frame of this fellowship program will be available in the journal "Environmental Innovation and Societal Transitions" probably in autumn 2013.

Most of these LLs deal with implementing working prototypes of renewable energy technologies into homes in order to gain experience of the end-users from their normal living environment. The information about the renewable energy technologies are gained by sensors using the smart grid concepts in the LL building. LLs monitor the interactions of the users with the innovations at home, in order to provide information for the researchers to test and evaluate the innovations. In sum, the aim of renewable energy LLs is the development of sustainable, renewable energy innovations by monitoring the interaction of users/customers with the implemented technologies in their homes. The LLs intend to prototype, validate and co-create innovations and try to execute longitudinal testing as well. For this purpose they use so called LL “test houses”, as well as attempt to install their technologies into homes over the long run.

## 5.3 User Involvement Techniques

The interactive co-creation with the end-users can be executed by the involvement of different groups of actors, involving also SMEs in the development process, on the supply side. The role of SMEs and research institutes from the construction industries is to provide inputs for testing and to stimulate discussion and creativity in such a way as to leave space for interactivity with the end-users. The role of LLs is then to organize co-creation with the users to gain information about the usage of new technologies in their normal, natural environment. Therefore, the main added value of LLs is the development and testing of technologies in this environment.

End users are therefore organized and involved to participate in the LLs in different ways and contexts according to the type of information companies would like to gain. When users are involved in the development process at an early stage, then interactive open innovation workshops are organized. End-users are recruited and motivated to participate, and LLs gather the opinion of users and their needs for new developments. Besides these interactive open innovation workshops, the interaction forms of LLs with end-users are mostly through interactive exhibitions of renewable energy technologies, where users can express their opinion and give feedback to the researchers. At the later stages of the development process, when a first prototype is already available, the co-development, co-creation and interaction can be executed interactively with the end users by using the technologies/prototypes every day in its normal working environment. Technologies are also tested in the homes in order to perform longitudinal research using observation equipment to collect the reports of the users.

Based on the interviews executed, the participation of end users and experts in LLs should be organized by taking into consideration their cultural background and their socio-economic context. The purpose of this is to enhance the amount of information gained for developing the technologies by using the LL model, as well as to ensure the engagement of a definite group on the renewable energy technologies.

## 5.4 The Effect of LLs on Renewable Energy Innovation

As a result these homes, as LLs, can demonstrate significant carbon reduction. Mostly universities, innovation centres, SMEs providing technologies for testing, local councils and citizens collaborate together to interact and co-develop innovative technologies. The smart-grid concept also helps users to optimize their energy usage.

LLs are a possibility for SMEs as well as research institutes and multinational companies to demonstrate their latest developments in an interactive forum and exhibitions in order to get critical opinions from the immediate interaction of the experts of the relevant fields and the end-users. Companies may gain long-term advantages this way and open new target markets. The LL is a strategic possibility for targeting different market segments with different products developed co-creatively with the users. There is an example for this, as "Habitat" LL especially targets the communities of low-income in order to target them with developments meeting their needs and to improve their living environment and the energy efficiency of their house, by way of learning from them interactively in an LL environment.

Involving users as co-creators into the development process might also enhance engagement and raise awareness of the community towards renewable energy technologies. Therefore they have a specific role in enhancing the social acceptance of these technologies. As has been well demonstrated by several studies, one of the significant problems of enhancing the usage of renewable energy technologies is their gaining social acceptance.

**Conclusions**

Open innovation as a new type of division of labour is already a widespread, global reality. It leads to horizontally networked firms where the networking [inclusion in, exclusion from the network] is subjected to a global competition. Many sorts of interactive relationships with the customers, including LL type, interactive value creation in an integrated production process are also quickly developing. The move towards consumer dominance in the innovation dynamic is envisioned by different analysts. But they conceptualise this dominance rather differently. A rather projected vision of what an LL could be sees this dominance as the leading role of users in their co-operation in the production process in the real life milieu where the solution for them must actually function. While the LL method is very important for product development, it is even more important in envisioning and realising services.

The research results on LLs in the renewable energy industry show that LLs provide several types of added value in innovation; among other benefits, they strengthen the commitment of renewable energy users as well as the immediate interaction among the main actors in the renewable energy industry. This article also highlights the main user involvement techniques and their usage in the renewable energy industry at different stages of the development process. We hope that these research results, having taken into account the industry specificities, can be utilized in other branches of industry as well. The future of LLs depends on the ability to realize the prospective concept of LLs instead of only /re/naming more or less testing procedures as LLs as often occurs nowadays.

**References**

[1] Amar Bhidé: The Venturesome Economy, How Innovation Sustains Prosperity in a More Connected World, Princeton University Press, 2008

[2] Anna Stahlbröst: Forming future IT – The Living Lab Way of User-Involvement, Ph.D dissertation, Lulea University of Technology, 2008

[3] C. K. Prahalad – M. S. Khrisnan: The New Age of Innovation, McGraw Hill Books, 2008

[4] Donald E. Stokes: Pasteur's Quadrant, Washington DC. The Brooking Institution, 1997

[5] EC INFSO: Living Labs for User-driven Open Innovation, An Overview of the Living Labs Methodology, Activities, Achievements, 2009, p. 19

[6] Eric von Hippel: Democratizing Innovation. MIT Press, Cambridge, 2005

[7] Esko Aho: Creating an Innovative Europe, Office for Official Publications of the European Communities, Luxembourg, 2006

[8] Hans Schaffers et al.: CO-LLABS Technical Report WP1, 2008, pp. 8-13

[9] Henry Chesbrough: Open Innovation: The New Imperative for Creating and Profiting from Technology. Boston, Harvard Business School Press, 2003

[10] Henry Chesbrough: Open Services Innovation, Jossey-Bass, San Francisco, 2011

[11] Ingrid Mulder, Daan Velthausz, Martijn Kriens: The Living Lab Harmonization Cube: Communicating Living Lab Essentials, eJOV Executive – The Electronic Journal for Virtual Organizations and Networks Volume 10, "Special Issue on Living Labs", November 2008, pp. 2-6

[12] Ingrid Mulder, Pieter Jan Stappers: Co-creating in Practice: Results and Challenges, Collaborative Innovation Emerging Technologies Environments and Communities, Proceedings, Leiden, 2009

[13] Jens Schumacher, Karin Feurstein: Living Labs – the User as Co-Creator; ICE Conference proceeding, Sophia-Antipolis, France, 2007, pp. 3-6

[14] Katalin Kovács: Open Innovation and SMEs, Living Lab Best Practices and the Role of SMEs in Interactive User-Involvement, ERENET [Entrepreneurship Research and Education Network of Central European Universities] Profile Issue 22, March 2011

[15] Mulder, I., et al: Towards Harmonized Methods and Tools for Living Labs. Paper to be presented at e-Challenges, The Hague, The Netherlands, October 24-26, 2007

[16] Pieter Ballon et al: Test and Experimentation Platforms for Broadband Innovation, SMIT, Brussels, 2005

# Measurement and Simulation of Energy Use in a School Building

**Snezana Dragicevic, Aleksandar Peulic, Miroslav Bjekic, Radojka Krneta**

University of Kragujevac, Faculty of Technical Sciences
Svetog Save 65, 32000 Cacak, Serbia
snezana.dragicevic@ftn.kg.ac.rs, aleksandar.peulic@ftn.kg.ac.rs, miroslav.bjekic@ftn.kg.ac.rs, radojka.krneta@ftn.kg.ac.rs

*Abstract: This paper presents the development of a wireless temperature monitoring system and the application of measurement data for computer model validation, and its application to the simulation of energy use in a school building in Cacak, Serbia. The system for temperature monitoring was realized with a GPS/GPRS (Global Positioning System/General Packet Radio Service) system for low power data acquisition, using an MSP430 Texas Instruments microcontroller. With respect to heat loss analysis, the continual measurement of ambient and inside temperatures with a sampling time of one hour has been performed. For the simulation, DesignBuilder software is used. The simulation model, which reproduces the temperature measurement of school buildings, was developed and tested for energy analysis. Results are used to develop generalized guidelines for the determination of the efficiency of energy saving measures and the evaluation of low-energy buildings.*

*Keywords: temperature measurement; energy efficiency; low power microcontroller; DesignBuilder simulation*

## 1 Introduction

High-energy consumption is one of the most serious problems in the world today. Recently, this topic has encompassed not only economic but also ecological and social importance. School buildings are typical structures, and their optimal energy consumption is a matter of public interest. Defining measures to improve the energy efficiency of school buildings is very important. There are several reasons for the implementation of priority measures for the energy optimization of school facilities: the number of these buildings, the need to maintain ambient comfort during the day and during most of the year, and the large amount of thermal and electrical energy used to maintain that comfort. Schools usually have a high level of energy consumption due to their considerable heating requirements and high electricity usage for lighting and equipment. The typical annual heating consumption of some European school buildings is reported as 96 kWh/m$^2$ in Ireland, 192 kWh/m$^2$ in Slovenia, and 157 kWh/m$^2$ in the UK [1]. Specific energy consumption in buildings in Serbia is 3-4 times larger, and there is great potential for an increase in energy efficiency.

The heating energy demand of a school building is influenced by construction, building services, weather, surroundings and the way it is used. The impact of the individual influencing factors is not considered. Stationary computer models of buildings are usually too coarse to identify details of all the temperature distributions or energy flows that are relevant for the consideration of energy demand and comfort [2, 3]. Furthermore, dynamic simulations can be very complex. However, in practice most of the numerous parameters cannot be determined precisely, because they are not fixed by the construction plans and/or are difficult to measure in a building [4, 5].

This paper presents the development of a wireless temperature monitoring system, and the application of the measurement data for computer model validation, which is used for the dynamic simulation of energy use in schools. Validated computer models for energy building simulation have an advantage in that the accuracy of the temperature prediction of the model over a short time scale is known. This allows the making of reliable predictions on time-dependent effects.

# 2 The School and Wireless Measurement Procedure

The selected school is a primary school, "Vuk Karadzic", in Cacak, Serbia. The building has a total area of 4,236 m$^2$, an external facade surface of 2,700 m$^2$ and a facade window area of 487 m$^2$. Measurements of internal and external temperatures are well suited for precise predictions of heating energy demand and comfort of buildings and suggestions of modifications to them [6]. Continual measurements of ambient and outside temperatures, with a sampling time of one hour, were performed during the winter months. The measurements of ambient temperature were performed in two classrooms: one with old windows with wooden frames (overall coefficient of heat transfer of 3.49 W/m$^2$K), whereas in the other there are new PVC windows (overall coefficient of heat transfer of 1.49 W/m$^2$K). These two classrooms have equal dimensions of 59.40 m$^2$ and very similar positions, so it can be assumed that other relevant thermal parameters are also almost the same. The surface area of the windows in both classrooms is 18.48 m$^2$.

The temperature measurement system is performed by a Global Positioning System/General Packet Radio Service (GPS/GPRS) based wireless acquisition system. The basis of the hardware part of the GPRS-based system for data acquisition (GPRSuC) from remote locations consists of a low power

MSP430F147 microcontroller and Telit GM-862 GPRS/GSM/GPS module (Fig. 1) [7, 8]. In addition to the standard functions of the devices used in M2M (Machine to Machine) communication, this module has a GPS receiver as well as a dedicated GPS port on which the data obtained from the GPS are shown in NMEA (National Marine Electronics Association) format. This makes the entire platform applicable not only for monitoring of the temperature, but also for the acquisition of data from measuring sites whose locations are not fixed. In addition to GPRS, GPS Telit is also able to send text messages through SMS (Short Message Service).

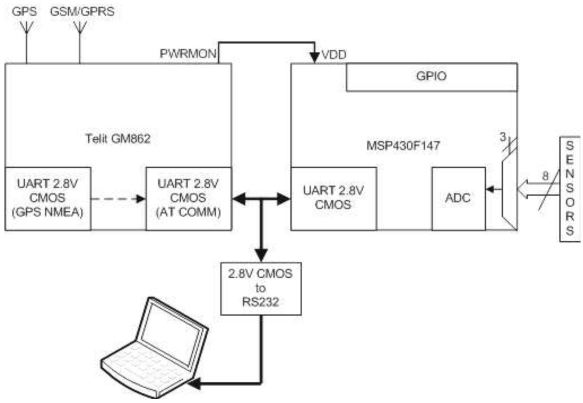



Figure 1
The structure of GPRS microcontroller system

Communication takes place via the AT port between the microcontroller on one side and the Telit module on the other using AT commands. GPS sentences are encapsulated in AT commands. AT commands are also used to send text messages through SMS. The system's operating range is -10 °C to + 55 °C, which can be a potential problem if GPRSuC is used in environments with very low temperatures.

A power supply voltage is appropriate for battery powered systems and it is set at 4.2 V. The GSM modem is made in a way that the RF transmission is not continuous; rather, it is packed into bursts at a base frequency of about 216 Hz. Consequently, relative current peaks can be as high as 2 A, and therefore the power supply is designed with a linear voltage regulator LM317K with a maximum continuous current of 2 A.

Microcontroller selection is conditioned by the requirements for small consumption and low voltage power supply. A maximum consumption of 560 uA at 3.0 V supply voltage and a frequency of the external oscillator of 32768 KHz make this microcontroller one of the best on the market from the point of view of power consumption. Low consumption is important because the microcontroller is powered over the PWRMON pin of the Telit module, which is normally used for an indication of operation. The PWRMON pin establishes a high voltage level, which provides sufficient power for the microcontroller MSP430F147, 900 ms after powering up the Telit module. Digital inputs and outputs are 2.8 V CMOS type, so it is not necessary to implement a circuit for level shifting between the microcontroller and UART peripherals of the Telit module. The microcontroller has a 12-bit, 8 channel mixed A/D converter with successive approximations.

The firmware is written in C, and its structure is shown in Figure 2. The Telit module acts as a slave carrying out the commands sent by the microcontroller.

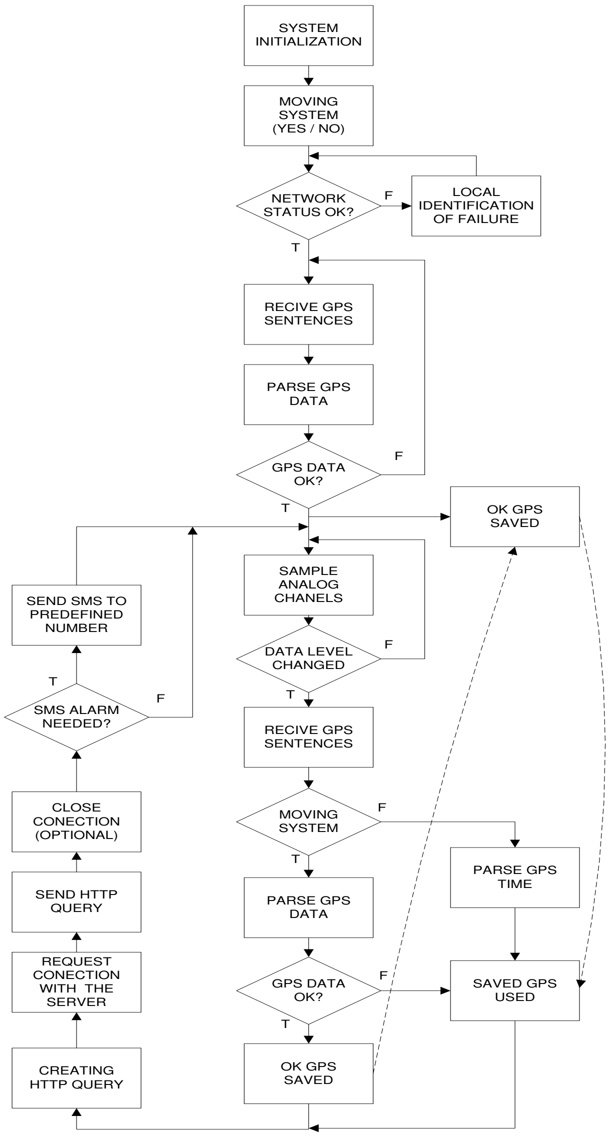



Figure 2
Flowchart of the firmware

Since the platform is designed to be suitable for the monitoring of mobile locations, GPS accuracy is very important. Since the system has an independent power supply and consumption is highest in GPRS communication, it is very important to decrease the level of communication without losing the significance of the data. A real system prototype is shown in Figure 3.

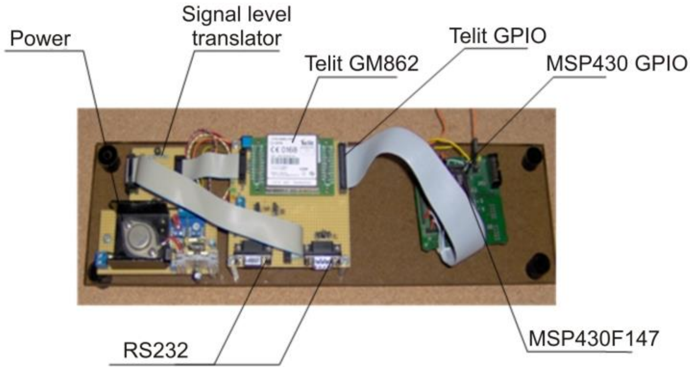



Figure 3
A prototype of temperature monitoring system

General Packet Radio Services' standard allows data transfer in a completely different way from the Circuit Switch Data (CSD) type of transmission. In CSD, data are transmitted by directly establishing a connection with other, remote modems, so all devices in-between are used to provide a simulation of the physical connection between the final points (point-to-point connection). Besides the obvious disadvantage in terms of low utilization of network resources, there are also problems of long delays in establishing a connection and high fees for the use of network resources, based on the time period of the established link, not the amount of data as with GPRS. One message from the remote acquisition system consists of eight measured parameters, a time stamp and identification field, and has nearly 110 bytes, so it is appropriate to use the GPRS system [9]. It is important to note that the system has no fixed IP address. Mobile operators provide a fixed IP address service, and it is possible to achieve communication in both directions with changes in the overall software of the system.

A practical part of this study is to present the collected data. It is only possible to pass parameters and values in some other way if a specified URL http request is made. The entire software solution has been realized using open source J2EE technology. The part of the software used to exchange data with the hardware is the interface, with no visual interpretation, and it is only executed when the HTTP request is passed to the Servlet by a device that forwards data. It has been developed as a Java Servlet, the hardware of which calls using the HTTP GET request to obtain the appropriate data Servlet, which performs the same processing and saves the results in the MySQL database. The Servlet can be accessed with the use of HTTP requests from anywhere in the world.

As the container, the Apache Tomcat version 6.0.16 is used. The project is implemented as a Web application, which is located on the server in the Computer Science Laboratory of the Technical faculty in Cacak, Serbia, and which provides current monitoring of temperatures and displays their values in real time. Since a system sends GPS coordinates, GIS support is used for better data presentation. As geographical support for the project, the Google Maps API is used. The principle of working with Google Maps API is that a complete GIS system is on Google's server. The user passes the coordinates and parameters for the display to the corresponding server, which replies by sending the required graphical content. Google Maps usage is free of charge and offers different types of presentation. Mouse clicking on the marker surface displays the current parameter values for the checking system (Fig. 4). The web application is built to monitor more than one data acquisition system. There is also a possibility of historical observation of measured parameters.

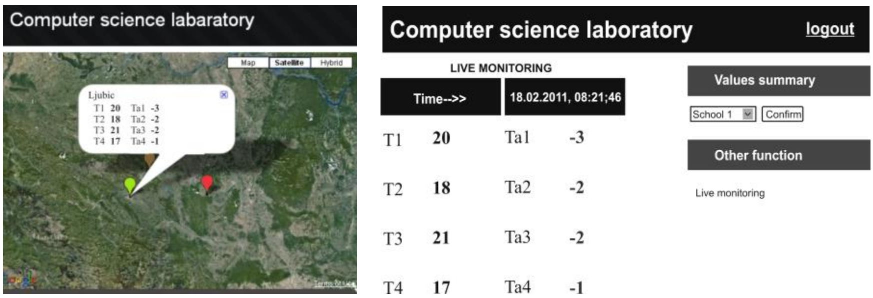



Figure 4

Multiple systems monitoring with data presentation on Google map and laboratory portal

# 3 Simulation Model and Validation

Building energy simulation is the analysis of the energy performance of buildings using computer modelling and simulation techniques. The calculation of building loads and energy consumption are used to determine the energy characteristics of the building and its building systems. Building energy simulation is employed to design the building to the requirements of local building regulations, codes and standards. Subsequently, it can supplement energy auditing to check the energy performance of the as-built building. With building energy simulation, complicated design problems can be investigated and their performance can be quantified and evaluated. It is also a useful tool for developing a better understanding of the building's performance [10].

For the simulation of the thermal behavior of the selected school building, the simulation software DesignBuilder[1] has been used (Fig. 5). This software is based on the Energyplus code and is provided with a 3D interface and meteorological database. It is based on the most popular features and capabilities of BLAST and DOE-2, as a stand-alone simulation.

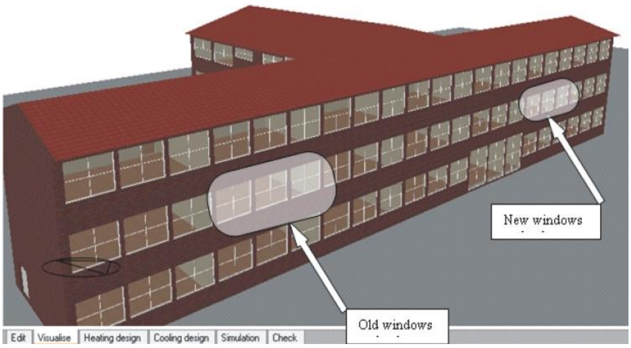



Figure 5
DesignBuilder model of school building

For all simulations, information on the wall, roof and flooring construction, as well as the windows and building services, are taken from the construction plans. The reproduction of the user behavior was necessary to yield the best possible validation of the simulation models. The basic parameters used for energy simulations are given in Table 1.

Table 1
Description of Classrooms

| **Overall** | | | | |
|---|---|---|---|---|
| Construction | Layers | Material | Thickness (m) | U value ($W/m^2K$) |
| External walls | Outermost layer<br>Layer 2<br>Layer 3<br>Innermost layer | brickwork<br>air gap<br>brickwork<br>plaster | 0.105<br>0.050<br>0.105<br>0.013 | 1.562 |
| Internal partitions | Outermost layer<br>Layer 2<br>Innermost layer | gypsum plastering<br>concrete block<br>gypsum plastering | 0.013<br>0.100<br>0.103 | 1.212 |
| Ceilings | Single layer | ceiling tiles | 0.020 | 1.692 |
| Glazing | 44% of gross floor area | | | 3.187 |

[1] http://www.designbuilder.co.uk

| **HVAC System** | |
|---|---|
| Thermostat Settings | 20°C heating |
| Schedule | 8:00 am - –6:00 pm, "off" night and weekends |
| **Internal gains** | |
| Occupancy | 3 $m^2$/person |
| Lighting | 3.4 W/$m^2$, 100 lux |
| Receptacles | <100 W (periodic usage of overhead projector, computer) |
| Schedule | 8:25 am– 3:30 pm weekdays, with some teacher occupancy before and after regular class hours |

Figure 6 shows the inside temperatures measured in the selected classroom, as well as the ambient temperatures. In the same figure, the results of the hourly temperature analysis in DesignBuilder are presented in order to compare the results of the simulations with measured data. Interior and exterior temperatures were predicted using DesignBuilder 1.8 software. All material and construction details, as discussed previously, have been applied to the simulation program, and a real weather data file for Cacak has been used in the simulation.

The good agreement of the simulations with the measurements from the temperatures analyzed indicates the successful modelling of the school building envelope presented and its heat losses. A reproduction of user behavior was necessary to yield the best possible validation of the simulation models.

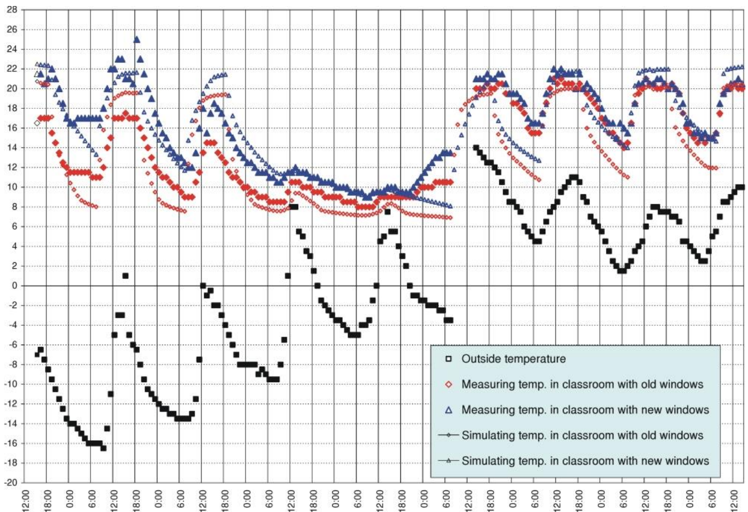
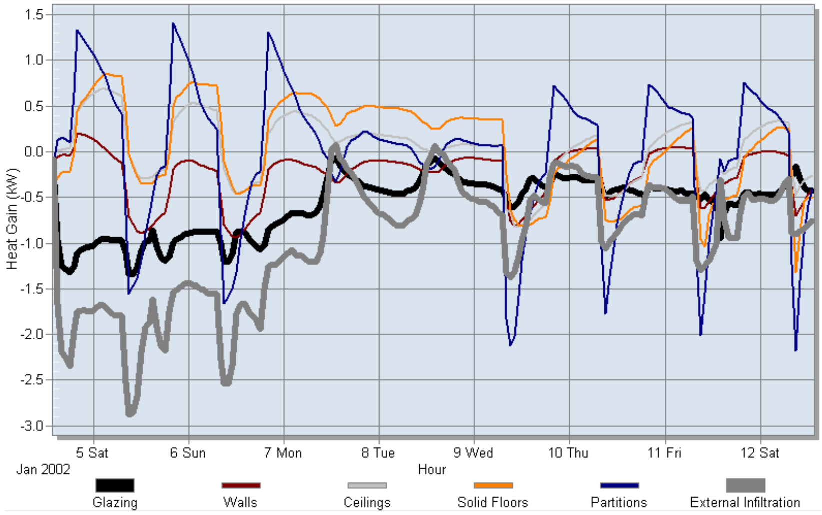
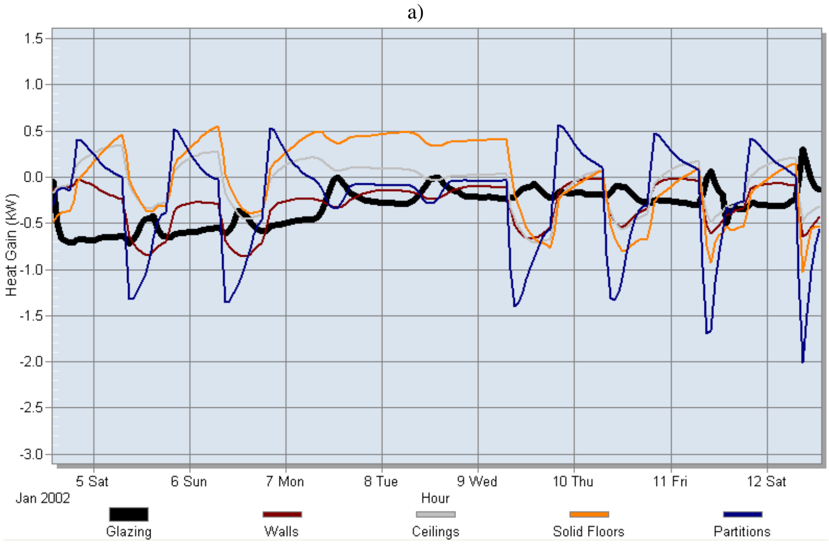



Figure 6
Simulated and measured indoor and outdoor air temperatures

In addition, dynamic simulations of heat losses from the school building are performed. The results from the DesignBuilder dynamic simulation include hourly heat losses and gains through windows, walls, ceilings, solid floors and partitions for classrooms with old windows and for classrooms with new windows (Fig. 7). Heat losses through windows with external infiltration are specially calculated.




a)



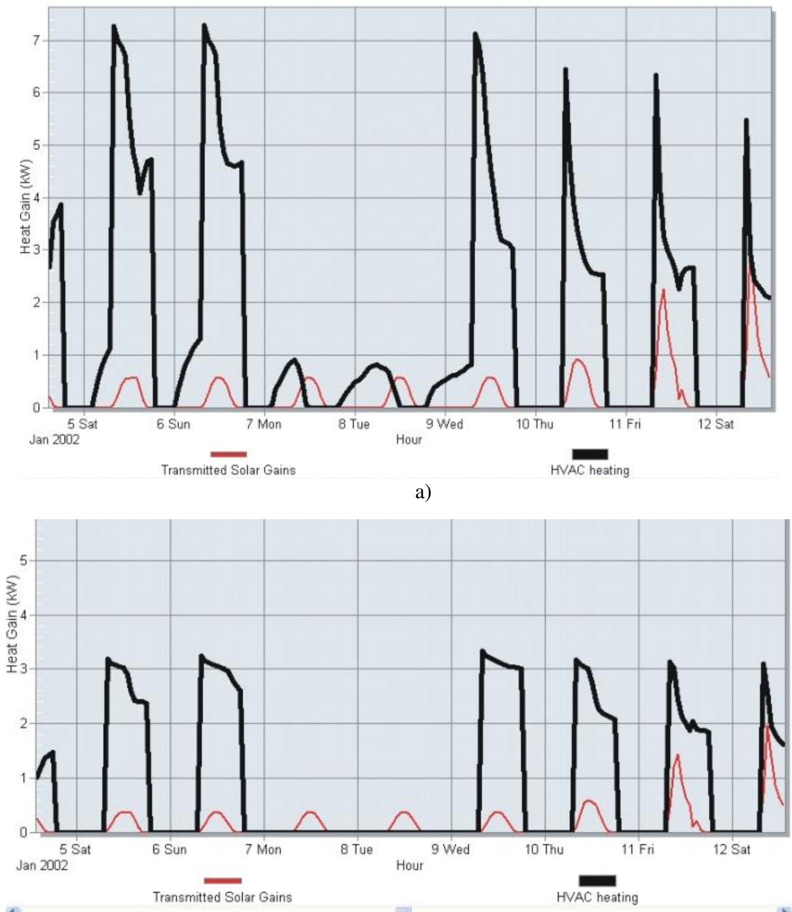

b)

Figure 7

DesignBuilder simulation results for classroom heat gains: a. Old window; b. New window

The replacement of old windows can effectively decrease heat losses by up to 50% for an ambient temperature from -5 °C to -15 °C, while for higher ambient temperatures the heat losses could decrease by up to 18%.

Transmitted solar gains and HVAC heating for the classrooms with both old and new windows are shown in Figure 8. Heat gains through windows are negligible because the simulation period did not include any sunny days.




a)

b)

Figure 8
DesignBuilder simulation results for classroom HVAC heating and solar gains:
a) Old window; b) New window

The simulation results confirm the adequacy of window replacement from a viewpoint of heating energy savings. At an ambient temperature of 0 °C, HVAC heating for the classroom with old windows is twice as great as the classroom with new windows; at -18 °C, the HVAC heating for the classroom with old windows is 2.41 times greater.

The results of the DesignBuilder dynamic energy simulation of the school building, for the cases with the old and new windows, monthly and annual energy consumption for space heating could be estimated. Table 2 shows average monthly heat energy consumption in the winter months for the school building with old and new windows.

Table 2
School heat energy consumption in the winter months (MWh)

| Month | Old windows | New windows |
|---|---|---|
| October | 0.315 | 0.040 |
| November | 6.784 | 2.897 |
| December | 23.917 | 16.368 |
| January | 26.502 | 17.799 |
| February | 16.690 | 10.087 |
| March | 2.102 | 0.289 |

Implementation of the proposed energy efficiency improvements in the school building would provide annual energy savings of 29 MWh, i.e. 37%, and reduction of the carbon-dioxide emissions of 10.5 tCO2 into the atmosphere per year.

## Conclusions

In this paper, a flexible measurement system for the temperature monitoring of a school building is proposed. The system communication is completely wireless, easily operable and has a low power requirement. Another consideration of this study was the suitability of the DesignBuilder software for simulating the thermal performance of the selected school. The developed and validated model is used for building energy simulation in order to analyze the effects of the replacement of windows. The building model is validated according to the energy demand and indoor temperature, and can, therefore, be used to analyze the impacts of various parameters on the heating energy demand and comfort. The results show satisfactory agreement with the measured temperature profiles for the analyzed classrooms. With this building model, the impact of the main influencing factors on the real heating energy demand can be analyzed quite precisely.

## Acknowledgements

This paper was supported by the Ministry of Education and Science of the Republic of Serbia within the project “Improvement of energy characteristics and quality of interior space in the buildings of educational institutions in Serbia and the impact on health” under N$^{o}$. 42008 (2011-2014).

## References

[1] Airaksinen, M.: Energy Use in Day Care Centers and Schools, Energies, 4, 2011, pp. 998-1009

[2] Desideri, U., Proietti, S.: Analysis of Energy Consumption in the High Schools of a Province in Central Italy, Energy and Buildings, 34, 2002, pp. 1003-1016

[3] Yiqun, P., Mingming, Z., Gang, W.: Whole Building Energy Simulation and Energy Saving Potential Analysis of a Large Public Building, Journal of Building Performance Simulation, 4, 2011, pp. 37-47

[4] Gieseler, U. D. J., Heidt, F. D., Bier., W.: Combined Thermal Measurement and Simulation for the Detailed Analysis of Four Occupied Low-Energy Buildings, 8th International IBPSA Conference, Eindhoven, The Netherlands, 2003, pp. 391-398

[5] Oka, S., Sedmak, A., Djurovic-Petrovic, M.: Energy Efficiency in Serbia - Research and Development Activity, Thermal Science, 10, No. 2, 2010, pp. 5-32

[6] Krneta, R., Dragićević, S., Radovanović, M., Bjekić, M., Energy Consumption in Schools in Serbia – Review of Present State in Cacak Municipality, Journal on heating, refrigerating and air-conditioning, Vol. 34, No. 1, 2005, pp. 49-53

[7] Alkar, A. Z.: An Internet-based Wireless Home Automation System for Multifunctional Devices, IEEE Trans. Consumer Electron., 51, No. 4, 2005, pp. 1169-1174

[8] Gutierrez, J. A., Naeve, M., Callaway, E., Bourgeois, M., Mitter, V., Heile, B.: IEEE 802.15.4: a Developing Standard for Low-Power Low-Cost Wireless Personal Area Networks, IEEE Network, 15, No. 5, Sept-Oct. 2001, pp. 12-19

[9] Alheraish, A.: Design and Implementation of Home Automation System, IEEE Trans. Consumer Electron., 50, No. 4, 2004, pp. 1087-1092

[10] Shetal, W., Sharples, S.: A Building Simulation Sustainability Analysis to Asess Dwellings in a New Cairo Development, Fourth National Conference of IBPSA-USA, New York City, New York, August 11-13, 2010, pp. 94-101

# Managing Rational and Not-Fully-Rational Knowledge

**Jozef Hvorecký, Jozef Šimúth, Branislav Lichardus**

Vysoká škola manažmentu, Panónska cesta 17, 851 04 Bratislava, Slovakia
jhvorecky@vsm.sk, jsimuth@vsm.sk, blichardus@vsm.sk

*Abstract: Knowledge management (KM) is a range of strategies and practices in organizations to identify, create, represent, distribute, and enable the adoption of insights and experiences. Knowledge is present in organizations in two forms: explicit (well-structured and unambiguously captured) and tacit (vague or informal, based on experience and beliefs stored in human brains). These two types split knowledge management into its "hard" and "soft" components. Each of them can contribute to an organization's development and prosperity but must be controlled in different ways. In hard knowledge management, the elements of knowledge, insight and experience are embedded in organizational processes and practice. To control them, traditional (rational) managerial approaches can be applied. However, people do not always act rationally. To benefit from knowledge embodied in individuals, more sophisticated strategies should be used. First we show examples of lower rationality studied by earlier researchers. Then, we use the SECI model to disclose situations deserving managers' special care. We demonstrate the presence of non-rationality in all SECI stages and exemplify its manifestations. Our conclusions can help managers concentrate on the core problems of knowledge management and apply it more efficiently.*

*Keywords: Knowledge management; Rationality; Irrationality; Emotions; SECI Model; Tacit & explicit Knowledge, Managing tacit knowledge*

## 1 Irrationality in Management

Since the Enlightenment, rational thinking and reasoning has been considered the indisputably best method of decision-making. In textbooks [e.g. 1, p. 631], rational methods are presented first. Usually they are divided into five stages:

- Recognition and definition of a problem or opportunity;
- Search for alternative courses of action;
- Gathering and analysis of data about alternatives;
- Evaluation of alternatives;
- Selection and implementation of a preferred alternative.

Formal approaches are effectively used as bases for designing and implementing of knowledge systems and expert systems [2], for constructing optimization algorithms, for designing robot's movements and for planning exactly specified processes.

On the other hand, recent research indicates that the level of rationality in human's decision-making declines with the growing level of managerial position. Isenberg [3] noted that in making their day-to-day and minute-by-minute tactical maneuvers, senior executives tend to rely on several general thought processes such as intuition; managing a network of interrelated problems; dealing with ambiguity, inconsistency, novelty and surprise; and integrating action into the process of thinking. Agor [4] came to a similar conclusion when he found that without exceptions, top managers in every organization differ significantly from middle and lower mangers in their ability to use intuition to make decisions on the job. He adds that women and people with Asian background demonstrate this trait more than other groups.

Glass [5] observed that intuitive approaches are neglected: "*Given our choice of decision-making techniques, most of us would use quantitative approaches first and rational ones second; intuition would come at or near the bottom of the list*". He shows that not-fully-rational decisions must be made even in fields in which rationality seems to be dominant, such as programming and software development and stresses: *We invent other names for it, cloaking its apparent irrationality in socially acceptable terms. For example, we speak of "gut" decision making—decisions coming from some deeply felt belief that goes against the grain of the environment surrounding the decisions*.

There exist examples of using not-fully-rational approaches in managerial activities regardless of their branch of industry or administration. Brunson [6] shows that choices are facilitated by narrow and clear organizational ideologies, and actions are facilitated by irrational decision-making procedures which maximize motivation and commitment. Guo [7] understands the problems caused by not-fully-rational procedures in strategic marketing but stresses their positive contribution, too:

*1) Irrational factors guide the strategy of the manager in gathering and compiling information.*

*2) Irrational factors support strategy on the part of the manager when carrying out information analysis.*

*3) Irrational factors help improve the efficiency and quality of decision making.*

*4) Irrational factors promote the smooth implementation of strategic decision.*

Dann and Pratt [8] believe that the non-conscious information processing system evolved early in humans and is based on automatic and relatively effortless processing of information. According to them, researchers view intuition as one of the products of this non-conscious, or automatic, system of information processing. They underline that the term "intuition" is used in different ways: either addressing the process running in human brain or its outcome or both. They propose an exhaustive list of its interpretations. It leads them to introducing the terms *intuiting* for the process and *intuitive judgments* for its outcomes.

In describing the intuiting, they identify its following features:

- It is subconscious,
- It involves making holistic associations,
- It is fast,
- It results in affectively charged judgments.

Their research neighbors to our field of interest – knowledge management – as they also take into account the domain knowledge factors, in particular the heuristic or expert ways of derivation of new knowledge from the existing one.

All of the authors underline the positive contribution of the not-fully rational approach in their research fields regardless whether they call them "emotions", "intuition" or "irrationality". In their research, the concepts do not mean "no wisdom" or "nonsense". They rather specify a kind of reasonableness not fully controlled by our erudition and formal reasoning.

In this paper, we study the role of such interpreted irrationality in knowledge management, its manifestations and implications to managers' activities.

# 2 Knowledge in Managers' Perception

Defining something as subtle and intangible as knowledge is almost impossible. The likely best way is posting a series of alternative definitions, each of them applicable under different conditions. The Oxford English Dictionary's definition [10] specifies knowledge as:

- Expertise and skills acquired by a person through experience or education;
- What is known in a particular field or in total;
- Theoretical or practical understanding of a subject;
- Awareness or familiarity gained by experience of a fact or situation;
- Facts and information.

The first description is appropriate when describing the capabilities of educated persons; the second is more appropriate for defining a (scientific) discipline; the third one speaks about the "depth" of our knowledge. The fourth indicates that knowledge can grow, while the fifth says that it can also be expressed in fixed terms.

Knowledge helps people solve their problems. Different problems require distinct pieces of knowledge, as Table 1 shows. Almost all elements of knowledge mentioned above are illustrated in the table: expertise (in surgery), familiarity with the situation (setting up the diagnosis), understanding of the field (taxation), facts and information (tax ranges), theoretical and/or practical understanding of the subject (cleaning the oven).

Table 1
Examples of knowledge necessary for solving problems

| Problem | Knowledge | Solution |
|---|---|---|
| Calculating salaries | Person's income,<br>Tax regulations,<br>Calculations | Net Income,<br>Tax |
| Appendicitis | Setting up the diagnosis,<br>Surgery experience,<br>After-operation treatment | Healthy patient |
| Dirty oven | Household skills,<br>Detergent application | Clean oven |
| Talking to dead | Spiritual practices | Evoked ghosts |

*Source: Hvorecky & Kelemen [11]*

As one can also see, the validity of the solution does not depend merely on the individuals' gained information, experience, and skills, but also on their beliefs. Only people who believe in ghosts are ready to consider "spiritual practices" as a category of knowledge; others will likely complain. The opponents will also reject the both proposed methods and solutions because (according to their conviction) spirits cannot be evoked.

The reason for the dichotomy is rooted in different quality of our knowledge. Some of its pieces are undisputable (e.g. $1 + 2 = 3$) as they have been verified to the maximum human potential. All professions have their quantum of *unconditional knowledge* – the basis of the discipline. Some pieces of knowledge are *accepted* as correct by mutual agreement. For example, according to the ISO standard, the boiling point of water is $100^{o}$ C, while the imperial system posts it at $212^{o}$ F. Finally, some elements are based on individual faith only, such as the above-mentioned existence of ghosts. They can be characterized as our *believed* knowledge.

Using their personalized pieces of knowledge, different people tend to process the same situation differently and to act in different ways. Managers should be aware of these differences as undisclosed substantial divergence among them may harm any team's cooperation. For example, the Mars Climate Orbiter space experiment crashed due to discrepancies in computations of its orbit, caused by mixing up the metric and imperial measuring systems [12].

The necessity of managing the variety of pieces of knowledge dispersed in humans led to the birth of knowledge management – a discipline studying the optimal approaches to its exploitation in organizations.

# 3 Knowledge Management

Knowledge Management (KM) is a range of strategies and practices organizations use to identify, create, represent, distribute, and enable the adoption of insights and experiences [13]. In accordance to this discipline, knowledge is present in two forms [14]: *explicit* (well-structured and unambiguously captured) and *tacit* (vague, informal, and based on experience and beliefs).

Explicit knowledge can be stored on paper, media or by other appropriate means. Computer programs represent the most advanced forms of explicit knowledge today. Knowledge is transformed into the abstract machines capable of getting data from their environment, remembering and processing them and producing results. The most advanced forms mimic human reasoning. Transferring our explicit knowledge into production lines, machines and robots is one of the principal trends of contemporary science and technology. Mathematical and chemical formulas, optimization and validation methods, recipes and operational instructions – all are examples of explicit knowledge. Their sources and bearers can be quite easily identified. This allows us to implement this part of knowledge management more easily.

Tacit knowledge is stored in human brains only. One can register its presence only when it is applied. An example is the interpretation of statistical data. Different individuals are likely to read the same data in different ways depending on their experience, familiarity with the controlled environment, emotions, political views, etc. Even if some guidelines on interpreting statistical data can be proposed, there is no universal method, and the result depends substantially on its interpreter. Occasionally, the person might not even be aware of possessing a piece of tacit knowledge. Then, his/her activity may seem random and its outcome simply good luck.

Tacit knowledge represents a considerable portion of our knowledge. In accordance to Abidi et al [15], it may contribute to two thirds of the decisions of logistics workers. Other authors [16] estimate its prevalence as being as great as 90:10. Due to the hidden character of tacit knowledge, similar ratios represent estimations based on their particular author's conviction.

Two types of knowledge divide knowledge management into its "hard" and "soft" sections. In hard knowledge management, the well-specified, exact and captured elements of knowledge, insights, and experiences are embedded in organizational processes and practice. Traditional (highly rational) managerial approaches can be applied for their control. One of the aims of hard knowledge management is to transfer selected elements of tacit knowledge into explicit ones using of knowledge engineering [2].

Below, we concentrate on the management of tacit knowledge. We show that it is still possible to apply some traditional methods to it. For example one can map "who knows what" – naturally with the limitations rooted in lower chances to verify its existence and depth and in the impossibility to catch it *all*.

# 4 The SECI Model

Our research has a methodological character. Its main aim is to build a supporting tool for the "who knows what" mapping. As the "who" part is indivisibly tied to the particular organization and its goals, we concentrate on "what" components and develop a list of capacities to be searched for. Such a list might serve as a guideline during the mapping.

The SECI model [17] is a subject-independent model describing the relationship between explicit and tacit knowledge. Its original purpose is to demonstrate the way of knowledge development inside organizations.

| | TACIT KNOWLEDGE | EXPLICIT KNOWLEDGE |
|---|---|---|
| TACIT KNOWLEDGE | **Socialization** | **Externalization** |
| EXPLICIT KNOWLEDGE | **Internalization** | **Combination** |

Figure 1
The SECI Model

Bearers of tacit knowledge interact with bearers of (possibly different) tacit knowledge during *Socialization*. It is performed by interpersonal communication and/or intrapersonal insights. This is the most traditional form of learning and is present in any human community.

To achieve a person-independent knowledge, people try to express their internal understanding of objects and methods in a commonly accepted way using various forms of *Externalization*. That results in discussion of a subject in a standardized, comprehensible format. These presentations (numbers, texts, graphs, formulas, etc.) create a basis for the wider distribution of knowledge as the "dialogue" between the author and consumer of the piece of knowledge does not depend on their geographic location and time distance.

The pieces of knowledge expressed in their formal notation can be processed by their receivers. Such *Combination* may lead to new pieces of knowledge. Computers and robots are also capable of executing combination when it is incorporated in their controlling programs. On the other hand, a machine-performed combination represents just a part of all actions in this area because people perform intellectual activities which belong to this category and cannot be executed by computers yet.

In the last stage, people try to interpret the outcomes of their activity and want to comprehend them. Through *Internalization*, the new piece of knowledge becomes an integral part of our individual knowledge ready for its future application.

The knowledge-acquiring processes runs:

- *Inside each of the four quadrants*: During Socialization, we learn by communicating thoughts and experiences with our partners. We absorb their style of thinking, study their mentality and effects of emotions on them etc. During Externalization we learn to visualize our ideas and demonstrate them in a legible manner. We study which approaches are successful and which are not; we remember them in order to excel later. During Combination we learn to control our moves to get fair results and search for more efficient combination methods. During Internalization we adjust the new piece of knowledge into our already existing knowledge system and start comprehending its role in it.
- *The clockwise order* indicated by the initial letters S-E-C-I shows that learning runs in cycles. The ideas are born in our minds (S). Then we try to express them in a more concise way (E). This preliminary outcome is then elaborated in order to test its validity, acceptability and usefulness (C). Finally, we "shape" the new piece to a contour changing it into a part of our internal knowledge weaponry (I).
- *Eternally*: Every new piece of knowledge is presented to the community and discussed. Its "socialization" begins and may lead to new ideas. As a result, the knowledge processing acquires the form S-E-C-I-S-E-C-I-S-…

Thus, the SECI model shows a life-cycle of knowledge with its multiple reincarnations. Due to its eternal elaboration, our knowledge turns out to be deeper, wider and abundant with the time. The model also demonstrates the evolutionary character of our knowledge which constantly switches between explicit and tacit ones.

# 5 Tacit Knowledge as a Regular Part of Knowledge-related Processes

The conviction of some authors about the prevalence of tacit knowledge over the explicit variety is likely based on the fact that tacit knowledge must be applied even during Combination, which is seemingly a pure "explicit – explicit" activity.

For example, all mathematical calculations are not just formal manipulations with its symbols. One must know which formulas are applicable, which of them will be the most likely lead to the result, how to order them, whether the result is meaningful, and so on. Table 2 shows some of the not-fully rational activities accompanying the particular SECI processes.

Table 2
Not-fully-rational activities behind the SECI model

| **Socialization** | **Externalization** |
|---|---|
| Story-telling<br>Discussion<br>Opposing common opinions<br>Listening to other opinions<br>Showing example behaviour<br>Teaching and training<br>Brain storming | Speaking and writing excellence<br>Capturing of the idea's core<br>Formalization<br>Introduction of a new notation<br>Posing "right" questions<br>Demonstrating skills |
| **Internalization** | **Combination** |
| Digesting of a new piece of knowledge<br>Practicing a new activity<br>Implementing a problem solving method<br>Learning a new formal notation<br>Becoming involved in the topic<br>Understanding potential "usefulness" of knowledge | Lateral thinking<br>Creating analogies<br>Selection of the right knowledge processing method<br>Identification of the new piece of knowledge |

Notice that our list is not exhaustive. It rather contains typical examples derived from the authors' experience. Depending on the type of the organization and the character of its functions, some of the activities are more important than others. To manage tacit knowledge, therefore, means completing the above list in accordance with the aims of the organization, to select principal "irrational" activities and to set up their priorities in human resource development and its control.

# 5 Introducing the Management of Tacit Knowledge

Realize that the tacit knowledge management processes include a large portion of rationality. In other words, managing irrationality can be a rather rational process. Still, compared to the management of explicit knowledge, managers must apply different approaches. Their methods must be less direct and more based on cooperation with their employees. A typical strategy can be specified using the medical sequence "prevention-diagnosis-treatment".

## 5.1 Prevention

The prevention starts with *selecting appropriate employees* to knowledge worker positions. They must be not only educated and skilled, but also interested in their profession and motivated to cooperate on the goals of the organization. Their job duties must offer them enough room for grasping and digesting additional knowledge and experience and for discussing their validity and value. Contemporary human resource management is aware of some of factors shown in Table 2 and pays attention to them. But it does not focus on all variables. For example, how do recruiters assess an applicant's courage to oppose common opinions? When they disclose it, will they accept it as a positive feature of the applicant or not? How will they decide whether the applicant would use this skill appropriately at work?

Some tacit knowledge elements can be verified quite easily, despite their not-fully-rational character. For example, speaking and writing skills can be tested with a relatively high precision. However, some others are not as easy to measure, e.g. the ability to capture the core of ideas. Some skills can only be judged by relevant specialists, e.g. quality of design. Again, due to the irrational character of the notion of quality, their judgment can still be misleading and can occasionally be neglected. In such cases, the reason for the refusal should be a rational one. Some elements of tacit knowledge may not be tested simply because even their owner is unaware that he or she possesses them, e.g. the ability to act as an example.

*Creating a proper and comfortable working environment* is another presumption. It should support the exchange of informal thoughts and an atmosphere of relevant criticism without fear of consequences. The team as a whole should cover the list of expected explicit and tacit knowledge and skills. Every team member should be familiar with the other team members' capability to execute them as well as with their priorities in their execution.

In order to form an efficient and effective team, *the triple character of knowledge has to be considered*:

- All unconditional knowledge must be present. If its pieces are absent, the team efforts must be concentrated on acquiring them, otherwise the task cannot be finished.
- Prior to the commencement of work, there must be an exhaustive mutual agreement on all task-related elements of accepted knowledge. Inconsistencies among them might lead to fatal consequences.
- Finally, the individuals' believed knowledge must be harmonized. If, for example, some team members question the ethical aspects of their work, their frustration could negatively influence their involvement up to the level of its sabotage.

## 5.2 Diagnosis

As a result of the untouchable character of knowledge, the significance of the knowledge worker's knowledge, skills and experience cannot be fully assessed during the prevention process. It can be only completed when its first results are presented in practice. Their evaluation should take into account their "useful irrationality", i.e. on the assessment of the knowledge workers' creativity and its application in their duties. Thoroughly designed and constantly monitored diagnostic methods must become a tool of every knowledge worker's manager. They must be based on a regular observation and evaluation of the work progress. They must be job-oriented and flexible enough to reflect the creative and innovative characteristics of the tasks.

The manager must control whether the particular team member is capable of accomplishing his/her duty and/or to fix what was done wrong. As the manager can hardly be a specialist in all specific fields, the situation can result in an absolute dependence of the manager on his/her subordinate(s). Managers without good discussion skills and patience to listen to the knowledge workers' opinions will be in a tough situation.

## 5.3 Treatment

Managing tacit knowledge requires permanent collaboration and mutual support. Flattening is a typical strategy of contemporary organizations. Their authoritarian hierarchies are changing. Instead of giving and receiving orders, control in organizations is mostly based on trust. In organizations having a rigid and/or conservative culture, building "islands of positive deviation" is a possible strategy. Such cells should demonstrate model behavior. The top management should value and reward them in order to demonstrate its devotion to their principles.

In general, the proper strategies are based on facilitating the knowledge workers' motivation rather than on giving orders to them. Success also depends on the quality of diagnostic methods because early-discovered problems can be fixed with less pain. The treatment methods must accent collaboration, freedom to act and feelings of responsibility for the outcome in each team and all of its members.

**Conclusions**

In their function, organizations apply many concepts and procedures which are not rational in the traditional sense. Knowledge management must accept this fact as a basis for its approaches. During our analysis, we have identified several not-fully-rational elements. All of them relate to tacit knowledge. The SECI model helped us to identify abilities and skills that support individual processes that participate in creating new knowledge. There is not (and can hardly ever be) a comprehensive list here; many others will be found in the future. As irrationality is a natural part of human nature, its position must be better specified; otherwise, knowledge management will not reach its potential. Consequently this would not bring its prospective benefits.

In our future research we will try to disclose them and, hence, to identify more specific methods of their diagnosis and treatment. We therefore understand our described research as a pilot study. Based on its results we have formulated our recommendations that could be instrumental to managers in focusing on the core problems of tacit knowledge management. In agreement with the authors named in our first chapters, we believe that even not-fully-rational aptitude of our brains can be developed by appropriate learning and training methods. Their combination with effective tacit knowledge management can help the organizations to promote and prosper.

## References

[1] David A. Buchanan, Andrzej A. Huczynski: Organizational Behavior. Prentice-Hall, Pearson Education, Harlow UK, 2010

[2] Jozef Kelemen et al: Knowledge in Context. Iura Edition, Bratislava, 2010

[3] Daniel J. Isenberg: How Senior Managers Think. Harvard Business Review, November-December 1984

[4] Weston H. Agor: How Top Executive Use Their Intuition to Make Important Decisions. Business Horizons, January-February 1996

[5] Robert L. Glass: Intuition Role's in Decision Making. IEEE Software, January/February 2008, pp. 95-96

[6] Nils Brunsson: The Irrationality of Action and Action of Rationality: Decisions, Ideologies and Organizational Actions. Journal of Management Studies, Vol. 19, Issue 1 (Jan. 1982) pp. 29-44

[7] Yongbo Guo: Study on Irrational Strategic Management. International Journal of Marketing Studies, Vol. 1, No. 2 (2009) pp. 92-95

[8] Erik Dane, Michael G. Pratt: Exploring Intuition and Its Role in Managerial Decision Making. Academy of Management Review, 2007, Vol. 32, No. 1, pp. 33-54

[10] Simpson, J., Weiner, E. (eds.): Oxford English Dictionary. Oxford University Press, 1989

[11] Jozef Hvorecký, Jozef Kelemen (eds.): Readings in Knowledge Management. Iura Edition, Bratislava 2011

[12] Wikipedia: Mars Climate Orbiter. http://en.wikipedia.org/wiki/Mars_Climate_Orbiter (Accessed on 8 June 2012)

[13] Wikipedia: Knowledge Management

http://en.wikipedia.org/wiki/Knowledge_management (Accessed on 8 June 2012)

[14] Paul M. Hildreth, Chris Kimble: The Duality of Knowledge. Information Research, Vol. 8, No. 1, October 2002

[15] Hella Abidi, Matthias Krump, Thomas Keuschen: Industry Qualifications Framework Ligistics: Explicit and Tacit Knowledge and Qualifications in Logistics Industry. In: Thorsten Blacker, Carlos Jahn, Wolfgang Kersten (eds.): Maritime Logistics in Global Economy, Josef Eul Verlag, Köln, 2011, pp. 327-345

[16] Beliz Ozorhon, Irem Dikmen, M. Talat Birgonul: Organizational memory formation and its use in construction. Building Research & Information, Volume 33, Issue 1, 2005, pp. 67-79

[17] Nonaka, I., Takeuchi, H.: The Knowledge-Creating Company – How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, London, 1995

# Combining Co-Training with Ensemble Learning for Application on Single-View Natural Language Datasets

**Jelena Slivka, Aleksandar Kovačević, Zora Konjović**

Computing and Control Department, Faculty of Technical Sciences
Trg Dositeja Obradovića 6, Novi Sad, Serbia
E-mail: slivkaje@uns.ac.rs, kocha78@uns.ac.rs, ftn_zora@uns.ac.rs

*Abstract: In this paper we propose a novel semi-supervised learning algorithm, called Random Split Statistic algorithm (RSSalg), designed to exploit the advantages of co-training algorithm, while being exempt from co-training requirement for the existence of adequate feature split in the dataset. In our method, co-training algorithm is run for a predefined number of times, using a different random split of features in each run. Each run of co-training produces a different enlarged training set, consisting of initial labeled data and data labeled in the co-training process. Examples from the enlarged training sets are combined in a final training set and pruned in order to keep only the most confidently labeled ones. The final classifier in RSSalg is obtained by training the base learner on a set created this way. Pruning of the examples is done by employing a genetic algorithm that keeps only the most reliable and informative cases. Our experiments performed on 17 datasets with various characteristics show that RSSalg outperforms all considered alternative methods on the more redundant natural language datasets and is comparable to considered alternative settings on the datasets with less redundancy.*

*Keywords: semi-supervised learning; co-training; ensemble learning; genetic algorithm*

# 1 Introduction

Semi-supervised learning is a class of machine learning techniques that employs both labeled and unlabeled data for training. The goal of semi-supervised learning is to achieve high accuracy while demanding less human effort. In this paper we focus on the major semi-supervised learning method called co-training [1]. Its successful application is guaranteed when the dataset possesses a natural division of the features in two disjoined subsets, called view, such that each view is sufficient for learning and conditionally independent to the other given the class label. This makes co-training application limited because the needed optimal feature split is unknown in the great majority of settings.

In this paper we propose a novel co-training based method, called the **R**andom Split **S**tatistic **alg**orithm (RSSalg), which exploits the advantages of co-training and is exempt from co-training requirement for the existence of adequate feature split. In our previous work [2, 3, 4] we published promising preliminary results for our RSSalg methodology. However, the disadvantage of our earlier method was the introduction of threshold parameters to the co-training setting, which greatly affects the performance of RSSalg and needs to be carefully manually tuned for its successful application. In this paper we develop a methodology for automatic determination of these thresholds.

This paper is organized as follows. Section 2 presents the related work. Section 3 describes our RSSalg methodology. Section 4 presents the conducted experiment and achieved results. The results are discussed in Section 5. Finally, Section 6 concludes the paper and outlines the possible directions of future work.

# 2 Related Work

One way to enable co-training application on the single view datasets is by designing a methodology that can provide a good approximation of the optimal feature split. This approach resulted with some promising results [5-10]. However, approximating the optimal feature split is a difficult task, as the relation between the characteristics of the views and the performance of co-training has not been sufficiently investigated. Moreover, research [10] indicates that in the real-world situations where co-training would be most useful, that is, in the situations where we posses only a small training set, the independence and sufficiency assumptions of the views cannot be reliably verified. In these situations, split methods can be unreliable, and thus the performance of co-training is uncertain as it may degrade the classification performance when its assumptions are not met [11].

In this paper, we compare our proposed method to a co-training algorithm that uses a random split of features, shown to be beneficial in case of redundantly sufficient feature sets [5, 6].

Also, we compare our method to a method called maxInd [7]. MaxInd approximates the optimal feature split for co-training by crating two maximally independent views, given the class label. This approach is based on the conditional independence requirement for the views [1]. Experimental results with maxInd have shown that the prediction accuracy of co-training does not always become better by simply choosing truly independent views, leading the authors to conclude that the relation between the characteristics of views and the performance of co-training should be investigated more in detail. These findings have also been confirmed in [8], where the authors randomly generated a number of splits and tested the most independent splits against the least independent ones.

Another way researchers approach the problem of enabling co-training application on single view datasets is by combining co-training with ensemble learning [12]. This approach is mainly based on using an ensemble of classifiers in the place of two single classifiers in the co-training algorithm [13-15]. Methods based on this approach are able to significantly boost the performance of co-training. However, they usually rely on having a relatively large initial training set in order to build the initial ensemble of diverse and accurate classifiers. Moreover, in the case of a high dimensional dataset, classifiers trained on small bootstrapped data samples using single feature view may face the 'large *p*, small *n* problem' [16, 17]. One of the goals of our setting is to combine the advantages of co-training style algorithms and ensemble learning while keeping the training set at the same size as would be used in the original co-training setting (just a few labeled examples). Instead of employing multiple classifiers inside the co-training algorithm, our methodology exploits different configurations of co-training in order to create an ensemble of classifiers from only a few labeled examples.

In [18] it is argued that ensemble learning and semi-supervised learning can be mutually beneficial. Based on this, some approaches exploit unlabeled data for ensemble diversity augmentation [19, 20]. As opposed to other ensemble approaches, our methodology combines multiple co-training classifiers. In this way, a hierarchical ensemble is constructed: each co-training consists of two base classifiers, and our RSSalg methodology combines multiple co-training classifiers. In [21] we have also explored an alternative way of combining multiple co-training classifiers by treating them as inconsistent and unreliable annotators in an unsupervised multiple-annotation setting.

# 3 Methodology

In our setting we are given a training set of labeled examples *L*, which is relatively small and a set of unlabeled examples *U*, which is relatively large. Our goal is to determine the unknown labels of the instances in the given test set *T*.

The first step in our Random Split Statistic algorithm (RSSalg) is to create an ensemble of *m* diverse co-training classifiers $\{CL\}_1^m$. This is achieved by running co-training until its termination *m* times on a given dataset, using a different configuration for each of the m runs in order to get a different classifier. Each time co-training is run independently of the other co-training runs and each run of co-training uses a different random split of features, thus producing a different co-training classifier. A random feature split is created by random selection of half of the features from the feature set as the first view, and the remaining half of the features is treated as the second view. Each independent run of co-training produces a different pair of base classifiers, as each time a different feature split is used. Based on their confidence, each pair of base classifiers selects different instances from the unlabeled set for labeling. Thus, it may happen that in different runs of co-training algorithm different instances are selected for labeling. Also, it may happen that the pairs of base classifiers from different co-training runs give a different label to the same instance. Consequently, in each co-training process a disparate enlarged training set $L$ (consisting of initial labeled data and data labeled in the co-training process) is formed. We will refer to an enlarged training set created during the $i^{th}$ co-training run as co-training result set $Lres_i$.

The second step in RSSalg is to form the final classifier by training it on the best instances from gained $m$ co-training result sets $\{Lres_i\}_i^m$. The statistic $S$, based on co-training result sets, is created: for each $Lres_i$ and each instance $e$ which is a member of $Lres_i$, the number of times instance $e$ occurs in all co-training result sets $n_e$ is determined (equation 1), and for each class $C_k$ we count the number of times instance $e$ is labeled $C_k$ (equation 2).

$$n_e = |\{Lres_i \mid e \in Lres_i, i \in \{1..m\}\}| \quad (1)$$

$$n_{eCk} = |\{Lres_i \mid e \in Lres_i \wedge label(e) = C_k, i \in \{1..m\}\}| \quad (2)$$

Each instance $e$ is assigned its most voted label $C$, determined by a majority vote of co-training base classifiers $\{CL\}_i^m$ (equation 3).

$$n_{eC} = \max(n_{eCk} \mid k \in \{1..m\}) \quad (3)$$

Falsely labeled instances would introduce noise into the training set for the final classifier. Our assumption is that the instance is reliable (in terms of high probability of being assigned the correct label) if the majority of co-training base classifiers $\{CL\}_i^m$, created in the first step of RSSalg, agree on the label of that instance. Thus, for each instance $e$, the label agreement percent $e_{agg}$ is calculated as follows: $e_{agg} = n_{eC}/n_e$. An instance is marked as reliable if its label agreement percent exceeds the defined label threshold $l_{ts}$.

In the extreme case, an instance might occur only once in co-training result sets ($n_e$=1), and label agreement percent of that instance would be 100%. However, we consider this instance uninformative, as there is no other co-training classifier which would contradict or agree on its label, and labeling an instance based on the prediction of only one co-training classifier, which can possibly be of poor performance, is unreliable. Therefore, another condition for selection of instances in the training set for the final classifier is defined: an instance must be informative. Our assumption is that instances that appear in most of co-training resulting sets are informative (in terms of having good bases to improve the learning process). To define which instances appear in most co-training resulting sets, an example threshold $e_{ts}$ is used. Let $n$ be the number of different instances noted in statistic $S$. For each instance $e$ from $S$ we calculate the occurrence percent $e_{occ}$ as follows: $e_{occ} = n_e/n$. An instance is marked as an informative instance if its occurrence percent exceeds the defined example threshold $e_{ts}$.

Thus, if we had optimal values of label threshold and example threshold, we could train our final classifier on instances marked both informative and reliable, where each instance would be assigned its most voted label. The result of the RSSalg is this final classifier which can be used to label previously unlabeled data. Pseudo code of RSSalg is presented in Figure 2.

Given:

- A small set $L$ of labeled training examples
- A much larger set $U$ of unlabeled examples
- A set $T$ of labeled test instances (used for model evaluation)
- Label threshold $L_{ts}$ and example threshold $E_{ts}$

for $i = 1..m$ iterations:

- Create two feature sets (views) $x_1$ and $x_2$ describing the examples by randomly splitting feature set $x$ in half.
- Run co-training algorithm using $L$, $U$, $x_1$ and $x_2$. Algorithm outputs $Lres_i$ (enlarged labeled set $L$ consisting from initial $L$ and examples from $U$ labeled by co-training algorithm)
- Update statistic $S$ - for each instance $e$ which is a member of $Lres_i$:
    - Increase the occurrence number of instance $e$, $n_e$.
    - for each class $C_k$, we count the number of times instance $e$ is labeled $C_k$ (equation 2)

for each instance $e$ occurring in statistic $S$:

- Calculate the most voted label of the instance, $C$: $n_{eC} = \max(n_{eCk} \mid k \in \{1..\, m\})$ and label instance $e$ as $C$.
- Calculate the label agreement percent: $e_{agg} = n_{eC}/n_e$
- Calculate the example agreement percent: $e_{occ} = n_e/n$ ($n$ -number of examples in $S$)

for each instance $e$ occurring in statistic $S$:

- If $e_{agg} > L_{ts}$ and $e_{occ} > E_{ts}$ add $e$ (labeled as $C$) to initial training set $L$

Train learner on the enlarged training set $L$ to get the final classifier. Apply final classifier to $T$.

Figure 2
Pseudo-code of RSSalg

An example case of forming the training set for the final classifier in RSSalg is shown in Figure 3. In this example, the label threshold is set to 80%, which means that the example is considered reliable if it is in at least 80% cases assigned the same label. The example threshold is set to 70%, which means that the example is considered informative if it is contained in at least 70% of the co-training result sets. Figure 3 shows that all examples from the initial training set $L$ are contained in the final training set, as these examples appear in all co-training result sets and are assigned the same label in all these sets (co-training algorithm only adds examples to the initial training set without modifying the initial data). Example $U_1$ is added to the final training set as it is contained in all the co-training result sets (it is contained in 100% co-training result sets, which is more than the value of 70% defined for the example threshold) and in 80% of the cases it was assigned the positive label (which is the same as the value of 80% defined for the label threshold), and thus in the final training set $U_1$ is labeled positive. Example $U_2$ is also added to the final training set and labeled negative as it is contained in 80% co-training result sets and it is 87.5% of the times labeled negative (in 7 out of 8 cases in which it is contained in the co-training result sets). Example $U_3$ passes the example threshold (it is contained in 80% co-training result sets), but is not considered reliable as it was 50% labeled positive and 50% labeled negative, and thus it is not included in the training set of the final classifier. Example $U_4$ is not included in the final training set as it does not pass the example threshold (it is contained in only 50% co-training result set), although it has a high label agreement (100% - in the five cases that it was labeled, it was labeled negative). Finally, the last unlabeled example $U_m$ is excluded from the final training set as it fails to pass both thresholds.

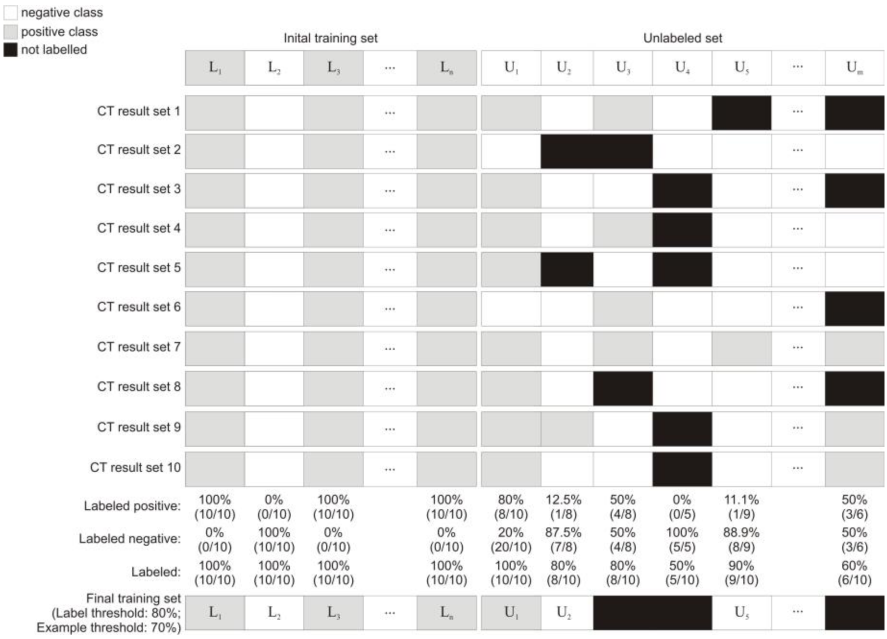



Figure 3

An example case of forming the training set for the final classifier

In order to find an optimal label threshold/example threshold pair ($l_{ts}$ and $e_{ts}$) which maximizes the performance of the model output by RSSalg, a genetic algorithm is used.

## 3.1 Automatic Determination of Label Threshold/Example Threshold Pair

Defining label/example threshold pair is a complex optimization problem. Genetic algorithms [22] are a class of adaptive search and optimization techniques, extremely efficient at searching large solution spaces. They have been shown to be a robust and effective method for the optimization of complex problems characterized by multiple optima, nondifferentiability and other irregular features [23], and thus they are suitable for our problem of optimizing label/example threshold pair.

Genetic algorithm is a search heuristic applied in order to find approximate solutions to complex optimization and search problems. It mimics the behavior of natural selection in order to reach an optimum solution to any multi-dimensional problem [22].

In our methodology each individual has two chromosomes: one represents the label threshold and the other represents the example threshold. Both chromosomes have a binary string structure. The values of threshold range from 0% to 100%. Simply converting the threshold values to their binary values and handling them as binary strings would allow recombination operators to possibly change them in such a way that they fall out of specified range. For example, if we use 8-bit binary strings to code our values, a value of 1% encoded as 00000001 could fall out of range by modification of the first bit: 10000001 would encode 129%. Thus, if we use the n-bit encoding, we divide the number of all possible values ($2^n$-1) in 100 intervals and assign each interval one value from 0% to 100%. Thus, when converting the number *x*, which falls into range of (*minVal*, *maxVal*), to its n-bit binary encoding, we use equation 4. For example, the example threshold of 55% is encoded in 8-bit binary string as 10001100.

$$Binary[Round(\ (x - minVal) \cdot (2^n - 1) / (maxVal - minVal)\ )] \quad (4)$$

Each individual represents the label/example threshold pair. As we want to produce the model with the highest possible accuracy, the logical fitness function for the individual would be the accuracy of this classification model achieved on the set intended for model evaluation. However, in the co-training setting we are limited to only a few labeled examples and we lack the labeled test data necessary to evaluate the accuracy of our model.

Setting the label threshold/example threshold pair causes some of the data from co-training result sets that exceeds these thresholds to be selected as training data for the final model creation, and some of the data to be omitted from this selection. In the example shown in Figure 3, examples $U_1$, $U_2$ and $U_5$ are included in the final training set, while $U_3$, $U_4$ and $U_m$ are shown to be omitted from the final training set. Based on the idea of out-of-bag estimation [24] to use the left-out examples from bootstrap samples to form accurate estimates of important quantities, we form the test set for model evaluation from the omitted data that did not fulfill the requirements to be included in training of the model, that is, from the data whose label agreement percent and example occurrence percent did not exceed label threshold and example threshold, respectively. Instances in thus formed test set are labeled by a majority vote of resulting co-training classifiers, in the same way as examples that did pass the thresholds. In the example shown in Figure 3, we evaluate the model trained on the examples $\{L_1,\ldots,L_n, U_1,U_2,U_5,\ldots\}$ on the test set $\{U_3,U_4,\ldots,U_m\}$.

However, some label threshold/example threshold pairs which might occur can cause all of the examples to be selected in the final training set, thus leaving no examples left for evaluation of the individual. Also, too small test sets $\{U_3,U_4,\ldots,U_m\}$, with high possibility of noise, could impose a bad estimation of the performance of the created model. Thus, we define a testing threshold - a minimal number of examples needed in the test set for the evaluation of an individual. This testing threshold is dependent on the size of the total number of examples in statistic *S*. For example, if we use a testing threshold of 20%, each individual that uses more than 80% examples from the statistic *S* for the model, and less than 20% of the examples form *S* for testing the model, it is considered as poorly estimated. In these cases, we transfer examples from the training set $\{L_1,\ldots,L_n, U_1,U_2,U_5,\ldots\}$ to the test set $\{U_3,U_4,\ldots,U_m\}$ until we have enough examples in the test set for estimation. The examples that are transferred from the training set are those estimated to be the least confident ones. We estimate which examples are the least confident ones based on the label agreement percent and example agreement percent: those examples that have the smallest sum of these two values are considered to be the least confident ones.

In the selection stage individuals of the current generation are chosen from the population and allowed to reproduce. We used the proportional (roulette wheel) selection [25]. For recombination, we used bi-parental uniform crossover [26] and single point mutation operator [22] in order to produce new individuals. In our setting elitism is used to preserve the best individual (as determined from fitness evaluations) in each generation in order to increase the speed of the search. Elitism reserves two slots in the next generation for the highest scoring chromosome of the current generation.

## 3.2 The Motivation behind RSSalg Methodology

The motivation for our methodology is that unlabeled data can be helpful to ensemble learning [18]. Our methodology exploits ensemble learning in the terms of combining the result sets $Lres_i$, created in the co-training process by different co-training classifiers. Ensemble learning is only effective if the classifiers in the ensemble are both diverse and accurate [27], which can be achieved by applying an unstable base learning algorithm. Co-training possesses this feature of instability as it is sensitive to the two underlying assumptions on the views [5, 28], and therefore to the feature split division. Thus, by running co-training using different random splits of features, we should gain an ensemble of diverse classifiers. The empirical studies show that the co-training algorithm used with random split may be beneficial, provided that there is enough redundancy in the data [5]. A major characteristic of natural language datasets is the high level of feature redundancy [29]. Thus, by using co-training as the base learner, with different random splits of features on natural language datasets, we hope to gain an ensemble of both accurate and diverse classifiers. However, on the less redundant datasets, co-training using a random split is unreliable. In the experiment performed in this paper, we test RSSalg on both groups of datasets.

Our RSSalg methodology has certain similarities with bagging [30]. The first phase in both bagging and our RSSalg methodology is creating the bootstrap replicas of the original training set. In order to perform bagging, we need a large training set. The set of bootstrap replicates is created by randomly drawing examples from that large training set. In bagging, a necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is their accuracy and diversity [27]. Bagging relies on the available training data in order to create the diversity. In the setting where we dispose with only a small amount of training examples, the diversity among the ensemble classifiers would be limited. Also, ensemble classifiers would be trained on bootstrap samples, which omit some of the training data, and thus their performance would be even worse than the performance of a weak classifier trained on the small labeled training set. Ensemble methods do not use unlabeled data as an additional source of knowledge but are rather designed when there is sufficient source of labeled data but only week learning algorithms.

In RSSalg, we dispose of the few training examples and a sufficiently large set of unlabeled examples. The set of bootstrap replicas is created from both labeled examples and unlabeled examples by applying co-training in order to incorporate unlabeled data into the original training set. In such a way we hope to increase the amount of training data and gain a more accurate classifier then the one trained on the original labeled set. The difference is that labels appointed by co-training to examples in bootstrap data are less reliable compared to the labels of examples in bootstrap data created by bagging.

The second step in bagging is to apply a base learning algorithm on each bootstrap replicate in order to gain an ensemble of learning models. Then, we can decide on a label of a previously unseen example by a majority vote of the resulting learning models. In RSSalg, we use majority voting in order to determine the final labels of the examples from bootstrap replicas. Even if the requirements for the successful application of co-training are met, in the co training process some noise is added to the initial training set through non-perfect classification. Moreover, the hard requirements of co-training are rarely met in practical situations, thus the performance of co-training is degraded by using non-ideal split. This introduces noise in bootstrap replicas created in the first phase of algorithm. Compared to other ensemble methods, bagging tends to be less accurate but stable and robust to classification noise [27]. Thus, by employing majority voting as in bagging, we hope to gain more accurate labels of examples in bootstrap replicas.

# 4 Empirical Evaluation

In this section we describe the experimental procedures and report the obtained results.

## 4.1 Datasets and their Preprocessing

We have tested our settings on tree natural language datasets: WebKB-Course [1, 5], News2x2 [5, 7] and LingSpam corpus [6, 8]. In addition, we have performed experiments on 14 UCI datasets. All of these datasets were previously used for co-training evaluation [10, 13]. We have selected these datasets that vary in size, dimensionality, redundancy and other characteristics in order to gain better insight on the effectiveness of our method in the real-world situations.

In case of natural language datasets we adopt the preprocessing technique used in [7] in order to compare our algorithm to the performance of co-training run using their artificial maxInd feature split. For dimensionality reduction, an English stop-word filter that removes 319 frequent words is used, and stemming is performed with Porter's stemming algorithm [31][1]. Based on document frequency, 200 most important features are chosen for each view. After dimensionality reduction, datasets are represented using the bag of words model with *tfidf* measurement [32]. It should be noted that both document frequency and *tfidf* measurement do not require the knowledge of the class label and therefore they do not violate the co-training setting in which labels are known for just a few initial training examples.

## 4.2 Evaluation Methodology

For evaluation, we adopt a 10-fold-cross validation procedure used in [7] in order to compare these algorithms more accurately. In this setting, data is divided in 10 stratified folds and in each of the 10 rounds of the validation process, a different fold is selected for random selection of required number of labeled training examples (the required number of training examples for each dataset is listed in Table 1, in the column denoted by |L|). The remaining data from that fold as well as 5 adjacent folds are used as unlabeled training data. The remaining 4 folds are used as test data. Thus, in each round, 40% of the data is used for testing, and 60% of the data is used for training. Each fold is used exactly once for the selection of labeled data, five times it is included as unlabeled data and four times it is used as a part of the test set.

In our experiments accuracy is used as the measure of performance. Also, a statistical significance test, a pair-wise t-test with p-value of 0.05, is applied to see if the differences in accuracy of the tested algorithms are statistically significant.

Co-training itself has a list of parameters that need to be configured. The number of examples for each class in the initial training set is chosen proportional to the

[1] In case of WebKB dataset, each sample was also filtered in order to remove the phone numbers, digits sequences, dates and non–alphanumeric characters which give no significance in predicting the class of the document.

class distribution in the dataset (for the only exception is the LingSpam dataset where we use the same settings as those in [7]). Parameters *n* and *p* that represent the numbers of examples per each class labeled by co-training inner classifiers at each iteration are also chosen proportionally to the class distribution in the dataset. By following [7], the size of the unlabeled pool *u* was 50 and the number of iterations used in each of the setting of co-training algorithm was 20. The number of different random splits used in RSSalg *m* was 50 for the natural language datasets and 30 for the UCI datasets. As the base classifier in co-training algorithm, we have used Naïve Bayes as it displays both speed and accuracy when applied on benchmark datasets.

We have used the following settings for the genetic algorithm: the generation size was 50, the number of iterations for genetic algorithm[2] was 10, the fixed probability of swapping bits in uniform crossover was 0.3, the probability of bit mutation was 0.02, and elitism was used. All these parameters were empirically chosen. The value of testing threshold was set to 20%. The datasets and their basic properties are listed in Table 1. The first four datasets in Table 1 (WebKB, LingSpam, News2x2 and Spambase from UCI) are natural language datasets, and the remaining 13 are UCI datasets. Both groups are sorted by parameter Gap.

Table 1

A summary of the datasets used in the experiment. Notation - **Dim:** the number of features describing the dataset; **|L|:** the size of the initial training set *L*, in the format of number of positive examples/number of negative examples; **$L_{acc}$:** accuracy achieved by a supervised Naive Bayes classifier trained on the initial set *L*; **|All|:** the size of the entire training set *All* (i.e., the sum of numbers of labeled and unlabeled examples), in the format of number of positive examples/number of negative examples; **$All_{acc}$:** accuracy achieved by supervised Naive Bayes classifier trained on the entire training set *All* (i.e., labeled examples and unlabeled examples with correct label); **Gap:** performance gap (also called the optimal gain in [10]) computed as $All_{acc}$ - $L_{acc}$.

| Datasets | Dim | \|L\| | $L_{acc}$ | \|All\| | $All_{acc}$ | Gap |
|---|---|---|---|---|---|---|
| WebKB | 400 | 5/5 | 78.6 | 138/492 | 96.4 | 17.8 |
| Spambase | 57 | 2/1 | 67.7 | 1672/1087 | 79.6 | 11.9 |
| LingSpam | 400 | 5/5 | 80.1 | 288/1447 | 88.9 | 8.8 |
| News2x2 | 400 | 5/5 | 81.1 | 600/600 | 89.6 | 8.5 |
| Hepatitis | 19 | 1/1 | 61.7 | 19/73 | 84.8 | 23.1 |
| Kr-vs-kp | 36 | 6/5 | 65.6 | 1001/916 | 87.2 | 21.6 |
| Credit-g | 20 | 1/1 | 53.6 | 420/180 | 74.1 | 20.5 |
| Heart-statlog | 13 | 3/2 | 65.7 | 90/72 | 80.5 | 14.8 |
| Cylinder-bands | 39 | 2/3 | 58.4 | 136/187 | 72.9 | 14.5 |
| Sonar | 60 | 1/1 | 55.5 | 66/58 | 68.8 | 13.3 |
| Ionosphere | 34 | 5/3 | 70.1 | 135/75 | 83.1 | 13.0 |
| Breast-cancer | 9 | 1/1 | 59.0 | 51/120 | 71.7 | 12.7 |
| Credit-a | 15 | 4/5 | 69.3 | 184/229 | 81.5 | 12.2 |

[2] Empirical experiments show that GA fastly converges after just a few iterations.

| | | | | | | |
|---|---|---|---|---|---|---|
| Tic-tac-toe | 9 | 3/6 | 58.8 | 199/375 | 70.7 | 11.9 |
| Breast-w | 9 | 1/2 | 86.3 | 144/274 | 97.4 | 11.1 |
| Mushroom | 22 | 3/3 | 84.7 | 2524/2349 | 95.3 | 10.6 |
| Diabetes | 8 | 2/1 | 64.8 | 300/160 | 75.0 | 10.2 |

In our experiments the following co-training methods were considered as alternatives for solving the problem of co-training application on single-view datasets:

1) Co-training applied with a natural feature split (in the case where this split was known, i.e. WebKB, News2x2 and LingSpam dataset), hereinafter referred to as **Natural**.

2) Co-training applied with a random split of features, hereinafter referred to as **Random**. In order to give more realistic results, co-training is used with several random splits (the same *m* splits used in RSSalg) and the results are averaged.

3) To experiment with an alternative method of combining co-training classifiers gained in the first step of RSSalg, we use majority voting method, hereinafter referred to as **MV**.

4) Co-training applied with an artificial maxInd feature split introduced in [7]. MaxInd is not always successful in combination with Naïve Bayes and it displays better performance when applied with other base classifiers, such as RBF Nets and SVM [7]. Because of this, we report the best performance of maxInd achieved when using one of these three classifiers for each dataset, hereinafter referred to as $\mathbf{maxInd_{best}}$.

5) The RSSalg introduced in this paper, optimized with the genetic algorithm procedure presented in section 3.1., hereinafter referred to as **RSSalg**.

6) RSSalg with label threshold/example threshold pair optimized with genetic algorithm that uses the accuracy of the resulting classification model, achieved on the set intended for model evaluation, as the fitness function, hereinafter referred to as $\mathbf{RSSalg_{best}}$.

## 4.3 Experimental Results

The empirical studies [5] show that co-training algorithm using a random feature split may be beneficial if there is enough redundancy in the data. A major characteristic of natural language datasets is the high level of feature redundancy [29]. On the other hand, most of the UCI datasets are manually constructed with carefully chosen features and there should not be much redundancy in such created features. Thus, we will present and discuss the results achieved on natural language datasets and on the less redundant UCI datasets separately.

In Table 2 we show the accuracy and t-tests with 95% confidence obtained by each co-training method on all datasets.

Table 2
Comparisons of RSSalg and $RSSalg_{best}$ vs. four alternative co-training methods. The percent accuracy and standard deviation are reported based on 10-fold stratified cross-validation, and t-test with 95% confidence obtained by each co-training methods on all datasets. The Natural is applicable only for the first three datasets (WebKB, LingSpam, and News2x2) that have the known natural feature split. The results denoted by "v" indicate cases where RSSalg was significantly more accurate than the alternative co-training method, while a "*" indicates the significant difference in favor of the alternative method and no symbol represents a tie (no statistically significant difference). The overall results of the t-test are summarized in the rows "t-test" (separately for natural language datasets and UCI datasets) as the number of wins/ties/loses of RSSalg versus alternative co-training methods (e.g. 3/1/0 in the t-test row and the Random column means that RSSalg was statistically significantly more accurate than Random on 3 datasets, RSSalg and Random resulted in no statistical difference on 1 datasets, and zero (0) means no dataset indicated RSSalg was statistically significantly less accurate than Random in the experiment).

| Datasets | Natural | Random | MV | $MaxInd_{best}$ | RSSalg | $RSSalg_{best}$ |
|---|---|---|---|---|---|---|
| WebKB | 87.2±6.7 | 84.2±7.6 | 87.7±3.5 | 78.3±9.1 v | 87.3±5.1 | 90.7±3.3 |
| Spambase | - | 67.8±15.8 v | 77.4±4.7 | 68.9±8.2 v | 78.2±7.7 | 81.5±4.1 |
| LingSpam | 70.3±13.3 v | 76.6±8.5 v | 81.1±7.4 v | 83.9±1.1 | 88.5±7.0 | 91.1±5.9 |
| News2x2 | 82.9±4.6 v | 80.0±7.0 v | 86.3±2.4 v | 76.2±12.8 v | 89.1±3.1 | 90.6±1.8 |
| t-test | 2/1/0 | 3/1/0 | 2/2/0 | 3/1/0 | - | 0/4/0 |
| Hepatitis | - | 80.3±8.0 | 83.3±4.3 | 80.8±7.9 | 82.6±3.5 | 86.5±3.4 * |
| Kr-vs-kp | - | 54.4±5.0 v | 55.3±4.5 | 60.1±6.2 | 58.3±5.7 | 67.1±4.2 * |
| Credit-g | - | 62.0±5.5 | 64.4±5.5 | 68.1±1.8 * | 62.7±6.8 | 70.2±0.7 * |
| Heart -statlog | - | 79.4±8.2 | 81.8±2.0 | 80.8±4.5 | 81.1±3.2 | 83.3±2.2 |
| Cylinder -bands | - | 52.5±5.2 | 52.9±6.5 | 56.3±5.7 | 54.3±4.8 | 61.6±2.5 * |
| Sonar | - | 54.9±6.0 | 56.5±5.5 | 56.7±9.1 | 56.7±8.0 | 61.2±5.8 |
| Ionosphere | - | 69.4±12.6 | 73.1±4.9 | 78.3±7.7 | 74.6±7.3 | 79.6±5.8 |
| Breast -cancer | - | 66.7±6.1 | 68.2±4.5 | 67.5±5.4 | 67.0±6.8 | 70.4±5.3 |
| Credit-a | - | 69.2±15.0 | 73.4±11.0 | 76.1±2.6 | 72.0±8.4 | 77.6±4.8 |
| Tic-tac-toe | - | 61.5±3.2 | 63.2±2.5 | 62.0±1.7 | 61.6±3.0 | 64.1±2.9 |
| Breast-w | - | 96.8±0.8 | 96.9±0.7 | 96.7±0.7 | 96.5±1.0 | 97.5±0.4 * |
| Mushroom | - | 88.2±3.2 | 89.1±1.0 | 88.4±1.3 | 88.6±1.4 | 89.2±0.9 |
| Diabetes | - | 61.4±7.3 | 64.1±3.3 | 65.3±1.1 | 63.9±3.7 | 67.7±1.8 * |
| t-test | - | 1/12/0 | 0/13/0 | 0/12/1 | - | 0/7/6 |

The aggregated t-test results obtained by each method against each other in terms of wins-ties-losses on natural language datasets and less redundant UCI datasets are reported at Tables 3 and 4, respectively. Methods include not only the 6 co-training methods, but also 2 additional Naive Bayes classifiers trained on the small labeled set *L* and on a much larger training set *All* respectively.

Table 3
Aggregate number of wins/ties/loses of each method against other methods over four natural language datasets. Comparisons versus NB_L (i.e., Naive Bayes classifier trained on the initial set *L*) and NB_All (i.e., Naive Bayes classifier trained on a much larger training set *All*) are reported in the last two rows/columns. For example, 1/2/0 in the row MV and column Natural means that MV was statistically significantly more accurate than Natural on 1 dataset, MV and Natural resulted in no statistical difference on 2 datasets, and MV was statistically significantly less accurate than Natural on 0 datasets.

| | Natural | Random | MV | $maxInd_{best}$ | RSSalg | $RSSalg_{best}$ | NB_L | NB_All |
|---|---|---|---|---|---|---|---|---|
| Natural | - | 0/2/1 | 0/2/1 | 1/1/1 | 0/1/2 | 0/1/2 | 1/2/0 | 0/0/3 |
| Random | 1/2/0 | - | 0/3/1 | 1/2/1 | 0/1/3 | 0/0/4 | 1/3/0 | 0/0/4 |
| MV | 1/2/0 | 1/3/0 | - | 3/1/0 | 0/2/2 | 0/2/2 | 3/1/0 | 0/1/3 |
| $maxInd_{best}$ | 1/1/1 | 1/2/1 | 0/1/3 | - | 0/1/3 | 0/0/4 | 0/4/0 | 0/0/4 |
| RSSalg | 2/1/0 | 3/1/0 | 2/2/0 | 3/1/0 | - | 0/4/0 | 4/0/0 | 0/3/1 |
| $RSSalg_{best}$ | 2/1/0 | 4/0/0 | 2/2/0 | 4/0/0 | 0/4/0 | - | 4/0/0 | 0/3/1 |
| NB_L | 0/2/1 | 0/3/1 | 0/1/3 | 0/4/0 | 0/0/4 | 0/0/4 | - | 0/0/4 |
| NB_All | 3/0/0 | 4/0/0 | 3/1/0 | 4/0/0 | 1/3/0 | 1/3/0 | 4/0/0 | - |

Table 4
Aggregate number of wins/ties/loses of each method against other methods over 13 UCI datasets. Comparisons versus NB_L (i.e., Naive Bayes classifier trained on the initial set *L*) and NB_All (i.e., Naive Bayes classifier trained on a much larger training set *All*) are reported in the last two rows/columns. For example, 1/12/0 in the row RSSalg and column Random means that RSSalg was statistically significantly more accurate than Random on 1 dataset, RSSalg and Random resulted in no statistical difference on 12 datasets, and RSSalg was statistically significantly less accurate than Random on 0 datasets.

| | Random | MV | $maxInd_{best}$ | RSSalg | $RSSalg_{best}$ | NB_L | NB_All |
|---|---|---|---|---|---|---|---|
| Random | - | 0/13/0 | 0/9/4 | 0/12/1 | 0/4/9 | 7/4/2 | 0/2/11 |
| MV | 0/13/0 | - | 0/13/0 | 0/12/0 | 0/7/6 | 7/4/2 | 0/2/11 |
| $maxInd_{best}$ | 4/9/0 | 0/13/0 | - | 1/11/0 | 0/8/5 | 6/6/1 | 0/2/8 |
| RSSalg | 1/12/0 | 0/13/0 | 0/12/1 | - | 0/7/6 | 4/7/1 | 0/3/10 |
| $RSSalg_{best}$ | 9/4/0 | 6/7/0 | 5/8/0 | 6/7/0 | - | 12/1/0 | 0/5/8 |
| NB_L | 2/4/7 | 2/4/7 | 1/6/6 | 1/7/4 | 0/1/12 | - | 0/0/13 |
| NB_All | 11/2/0 | 11/2/0 | 10/3/0 | 10/3/0 | 8/5/0 | 13/0/0 | - |

In summary, in our conducted experiments the proposed method RSSalg had the following properties as compared to alternatives on natural language datasets:

1) On two datasets the proposed method outperformed Natural and they resulted in a statistical tie on one remaining dataset (WebKB). These results indicate that RSSalg is generally better then Natural.

2) On three datasets RSSalg outperformed Random and they resulted in a statistical tie on one remaining dataset (WebKB). These results indicate that RSSalg is generally better then Random.

3) On two datasets RSSalg outperformed MV (LingSpam and News2x2) and they resulted in a statistical tie on two remaining datasets. These results indicate that, RSSalg is generally better then MV.

4) On three datasets RSSalg outperformed $MaxInd_{best}$ and they resulted in a statistical tie on one remaining dataset (LingSpam). Therefore, it is generally better than maxInd because RSSalg outperformed maxInd's best performance on the majority of datasets.

5) On all four datasets RSSalg and $RSSalg_{best}$ result in a statistical tie. As $RSSalg_{best}$ represents the upper bound of performance of RSSalg on the test set, we can conclude that our method for automatic threshold determination was successful.

6) On all four datasets RSSalg outperformed the Naive Bayes classifier trained on the small labeled set *L*. In contrast, none of the alternative co-training methods outperformed the Naive Bayes to that level.

7) On three datasets RSSalg resulted with a statistical tie with Naive Bayes trained using a much larger set of labeled data *All*, and it lost to NB_All on one dataset (WebKB). None of the alternative methods were able to achieve this on the majority of datasets (only MV achieved a statistical tie with Naive Bayes classifier trained on a large labeled set on one dataset, Spambase).

In our experiments co-training using a natural feature split outperformed Naive Bayes classifier trained on the small labeled set *L* on only one dataset (WebKB).

In our experiments on the group of natural language datasets, maxInd was not beneficial. Its best setting ($maxInd_{best}$) has resulted with a statistical tie with Naive Bayes classifier trained on the small labeled set *L*.

On natural language datasets, in our experiments MV outperformed maxInd on three datasets and resulted with a statistical tie with maxInd on the remaining dataset (LingSpam). MV resulted with a statistical tie with Natural and Random on most datasets. However, as opposed to Natural, Random and maxInd it outperformed Naive Bayes classifier trained on the small labeled set *L* on the majority of datasets (it is only tied with NB_L on LingSpam dataset).

Generally, in our experiments on natural language datasets, RSSalg performed the best, followed by MV, and finally Natural, Random and maxInd which turned out to be of similar performance.

In the summery of our experiments on UCI datasets, the proposed method RSSalg had the following properties as compared to alternatives:

1) On one dataset (Kr-vs-kp) RSSalg outperformed Random and they resulted in a statistical tie on the remaining 12 datasets. These results indicate that RSSalg performs similarly to Random.

2) On all 13 datasets RSSalg and MV resulted in a statistical tie. Thus, RSSalg and MV have a similar performance.

3) RSSalg and $MaxInd_{best}$ resulted in a statistical tie on 12 datasets and RSSalg loses to $MaxInd_{best}$ on one remaining dataset (Credit-g). Thus, RSSalg performs similar to $MaxInd_{best}$.

4) On 7 datasets RSSalg results with a statistical tie with $RSSalg_{best}$ and RSSalg loses to $RSSalg_{best}$ on 6 datasets. As $RSSalg_{best}$ represents the upper bound of performance of RSSalg on the test set, we can conclude that our automatic threshold determination was successful on the majority of datasets. However, it is less effective compared to its performance on natural language datasets.

5) On 4 datasets RSSalg outperformed the Naive Bayes classifier trained on the small labeled set *L*, resulted in a statistical tie with it on 7 datasets, and lost to Naive Bayes classifier trained on small labeled set *L* on two datasets. Thus, given less redundant datasets, it is unreliable weather RSSalg will be beneficial.

6) On three datasets RSSalg resulted with a statistical tie with Naive Bayes trained using a much larger set of labeled data *All*, and it lost to NB_All on 10 datasets.

Generally, in our experiments on UCI datasets, Random, maxInd, and MV show similar performance. RSSalg showed a slightly worse performance in that it outperformed the Naive Bayes classifier trained on the small labeled set *L* on fewer datasets, compared to other benchmark algorithms.

It should be noted that $RSSalg_{best}$ has the best performance of all benchmark algorithms on UCI datasets:

1) $RSSalg_{best}$ outperformed Random on 9 datasets and they resulted in a statistical tie on the remaining 4 datasets.

2) $RSSalg_{best}$ outperformed MV on 6 datasets and they resulted in a statistical tie on 7 datasets.

3) $RSSalg_{best}$ outperformed $maxInd_{best}$ on 5 datasets and they resulted in a statistical tie on 8 datasets.

4) On 12 datasets $RSSalg_{best}$ outperformed the Naive Bayes classifier trained on the small labeled set *L* and resulted in a statistical tie with it on one dataset. In contrast, none of alternative co-training methods outperformed the Naive Bayes to that level, and some co-training methods even degraded the performance of the Naive Bayes classifier on certain datasets.

5) On 5 datasets RSSalg resulted with a statistical tie with Naive Bayes trained using a much larger set of labeled data *All*, and it lost to NB_All on 8 datasets.

Thus, a better method for automatic determination of thresholds for RSSalg on the less redundant datasets would greatly improve its performance.

The limitation of RSSalg is its time complexity. However, it runs completely off-line without any human interaction, which makes this issue less of a problem. Also, RSSalg can be easily executed in parallel, as multiple runs of co-training in RSSalg are completely independent.

Random should have the lowest time complexity of all benchmark algorithms. The time complexity of the MaxInd algorithm should increase as the number of features in the dataset increases, as it requires calculating pair wise *CondMI*s for all pairs of words that are in different sets in order to determine maximally independent views [7]. Finally, MajorityVote is less complex then RSSalg as it does not require genetic algorithm optimization or training of the final classifier as in RSSalg. However, MV requires an additional step that is not performed in RSSalg, namely applying each of the obtained co-training classifiers on the test set. As RSSalg, MV requires multiple independent co-training runs, which can easily be executed in parallel.

## Conclusion

In this paper we have proposed RSSalg, a methodology designed enable successful application of co-training on single-view datasets and boost its performance. Our method is most successful on datasets with enough feature redundancy, such as natural language datasets. RSSalg relies on co-training applied with different random feature splits in order to form the accurate enlarged training set.

We have compared the accuracy of the proposed method to several alternative co-training methods applicable to single-view datasets. In addition, the new method is compared to co-training with a natural feature split in cases where such a split was known. Our algorithm outperformed all considered alternative methods for co-training on the more redundant natural language datasets, while it was comparable to the considered alternative settings on the less redundant UCI datasets.

Our genetic algorithm technique of determination of most accurate and informative examples would perform its best if we could define the fitness function as actual accuracy achieved on the test set. However, in the co-training setting we are limited to only a few labeled examples and we lack the labeled test data necessary for this fitness function. We have used this model $RSSalg_{best}$ as indication of the upper bound on the performance of our RSSalg. Our results show that the performance of RSSalg reaches the performance of $RSSalg_{best}$ as its upper bound accuracy on natural language datasets, while it is less successful on the less redundant UCI datasets. $RSSalg_{best}$ has outperformed all considered co-training methods on UCI datasets; thus, a better method for automatic determination of thresholds for RSSalg on the less redundant datasets would greatly improve its performance, and this remains a task for the future.

Our empirical experiments with $RSSalg_{best}$ suggest that, given optimal parameters, RSSalg is consistent on various classification domains and does not require any specific requirements about the dataset to be met. However, the determination of optimal parameter values, without using labeled examples, remains to be improved in the future.

The downfall of genetic search is that it suffers from serious variance (i.e., very different results may be obtained by different runs). In the future we plan to undertake an experimental study in order to analyze the difference caused by different runs of genetic search.

Another limitation is the introduction of an additional parameter in the co-training setting, represented as the number of co-training classifiers created in the first step of RSSalg. In the future we plan to experiment with different parameter settings in order to investigate their impact on RSSalg.

In the experiments presented in this paper we have considered only binary classification problems, and in future research we plan on extending our method to multiple-category label problems. Also, we plan on undertaking more experimental studies in order to compare our method to ensemble approaches that exploit unlabeled data, such as co-forest [14]. Finally, we hope to integrate our solution into the information system for monitoring the scientific research activity of the University of Novi Sad (CRIS UNS)[3]. Our solution would serve as an additional support of system for automatic extraction of metadata from scientific publications [33]. The goal is to overcome the problem of manual annotation of a large number of scientific papers. Finally, we hope to apply our algorithm as the solution for automatic mining of methodologies from scientific articles [34].

**Acknowledgement**

Results presented in this paper are a part of the research conducted within the Grant No. III-47003 provided by the Ministry of Education and Science of the Republic of Serbia. The authors are thankful to Dr. Felix Feger and Dr. Irena Koprinska for sharing the details of their work.

**References**

[1] A. Blum and T. Mitchell: "Combining Labeled and Unlabeled Data with Co-Training", *Proc. 11th Annual Conf. Computational Learning Theory*, pp. 92-100, 1998

[2] J. Slivka, A. Kovačević and Z. Konjović: "Co-Training-based Algorithm for Datasets without the Natural Feature Split", *Proc. IEEE Int'l Symp. Intelligent Systems and Informatics*, pp. 279-284, Serbia, 2010

[3] J. Slivka, A. Kovačević and Z. Konjović: "Multi-Label Classification Experiments with Co-Training-based Algorithm" *Int'l Conf. Internet Society Technology and Management*, Serbia, 2011

[4] J. Slivka, A. Kovačević and Z. Konjović: "Co-Training-based Algorithms Applied to Subjectivity Detection Task" *Int'l Conf. Internet Society Technology and Management*, Serbia, 2012

[5] K. Nigam and R. Ghani: "Understanding the Behavior of Co-Training", *Proc. Workshop on Text Mining at KDD*, 2000

[6] J. Chan, I. Koprinska and J. Poon: "Co-Training with a Single Natural Feature Set Applied to Email Classification", *Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence*, pp. 586-589, 2004

[3] http://cris.uns.ac.rs/

[7] F. Feger and I. Koprinska: "Co-Training Using RBF Nets and Different Feature Splits", *Proc. Int'l Joint Conf. Neural Network*, pp. 1878-1885, 2006

[8] M. Terabe and K. Hashimoto: "Evaluation Criteria of Feature Splits for Co-Training", *Proc. Int'l MultiConference of Engineers and Computer Scientists,* 2008

[9] A. Salaheldin and N. El Gayar: "New Feature Splitting Criteria for Co-Training using Genetic Algorithm Optimization" Lecture Notes in Engineering and Computer Science, 5997/2010, pp. 22-32, 2010

[10] J. Du, C. Ling and Z. Zhou: "When Does Co-Training Work in Real Data?", *IEEE Trans. on Knowledge and Data Engineering*, 23(5), pp. 788-799, 2010. ISSN 1041-4347, 2010

[11] D. Pierce and C. Cardie: "Limitations of Co-Training for Natural Language Learning from Large Datasets", *Proc. 2001 Conf. on Empirical Methods in Natural Language Processing,* 2001

[12] Z.-H. Zhou: "Ensemble Methods: Foundations and Algorithms" In: Boca Raton, FL: Chapman & Hall, 2012

[13] Z.-H. Zhou and Ming Li: "Tri-training: Exploiting unlabeled data using three classifiers", *IEEE Trans. Knowledge and Data Engineering* 17, pp. 1529-1541, 2005

[14] M. Li and Z. H. Zhou: "Improve Computer-aided Diagnosis with Machine Learning Techniques using Undiagnosed Samples" *IEEE Trans. on Systems, Man and Cybernetics*, 37(6), pp. 1088-1098, 2007

[15] M. F. Abdel Hady and F. Schwenker: "Co-Training by Committee: A New Semi-supervised Learning Framework" *Proc. IEEE Inte'l Conf. Data Mining Workshops*, pp. 563-572, 2008

[16] M. West: "Bayesian Factor Regression Models in the "Large p, Small n" Paradigm", In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 7*, pp. 723-732, Oxford University Press, 2003

[17] Y. Yaslan and Z. Cataltepe: "Co-Training with Relevant Random Subspaces", Neurocomputing, Vol. 73(10-12), pp. 1652-1661, 2010

[18] Z-H. Zhou: "When Semi-supervised Learning Meets Ensemble Learning", *Proc. Int'l Workshop on Multiple Classifier Systems*, pp. 529-538, 2009

[19] M-L. Zhang and Z-H. Zhou: "Exploiting Unlabeled Data to Enhance Ensemble Diversity", Data Mining and Knowledge Discovery, pp. 1-32, 2011

[20] L. Izsó and P. Tóth: "Applying Web-Mining-Methods for Analysis of Student Behaviour in VLE Courses", Acta Polytechnica Hungarica, Vol. 5, No. 4, pp. 79-92, 2008

[21] J. Slivka, P. Zhang, A. Kovačević, Z. Konjović and Z. Obradović: "Semi-Supervised Learning on Single-View Datasets by Integration of Multiple Co-Trained Classifiers", 11th International Conference on Machine Learning and Applications (ICMLA 2012), *December 12-15, Boca Raton, Florida, USA*, 2012

[22] D. Goldberg: "Genetic Algorithms in Search, Optimization and Machine Learning", Addison-Wesley, 1989

[23] R. E. Dorsey and W. J. Mayer: "Genetic Algorithms for Estimation Problems with Multiple Optima, Nondifferentiability, and Other Irregular Features," Journal of Business and Economic Statistics, 13(1), 53-66, 1995

[24] L. Breiman: Out-of-Bag Estimation, Technical Report, Statistics Department, University of California, 1996

[25] T. Back and F. Hoffmeister: "Extended Selection Mechanisms in Genetic Algorithm", In: Belew, R. K. & Brooker, L. B., editors, *Proc. 4th Int'l Conf. on Genetic Algorithm*, pp. 92-99, 1991

[26] G. Syswerda: "Uniform Crossover in Genetic Algorithms", *Proc. 3rd Int'l Conf. Genetic Algorithms*, pp. 2-9, 1989

[27] T. G. Dietterich: "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization", Machine Learning, 40(2), pp. 139-157, 2000

[28] I. Muslea, S. Minton and C. Knoblock: "Active + Semi-supervised Learning = Robust Multi-View Learning", *Proc. ICML*, 2002

[29] T. Joachims: "A Statistical Learning Model of Text Classification for Support Vector Machines" *Proc. ACM-SIGIR Int'l Conf. Research & Development in Information Retrieval*, pp. 128-136, 2001

[30] L. Breiman: "Bagging Predictors", Machine Learning, Vol. 24(2), pp. 123-140, 1996

[31] M. F. Porter: "An Algorithm for Suffix Stripping", Program: Electronic Library and Information Systems, 14(3):130-137, 1980

[32] Gerard Salton and Chris Buckley: "Term-Weighting Approaches in Automatic Text Retrieval", Information Processing and Management, Vol. 24(5) pp. 513-523, 1988

[33] Kovačević, A., Ivanović D., Milosavljević B., Konjović Z., Surla D., 2011. "Automatic Extraction of Metadata from Scientific Publications for CRIS Systems" Program: Electronic Library and Information Systems, 45(4), pp. 376-396, 2011

[34] Kovačević, A., Konjović Z., Milosavljević B., Nenadic G., "Mining Methodologies from NLP Publications: A Case Study in Automatic Terminology Recognition", Computer Speech & Language, 26(2), pp. 105-126, 2011

# Automatic Recognition of Features in Spectrograms Based on some Image Analysis Methods

**Aleksandar Perović[1], Zoran Đorđević[2], Mira Paskota[3], Aleksandar Takači[4], Aleksandar Jovanović[5]**

[1]University of Belgrade, Faculty of Transportation and Traffic Engineering, Vojvode Stepe 305, 11000 Belgrade, Serbia, pera@sf.bg.ac.rs

[2]University of Belgrade, Faculty of Mathematics, Group for Intelligent Systems, Studentski trg 16, 11000 Belgrade, Serbia, dzoran1@gmail.com

[3]University of Belgrade, Faculty of Transportation and Traffic Engineering, Vojvode Stepe 305, 11000 Belgrade, Serbia, m.paskota@sf.bg.ac.rs

[4]University of Novi Sad, Faculty of Technology, Bulevar cara Lazara 1, 21000 Novi Sad, Serbia, stakaci@tf.uns.ac.rs

[5]University of Belgrade, Faculty of Mathematics, Group for Intelligent Systems, Studentski trg 16, 11000 Belgrade, Serbia, aljosha@matf.bg.ac.rs

*Abstract: This paper presents progress in the investigation and development of methods for the automatic localization, extraction, analysis and comparison/classification of the features in signals and their spectra. With diverse applications, different feature attributes turn out to be significant for the investigated phenomena. The general feature characteristics are morphologic and therefore suitable for a variety of algorithms focused on visual data processing, which we use in the automatic feature recognition. Our major applications were in the analysis of biological signals, and acoustic, sonar and radar signals; the methods presented here are applicable in other areas as well.*

*Keywords: Automatic detection of spectral features; Invariants of signal features; Brain Computer Interface; Noise elimination in radar signals*

## 1 Introduction

The automatic tracking of objects represented by signals from a variety of sensors (e.g. optical, infra-red, ultra violet, sonar, radar and others) generally requires previous application of feature determination, characterization, noise reduction, background reduction and automatized extraction. Object tracking in a variety of sensory environments could be related to real time monitoring of the dynamic changes of their spectroscopic correlates. Among generally present signal features, some are more relevant than the others. Fourier spectroscopy with its developments and generalizations plays an important role in the investigation of signals of diverse nature. Spectral features certainly demonstrate distinguishable formations in time Fourier spectra, so spectrograms are of special value. Here, we can consider classic Fourier analysis and uniform Fourier spectroscopy based on other orthonormal systems. Their importance in the analysis of signals could be crucial, enabling accurate classification and prediction, thus providing essential insight into the properties of investigated phenomena and systems. One relatively unexplored approach to the extraction of the features from signals of various origins is based on image processing techniques applied to spectral analysis. Although different types of adaptive techniques, such as feed forward neural networks or genetic algorithms, can be applied to optimize parameters of image transformations used in this process, the goal of the presented paper is to demonstrate the feasibility of such an approach and to provide an overview of our results.

Semantically rather distant contexts often can be treated with the same or similar manners of mathematical modeling and implementations, with partial or full automatization as one of the key aims. We have selected some interesting examples of such diverse relationships from our practice. Image filtering and enhancement techniques applied to the time spectra and their further composites with well distinguished features provide automatized recognition and automatized procedures needed in various problem solutions. Common features of the spectrograms and wavelet spectrograms are aggregations that have certain topologic characteristics, like contingency, expressed boundary and differences in the intensity from its nearest neighborhood, with certain time duration. On the opposite, there are examples where distinguishable features more resemble a random dot cloud with certain density aggregation. Finally, there are short lasting features, e.g. those corresponding to the short frequency pulses in signals, which are distinguishable as dots – small sized objects in the spectrograms. All these features are often surrounded by, or embedded in, a variety of noise and artifact formations from which they need to be separated/extracted during the recognition procedures, with a degree of fuzziness present in all important properties. From the methods and algorithms we have developed and implemented, here we present ones applicable to a variety of problems in signal analysis which are related to automatic object recognition. These methods are mostly applicable in signal analysis. The software is calibrated on synthetic and alternatively measured data. Our software and signals/images together with short illustrative presentations are partially available at http://poincare.matf.bg.ac.rs/~aljosha//GIS/GIS/sbgis.htm. As illustrations we use applications in biology (e.g. signals from brain implants, EEG/MEG, neuroacoustics, blood pressure, pharmacologic applications, microscopic CCD-FISH and fNMR), acoustics and radars. This mixture of diverse examples was chosen in order to show their invariants (invariability) to problem unspecific applications; they just stress the kinematic and geometric equivalents of the specific system dynamics.

The recognition of spectral features in the monitoring of cardiac parameters is used to reveal an approach of a serious cardiac crisis and characterizes important states of the system and their transitions. Spectroscopy features of electro encephalography (EEG) and magneto encephalography (MEG) can help to predict epileptic seizures. It is also used in brain-computer interfaces (BCI), which has attracted our attention since the early nineties. We had available EEG recordings of externally generated tone stimulation and imagined tones and music [1, 2]. The spectrograms of such recordings contain traces of imagined tones, which can be taken as the basis for the BCI command language. Thanks to the impressive achievements of the Rome group [3], the Graz group [4], the Tübingen group [5] and other research teams as well, BCI has become a reality recently, resulting in an explosion of interest in this area. Some researchers have suggested the use of high-frequency EEG for BCI applications [6, 7, 8], which could expand the brain command language capacity proportionally to the increase of the speed of changes of controllable brain electro-physiological states and their number [2, 9].

# 2 Method

In automatized recognition we treat features in spectrograms, their derivates and composite spectrograms, including some real time features, with the image processing tools. The sets/manifolds in the space of <*frequency, time, intensity*> could possess some (simple) topological properties that are important in the characterization of investigated time spectra and related semantics, such as electrophysiology (Event-Related Synchronization (ERS) and Desynchronization (ERD) has found wide applications in the highly frequency band-specific EEG and MEG [10]), or kinematics. On the other hand, images as organized sets in the <*x, y, intensity*> space often have photometric aspects that essentially characterize objects in images. We prefer to call this two-dimensional photometry a photo morphology. Hence, we have sets/manifolds in both metric spaces, time-frequency and spatial, where topologic/geometric invariants are in the focus of investigation. As we usually convert dynamics to the geometry or the other way around, we can identify those contexts when the same or similar types of invariants are used, or when we have common algorithms extracting these invariants and similarity classification. In this way, a part of the algorithm developed for the automatized analysis of morphological characteristics of, for example, chromosome and radar images, and their similarity and classification, is applicable to the study of spectroscopic features of signals, some needing adaptations and expansions. We shall illustrate such algorithms on both classes of examples, even mixing the two: the time-frequency and spatial. We discuss the implementations of the key functions involved, including algorithms which are basically simple but whose complexity grows with the variation of cases offered by experimental practice, thus needing increased efforts to deal with. Examples using recorded and synthetic signals have been selected to present problems, structures, methods, algorithms and solutions in a straightforward way, pointing to the aspects of interest in related problems; they all share the same corpus of analytic tools.

In order to be able to automatically detect objects in spectrograms, one first needs to separate those objects from the noise and artifacts. There are many methods for the noise removal from the signals and images, from sophisticated methods based on the Independent Component Analysis techniques to quite specialized methods we have used in the microscopic image processing, clutter in radar and sonar signals, other acoustic signal filtering, or filtering in images from cameras of different types. There are simpler and more complex situations, all requiring sophisticated noise reduction and elimination methods. In this way we have automatic noise threshold reduction, as shown in Fig. 6. However, in the case when the essential signal components are hard to separate from the noise, like when they are masked by or embedded in it, the problem of noise reduction and elimination becomes very sensitive and application dependent, with time dynamics in the noise threshold definition. Then, this step has to be subjected to a preliminary learning and intelligent treatment. The object kinetics within the observed portion of the representation space often involves object tracking as well, which, combined with the other aspects, increases problem complexity (for some examples of successful solutions of the object tracking problem, see [11, 12]). Obviously, kinetics can also be turned into a geometric form, so the tracked objects are related to their trajectories. Thus, the moving of the objects in space and the moving of the spectrogram features in time are tightly related.

We present aspects related to the feature structural morphology first, generalizing our methods for the analysis of microscopic image analysis; further on, we present adaptations of algorithms developed for the automatic object tracking and noise elimination based on the marine radar imaging.

## 2.1 Aspects Related to Feature Morphology

In Fourier spectroscopy, depending on the application, various criteria for minimum spectral stability are used, i.e. localization of sets of high/low intensity spots and lines in consecutive spectra, with topologic invariants that would distinguish them as features against e.g. smaller granular objects or random dot clouds, or all the way around. The key parameters must be determined before the application and usually involve prior knowledge of the underlying processes that are generating signals. For example, the shortest music tonal feature which is localized lasts around 0.1 s. Before the application of any further steps, all time spectrograms need to be recalibrated, unless the time spectra were initially WYSIWYG ("what you see is what you get"), when the recalibration is not needed. Similarly, constant features need no recalibration, except for the relative magnitude corrections. Generally, this is a rather complex problem. In [13] is described one partial solution which introduces some time delay, which is a method for spectrogram recalibration based on the geometry of the morphology of simple features as those present in many acoustic examples. In short, the above observations form the basis for further normalizations and for the measurement of similarity with the etalon objects and posterior feature classification with the context dependent criteria. Such procedures often need to be automatized as well.

Simplified, feature localization and structure representation can be performed as follows. If we denote a spectrogram or an image in resolution $m \times n$ with $A$, then for each row $x$, first select $y_1$ and $y_2 > y_1$such that $A(x, y_1) > threshold$ and $A(x, y_2 + 1) \leq threshold$ (where $threshold$ is a static, statistic, dynamic or learned parameter); repeating the same while $y < n$, would provide entry and exit points for each intersection of a contour with row $x$; similarly for all rows. In this way we obtain a simple localization of all contours in $A$. If artifacts are discerned by size or frequency range, they can be removed easily. After feature localization, contour stabilization-smoothing is applied if necessary, for example by removing the single convex pixels or by interpolation of the single concave pixels. Next, the topology of features is traced (Fig. 1) by two alternative processes. First, geodesy of the embedded isophotic (closed) curves determines the density gradients around feature local extremes and defines the dominant "meridian" of the feature. Alternatively, orthogonal vectors on points of the contour provide geodesic lines of the curved coordinate system, tailored to the topology of the feature. The results of the two methods are compared, providing a better contour tracing.

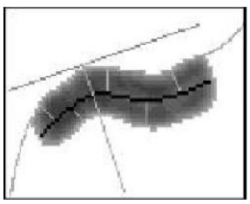


Figure 1
Feature curved coordinate system

Figure 1 shows a feature curved coordinate system tailoring and invariant capturing: positioning of the central meridian (black) and diameters perpendicular to it, determining positions of flexion points and flexion angles in the central meridian.

In the case of tonal, music or acoustic examples, the detected meridian curve possesses the information on the tonal time changes with all present dynamics (e.g. speed changes of detected vessels in submarine acoustic spectrograms, while in the case of inner music patterns or chromosomal morphologic invariants, these meridians organize the distribution of local extremes of photo density functions). Then, within tunable, reasonable approximation, the feature local extremes and contour local extremes are marked in the feature curved coordinate system. The central meridian is often curved and, together with flexion points and curvature in these points (the angle between the incoming and outgoing meridian fragments in the given point), presents an important feature invariant. The *Fishbone* demo at our web site illustrates the algorithm with normalization performed to the total feature rectification (activation of the step-by-step rectification is also available). Together with longitudinal and coordinate-wise expansion/compression factors, this can provide a good way of feature structural comparison (rectifying the corresponding features, calculating the corresponding flexure angle differences, with comparisons of corresponding segment lengths and relative positions of local extremes best fit). For the branching features, similar branch invariants could be taken into comparison considerations.

In [14] we described the similarity measures of objects in images based on geodesic (photomorphic) structural invariants (basically, the distributions of local extremes), implemented as a distance – metrics in the space of 3D manifolds. The same method can be used to characterize individual spectrogram patterns and measure their similarity using their photomorphic structural invariants. For two normalized spectrogram features $f_1$, $f_2$ with meridians $m_1$ and $m_2$, one way to define the morphology reflecting similarity of features, as a distance of their reasonable representations, can be

$$d(f_1, f_2) \;=\; \min\{\int_D |m_1(x) - c \cdot m_2(ax+b)|dx \,:\; a, b, c \in \mathbb{R}\}$$

where $a$ is a contraction, $b$ a translation parameter, $c$ an amplitude fitting parameter, or similar with some meridian simplifications. This also can be extended to the relative positions of local extremes, with the flexure angle differences as the second distance index. There are other ways to define combinatorial similarity of involved features, which are basically application dependent. Some examples of the real time spectrograms with the types of features we are dealing with including different common processing aspects generally applicable are shown below, with the purpose of the method details illustration. We have developed systems for the real time (RT) acquisition and real time analysis of a large number of signals.

## *Example 1*

In Fig. 2 we present acoustic recordings suitable for automatic recognition.

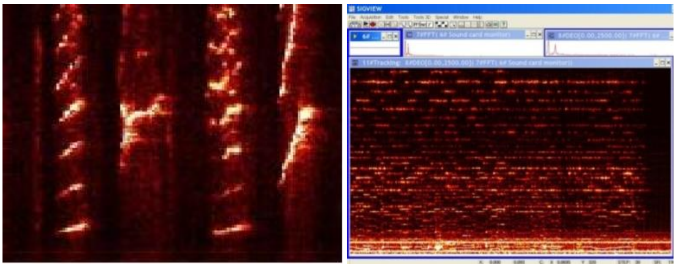


Figure 2
Examples of spectrogram features

The image on the left contains a spectrogram of a humpback whale song; the large structures correspond to melodious patterns, which are changing in both frequency and rhythm. On the right we have spectrogram of Die Kunst der Fuge, with all the tonal features of Bach's monumental music structure, in the last minute of performance by H. Walcha. Submarine sonar spectrograms can be rather similar to the ones presented above. Clearly, in the case when the percepted and imagined music tones are the same, we should expect spectrograms of a similar type and complexity, with spectrogram features of the imagined tones and music having similar properties as those of the acoustic origin, obviously with other features corresponding to the brain activity. We can distinguish inner tones, which are stable in frequency and thus flat in spectrograms, creating strong contrasts between high C versus low A as BCI signals, and tones with variable frequencies, with a kind of linear or nonlinear frequency inclination. Such features are in fact additional specialized filters detecting salient characteristics of the signal, as for example phonemic structures are detected in speech. The segments of the spectrograms of brain signals from different experiments with BCI are shown in Figs. 3 and 4.

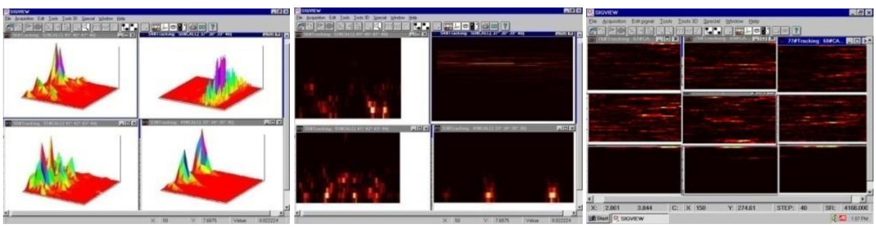


Figure 3
EEG spectral features

The left and center images in Fig. 3 show composite spectrograms of four EEG experiments exhibiting different types of imagined features prepared for automatic recognition with sporadic artifacts; the feature in the top right corner is lasting and stable in frequency, while the other three spectrograms exhibit the presence of dominant short lasting pulses; vertical - frequency range of 15 Hz, time scale of 10 s. The right image shows the fragments of real time spectrograms of EEG, containing features corresponding to the traces of inner tones. Vertically - frequency intervals; horizontally, time scale of 10 s. In Fig. 4, improved signal/noise provides object recognition directly, using partial linear dependence of two sources (spectrograms): the noise (random) is of a local nature while the signal components are present in both. These time spectra contain features related to imagined – inner tones, but these spectra show the presence of features corresponding to other processes in the brain as well. Searching for invariant structures in the spectrograms, we can obtain spectrograms containing only relevant features that will filter out all spectrogram structures not related to the inner tones. Before the similarity matching and measurements, in the case of inner tone classifier, we need to perform feature aggregation and disintegration. This process is application dependent: geometrically close objects should be aggregated into a single structure. On the other hand, features with a frequency inclination can be fragmented into individual tones (step functions) so that they can be subjected to matching with the set of calibration fuzzy tones.

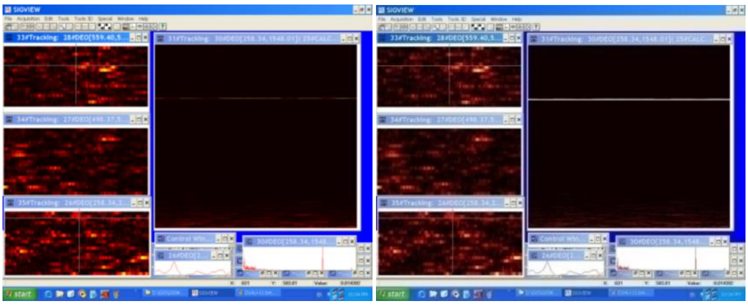


Figure 4

EEG recordings of inner music. Left: Parts of spectrograms of EEG channels with recording of inner (imagined) d2 tone; right: extracted inner tone in the right window in a composite spectrogram, shown enhanced bottom - right.

The automatic fuzzy classifier measures the degree of similarity with the pre-calibrated fuzzy tones, picking the best match (presented in [9, 13]). The method, based on local linear dependence and comparison to the set of special calibrated sets of adjustable fuzzy filters for more general application, is being further developed. Here, we have fuzziness at both modeling and intelligence levels. For some applications smoothing/roughening of the features is needed; these operations correspond to a variety of defuzzyfications.

## *Example 2*

In the following important examples of the arterial blood pressure (BP), we have further feature invariant determination; spectrogram features with relevant details are shown in Fig. 5. The exhibited morphology provides for different ways of capturing the characteristics of relevant features (the steps described above). In the top left picture, we can see tracing of the linear inclination of the high frequency feature whose non-homogenous morphology is shown in the lower window. Second order spectrum of the top feature longitudinal section is its essential invariant. On the right is displayed a BP spectrogram of features with their photometric representations, demonstrating the level of the noise surrounding these features and the problems related to the contour definition. These important patterns need a mathematically precise definition and characterization. Again, the (second order) spectrogram of the top left structure longitudinal section is its essential invariant, identical to the same invariant of the feature on the right. Here, pattern similarity matching is available as matching of the dominant lines in the corresponding second order power spectra.

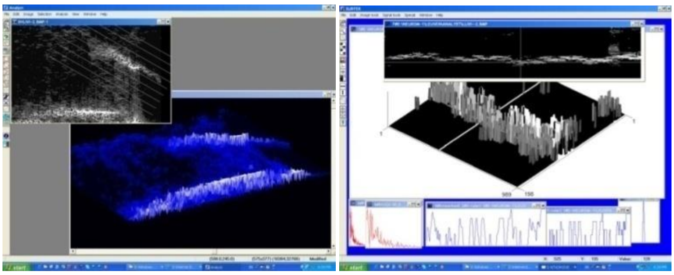


Figure 5

Recording of the blood pressure – BP; Features in the spectrograms with their photometric representations

### *Example 3*

Illustrations of the performance of discussed steps in the automatic feature recognition are shown in Figs. 6-8. For the investigation of some dynamic feature characteristics, e.g. those related to both time/rhythmic and frequency changes, or metric invariants, or those involving second order spectra, the analysis of features in their original shape, are necessary, together with some normalized form. Figure 6 shows an example of the step-by-step background noise elimination on the left. The automatic algorithm is based on threshold optimization: increasing the threshold until the image is fragmented into objects with larger than a parametric diameter and surface, subsequently reducing the threshold to enlarge detected objects till the level of surrounding noise increases above a parameter, while zeroing noncontiguous dot clusters. The automatic object contour definition is shown on the right, followed by the topology tracings of individual objects, which in turn is followed by the normalization based on its topologic structure. The normalization is necessary for the invariants determination and comparisons-matching. Originally developed for the structural study of microscopic images of chromosomes, this algorithm is adapted for time spectral feature localization, extraction and some normalization with feature comparison.

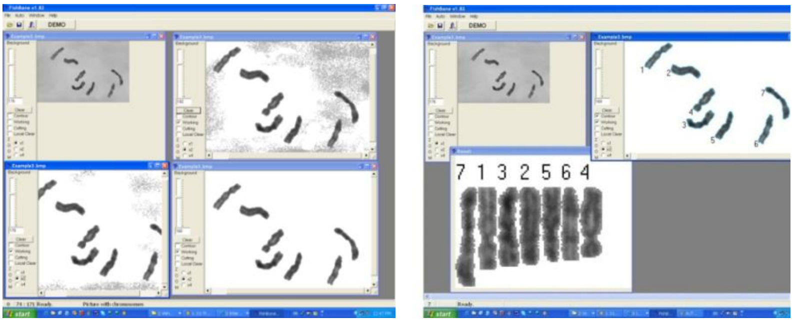



Figure 6

Left: background noise elimination; right: Automatic object contour definition

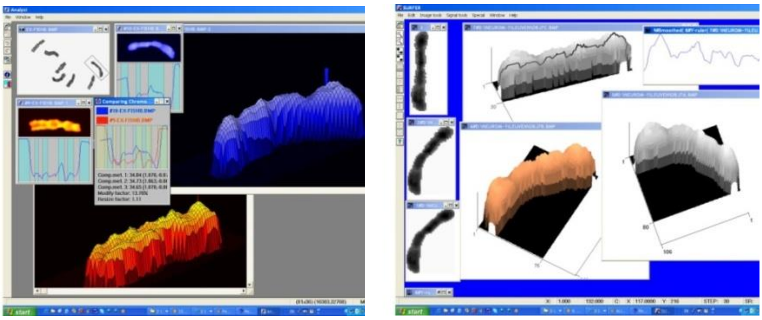


Figure 7

Left: The feature normalization and automatic comparison. Right: Feature step by step normalization with longitudinal sections exhibiting changes in morphology - dynamics.

## *Example 4*

One example of an acoustic spectrogram with locally well-defined and well-separated features, which are processed through the steps described above, is given in Fig. 8.

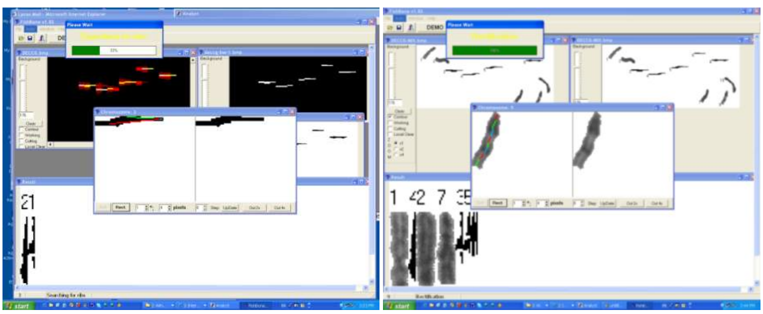


Figure 8

An automatic real time feature recognition; Synthetic spectrogram

Left: The real time spectrogram of a simple melody (deccg) with the automatic feature recognition; defuzzification was applied first, followed by normalization and the measurement of flexure angles and object extraction (lower). On the right, we have the same spectrogram enriched with a number of contiguous curved structures, non-constant in frequency, subjected to the same algorithm for the automatized feature recognition and classification, an adaptation of algorithm originally developed for automatic recognition of chromosomes in CCD-microscopic images.

## 2.2 Small Object Recognition

In this section we will give a brief overview of an alternative method for the efficient recognition of smaller, dot-like objects with the diameter < 10 pixels. Method can be applied to both matrices and vectors. Short frequency pulses are an important example of these. Spectral features which are stable and narrow in frequency might be examples of such sorts of vectors. Previously, we developed procedures for small object recognition and filtering by size based on the intensity discrimination (intensity of considered pixels). The method we present here is an improved Tomasi, Shi, Kanade procedure for the extraction of the characteristic features from a bitmap image (see [11] and [15]). It is robust and proved to be efficient, possessing all highly desirable properties, as illustrated in the subsequent figures. As an input we have a simple monochrome (0 = white, 255 = black) bitmap (matrix) $A$ of a fixed format (here presented with $400 \times 400$ pixel resolution). The components of $A$ signal amplitude values, or e.g. spectrogram intensities, will be denoted by $A(x, y)$, where $x$ indicates the corresponding row and $y$ indicates the corresponding column. Spatial $x$-wise and $y$-wise differences $I_x$ and $I_y$ are defined as follows:

$$I_x = \frac{A(x+1, y) - A(x-1, y)}{2}, \qquad I_y = \frac{A(x, y+1) - A(x, y-1)}{2} \tag{1}$$

The matrix $G$ of sums of spatial square differences is defined by

$$G = \sum_{x=p_x-\omega_x}^{p_x+\omega_x} \sum_{y=p_y-\omega_y}^{p_y+\omega_y} \begin{bmatrix} {I_x}^2 & I_x I_y \\ I_x I_y & {I_y}^2 \end{bmatrix}. \tag{2}$$

where $\omega_x = \omega_y$ is the width of the integration window (the best results are obtained with values between 2 and 4), while $p_x$ and $p_y$ are the indices corresponding to the indices $x$ and $y$ such that the formula (2) is defined; therefore, all inner pixels (i.e. pixels for which $I_x$ and $I_y$ can be defined) are included in the computation. We rewrite $G$ in the more compact form as

$$G = \begin{bmatrix} a & b \\ c & d \end{bmatrix}. \tag{3}$$

Using the above compact form (3) of $G$ we can compute its eigenvalues as

$$\lambda_{1,2} = \frac{a+d}{2} \pm \frac{\sqrt{(a-d)^2 + 4bc}}{2}. \tag{4}$$

Furthermore, for each inner pixel with coordinates (*x*, *y*) we define $\lambda(x, y)$ by

$$\lambda(x, y) = \min\big(\lambda_1(\mathrm{x}, \mathrm{y}), \lambda_2(\mathrm{x}, \mathrm{y})\big). \tag{5}$$

Finally, for the given lower threshold $T_{\min}$, the parameter $A_{\max}$ (in our examples $A_{\max}$ is equal to 255) set the value

$$\lambda_{\max} = \max\{\lambda(x, y) \mid (x, y) \text{ } is\ an\ inner\ pixel\}. \tag{6}$$

We define the extraction matrix $E$ by

$$E(x, y) = \begin{cases} \frac{A_{\max}}{\lambda_{\max}} \cdot \lambda(x, y), & \frac{A_{\max}}{\lambda_{\max}} \cdot \lambda(x, y) > T_{\min} \\ 0, & \frac{A_{\max}}{\lambda_{\max}} \cdot \lambda(x, y) < T_{\min} \end{cases}, \tag{7}$$

When two images or spectrograms are available (two consecutive shots or two significantly linearly independent channels) we obtain a solution in an even harder case for automatic extraction. Let $B$ and $C$ be two images where every pixel is contaminated with noise which has a normal Gaussian distribution, in which a stationary signal is injected, objects at coordinates $(x_1, y_1), \ldots (x_{10}, y_{10})$, all with an intensity of e.g. $m$ (within [0, 255] interval) and fluctuation parameter $p$; we generate the new binary image $A$ in two steps:

$$A(x, y) = \mathrm{abs}(B(x, y) - C(x, y)) \tag{8}$$

If $A(x, y) < p$ then $A(x, y) = 255$

else $A(x, y) = 0$;

The above simple discrimination reduces random noise significantly and reveals the signals together with residual noise. By performing the procedure defined by the equations (1) thru (7), we obtain the filtered image with extracted signals. The method is adaptable, using two parameter optimization (minimax): the minimalization of the integral surface of detected objects, then the maximization of the number of the small objects.

An alternative method for the detection/extraction of small features is based on a bank of Kalman filters. After the construction of the initial sequence of images, $Z_k$, the bank of one-dimensional simplified Kalman filters (see e.g. [16]) is defined using the iterative procedure as follows:

$$K_k(x, y) = \frac{P_{k-1}(x, y) + Q}{P_{k-1}(x, y) + Q + R} \tag{10}$$

$$\hat{X}_k(x, y) = \hat{X}_{k-1}(x, y) + K_k(x, y) \cdot (Z_k(x, y) - \hat{X}_{k-1}(x, y))$$

$$P_k(x, y) = (1 - K_k(x, y)) \cdot (P_{k-1}(x, y) + Q).$$

Initially, $P_0(x, y) = \hat{X}_0(x, y) = 0, Q = 1, R = 100$, where $Q$ is the covariance of the noise in the target signal and $R$ is the covariance of the noise of the measurement. Depending on the dynamics of the problem we put: the output filtered image in $k^{\text{th}}$ iteration is the matrix $\hat{X}_k$, the last of which is input in the procedure described by equations (1) to (7), finally generating the image with the extracted objects.

This method shows that it is not necessary to know the signal level if we can estimate the statistical parameters of noise and statistics of measured signal to some extent. In the general case, we know that its mean is somewhere between 0 and 255 and that it is contaminated with noise with the unknown variance.

The method of small object recognition, originally developed for the marine radar object tracking, works with vectors equally well. It is applicable to the automatic extraction of signals which are embedded in the noise and imperceptible (also in the spectra) in the case when we can provide at least two sources which are sufficiently linearly independent (their linear dependence on the signal components is essential for the object filtering – extraction), or in the situations when the conditions for application of Kalman filters are met.

## *Example 5*

In this example, we have introduced several dots (useful signals) with an amplitude of $a = 120$, and we have contaminated the image with random and cloudlike noise. The image on the left in Fig. 9 shows a bitmap with random contamination of the signal – dots. The image on the right in the same figure shows the resulting bitmap after the application of the procedure for noise reduction. After the initial setting $A_{\max} = 255$ and $T_{\min} = 124$, the extraction procedure yields the image shown below in Fig. 10.

The image on the left in Fig. 11 shows a similar example of the signal – dots contaminated with a cloudy noise containing granular elements which are similar in size and intensity to the signal. The image on the right shows the results of the reduction of noise: some new dots belonging to the noise cannot be distinguished from the signal – top and low right.

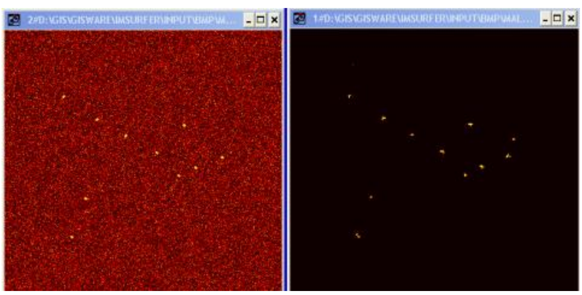


Figure 9

Reduction of noise

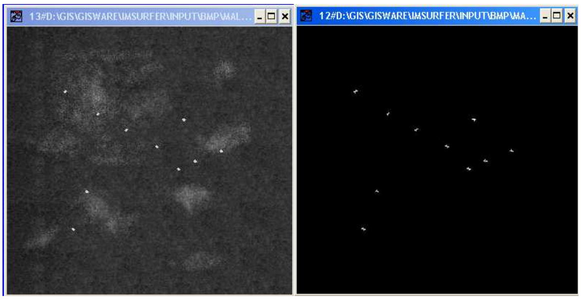


Figure 10

left: signal – dots, contaminated with cloudy noise; right: extraction of signal.

Note that the amplitude of the target signal is lower than the chosen lower threshold (images in Figs. 9 to 11).

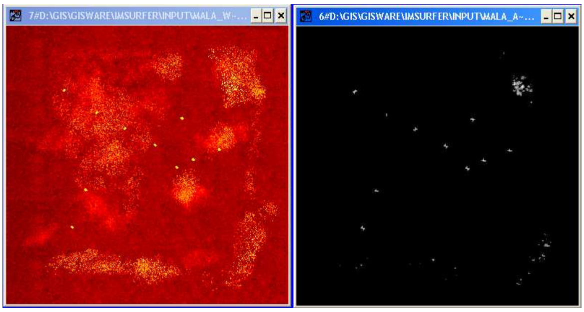


Figure 11

Signal extraction

## *Example 6*

Here we illustrate the application of the method of small feature extraction with two independent sources, shown in Fig. 12, with the signals embedded in the noise and the process of signal extraction.

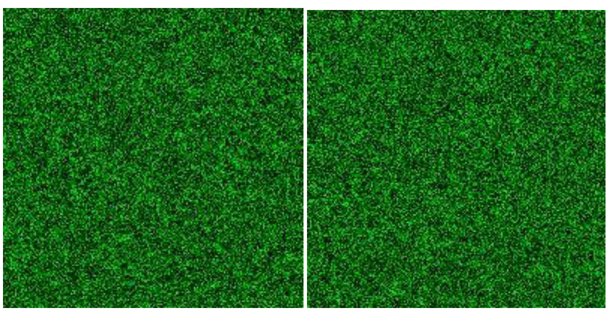


Figure 12

Two Gaussian noise images with the injected small objects below threshold

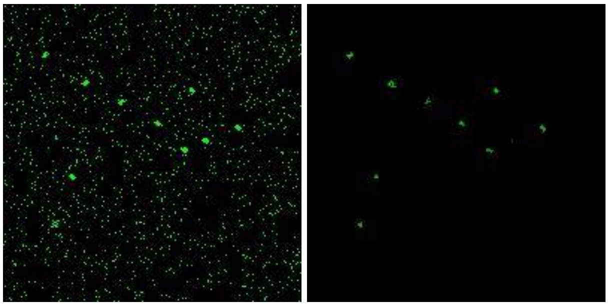


Figure 13
Left: Extraction of the objects based on simple discrimination. Note the presence of residual noise – smaller dots. Right: after application of the above method to the binary image on the left, the noise is completely reduced, yielding fully automatic small object recognition.

## *Example 7*

The application of Kalman filters in small object extraction. In the experiment shown, the initial sequence of images, $Z_k$, of the size $200 \times 200$ pixels is generated as follows. First, in each image we have introduced noise by $Z_k(x,y) = \mathrm{randn}(0{,}90)$; here "randn" generates random numbers in the interval $[0{,}255]$ using Gaussian distribution with $\mu = 0$ and $\sigma = 90$. Then, in every image we injected 10 objects (useful signals) at the same positions, each of them of a size around 10 pixels, with random (Gaussian) fluctuation of intensity around value 120. After the construction of the initial sequence of images, the bank of $200 \times 200 = 40000$ one-dimensional simplified Kalman filters is defined using the iterative procedure as above. The process of noise elimination and feature extraction is shown in the images in Fig. 15.

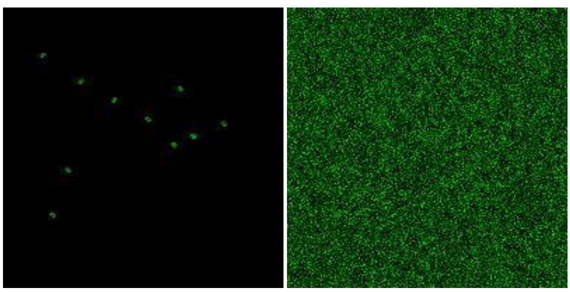


Figure 14
Left: Injected signal; Right: The image on the left injected with the Gaussian noise contamination.

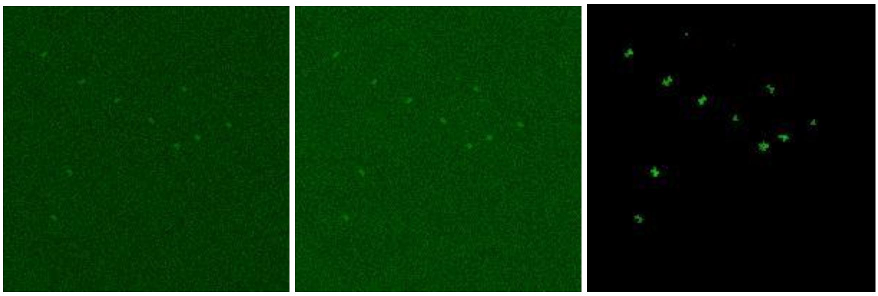


Figure 15

Left: The result of processing after the $21^{st}$ iteration of Kalman filter banks: Center: The result of processing after 34 iterations; Right: The result of processing after 36 iterations and application of the method described by formulae 1- 8, providing a complete object extraction.

### *Example 8*

Localization and extraction of the small size features in spectrograms of diverse origin. In Fig. 16 left and center, we show examples from [17] (similar examples are widely distributed in literature), which are used in brain connectivity pattern detection. The resolution here is: pix = 2 Hz *0.5 s; 2*1; 2*2. Note the size of the granulae in the shown spectrograms. The great majority of the important features are within 6x6 pix, and the method of small object recognition performs very well even with some noise contamination. In Figure 16, on the right is shown our RT reproduction (whistling) of the melody used by A. Ioannidis in his impressive presentation of MEG tomography, with the same melody stimulation (available on his site as well); all spectrogram granulae are within 4x4 pix size (Easy to generate with the available DEMO at our site). Good examples for the application of this method are spectrograms in the second, third and fourth quadrant in Fig. 3 left and center; in the presented context this method can be performed concurrently with the method from section 2.1 for parallel recognition of larger structures, as are those shown in the first quadrants of these images and features in the image on the right. The same is true in the case when both types of structures are overlapping, as in Fig. 17 with FISH signals. The extraction of the small objects within the cloudy structures in conjunction with the earlier described method based on the contour detection provides a means for the automated and concurrent detection of small and large structural components independently.

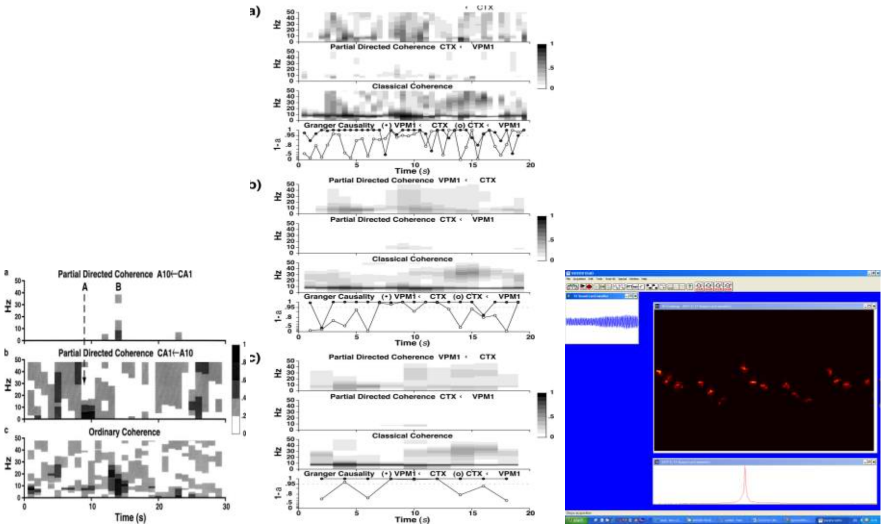



Figure 16

Left and center: brain connectivity relevant spectrograms from [17]: frequency range 50 Hz, time 35 s, granular dimensions easy readable, Resolution: pix = 20 Hz *4 s; a number of small size spectrogram objects are in the size of up to 5x5 pixels. Right: RT reproduction of A. Ioannides MEG example with spectrogram consisting of small features.

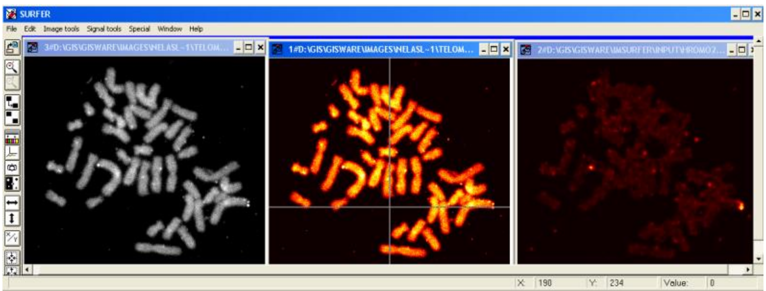



Figure 17

Extraction of FISH signals from chromosomes with the same method

## Conclusion and Discussion

This paper addresses the problem of the automatized recognition of features in signals and their Fourier or wavelet spectra and spectrograms. The algorithms presented use techniques developed for image processing and are suitable for morphologic investigations. These algorithms are able to localize and extract features and to determine their topologic and geometric characteristic invariants, which are used to represent and to classify the objects by applying subtle similarity measures. Small object recognition in cases of heavy contamination by noise of mainly random nature is successfully performed in rather general circumstances. This is applied well to the automatic recognition of short frequency pulses in the spectrograms. The methods presented are useful even for semiautomated or manual cases, when for example their automated application is limited because of some of the discussed reasons and can be applied for a detailed structural inspection and comparison of the features. Due to a modest complexity, all are real time (RT) applicable, even without the enhanced hardware.

Obviously, the real experimental practice always offers nice counterexamples that do not fit well into the predefined conceptual scheme, such as parts of the features in the BP spectrograms at the top of Fig. 5 (or granular noise indistinguishable from the objects – dots in Fig. 15). Photo morphology revealing the lower feature contour is very fuzzy. The left part of the top structure can hardly be called a feature at all, but rather a random cloud of dots. But the complete set of the dots is definitely functionally related. We have the appearance of a feature out of the randomness, which characterizes some micro-phenomena that are still not semantically bound at the macroscopic scale. This is a kind of reality where the method presented here might exhibit some problems. In circumstances like this, one should search for transformations that are able to convert information in the signal into simpler topological or geometrical structures. Nevertheless, our approach can be well applied to a variety of different cases of real time spectrogram features. Similarity and pattern classification in the continuous domains is properly modeled with metrics in the classic metric spaces, and the majority of our implemented similarity measures are of this kind. There are many other approaches. One mentioned earlier is the recognition of bumps in EEG spectra [18], which is done in the similar spirit as the perspective of this paper. The brain is investigated with nonlinear analysis methods too. The chaos theory is applied in the analysis of brain activity; the estimation of the fractal dimension in time domain gives a measure of signal complexity. It has been successfully used with some brain injuries [19].

Finally, it is worth mentioning that the theoretical development and steadily growing applications of pseudo-analysis are giving alternative methods for the mathematical design of the extraction criteria, automatic threshold design and so on; see for instance [20, 21].

**Acknowledgements**

This research has been partly supported by Serbian Ministry of Education and Science, projects III-41013, TR36001 and ON174009.

**References**

[1] A. Jovanović, *CD-ROM: CCD* Microscopy, Image & Signal Processing, (School of Mathematics, Univ. of Belgrade, Belgrade, 1997)

[2] A. Jovanović, A. Perović, Brain Computer Interfaces - Some Technical Remarks, International Journal of Bioelectromagnetism 9(3): 91-102, 2007

[3] F. Cincotti, D. Mattia, C. Babiloni, F. Carducci, L. del R. M. Bianchi, J. Mourino, S. Salinari, M. G.Marciani, F. Babiloni, Classification of EEG Mental Patterns by Using Two Scalp Electrodes and Mahalanobis Distance-based Classifiers, Method of Information in Medicine 41(4) (2002) 337-341

[4] G. Pfurtscheller, C. Neuper, C. Guger, W. Harkam, H. Ramoser, A. Schlög, B. Obermaier, and M. Pregenzer, Current Trends in Graz Brain-Computer Interface (BCI) Research, IEEE Transaction on Rehabilitation Engineering 8(2): 216-219, 2000

[5] F. Nijboer, E. W. Sellers, J. Mellinger, M. A. Jordan, T. Matuz, A. Furdea, S. Halder, U. Mochty, D. J. Krusienski, T. M.Vaughan, J. R. Wolpaw, N. Birbaumer, A. Kübler, A P300-based Brain-Computer Interface for People with Amyotrophic Lateral Sclerosis. Clinical Neurophysiology 119(8): 1909-1916, 2008

[6] A. Jovanović, Brain Signals in Computer Interface, (in Russian, Lomonosov, Russ. Academy of Science), Intelektualnie Sistemi, 3(1-2): 109-117, 1998

[7] R. J. Zatorre, A. R. Halpern, Mental Concerts: Musical Imagery and Auditory Cortex, Neuron, 47: 9-12, 2005

[8] J. K. Kroger, L. Elliott, T. N. Wong, J. Lakey, H. Dang, J. George, Detecting Mental Commands in High-Frequency EEG: Faster Brain-Machine Interfaces, in Proc. 2006 Biomedical Engineering Society Annual Meeting, (Chicago, 2006)

[9] Wlodzimierz Klonowski, Wlodzisław Duch, Aleksandar Perovic, and Aleksandar Jovanovic, "Some Computational Aspects of the Brain Computer Interfaces Based on Inner Music," Computational Intelligence and Neuroscience, Vol. 2009, Article ID 950403, 9 pages, 2009. doi:10.1155/2009/950403

[10] G. Pfurtscheller, F. H. Lopes da Silva, Event-related EEG/MEG Synchronization and Desynchronization: Basic Principles, Clinical Neurophysiology, 110(11): 1842-1857, 1999

[11] J. Y. Bouguet, Pyramidal Implementation of the Lucas Kanade Feature Tracker, preprint, http://robots.stanford.edu/cs223b04/algo_affine_tracking.pdf

[12] G. Bradski, Computer Vision Tracking for Use in a Perceptual User Interface, preprint, http://www.cse.psu.edu/~rcollins/CSE598G/papers/camshift.pdf

[13] S. Spasić, A. Perović, W. Klonowski, Z. Djordjević, W. Duch, A. Jovanović, Forensics of Features in the Spectra of Biological Signals, International journal for bioelectromagnetism 12(2): 62-75, 2010

[14] A. Jovanović, A. Perović, W. Klonowski, W. Duch, Z. Đorđević, S. Spasić, Detection of Structural Features in Biological Signals, Journal of Signal Processing Systems 60(1): 115-129, 2010

[15] J. Shi, C. Tomasi, Good Features to Track, preprint, http://www.ai.mit.edu/courses/6.891/handouts/shi94good.pdf

[16] G. Welch, G. Bishop, An Introduction to the Kalman Filter, University of North Carolina at Chapter Hill, Chapter Hill, TR 95-014, April 5, 2004

[17] K. Sameshima, L. A. Baccala, Using Partial Directed Coherence to Describe a Neuronal Assembly Interactions, J. Neurosci Methods 94: 93-103, 1999

[18] F. B. Vialatte, J. Sole-Casals, J. Dauwels, M. Maurice and A. Cichocki, Bump Time-Frequency Toolbox: a Toolbox for Time-Frequency Oscillatory Bursts Extraction in Electrophysiological Signals, BMC Neuroscience 10(46), doi:10.1186/1471-2202-10-46, 2009

[19] S. Spasić, M. Culić, G. Grbić, Lj. Martać, S. Sekulić, D. Mutavdžic, Spectral and Fractal Analysis of Cerebellar Activity after and Repeated Brain Injury, Bulletin of Mathematical Biology 70 (4): 1235-1249, 2008

[20] E. Pap. Pseudo-Analysis Approach to Nonlinear Partial Differential Equations. Acta Polytechnica Hungarica 5(1): 30-44, 2008

[21] E. Pap, V. Bojanić, M. Georgijević, G. Bojanić. Application of Pseudo-Analysis in the Synchronization of Container Terminal Equipment Operation. Acta Polytechnica Hungarica 8(6): 5-21, 2011

# Information and Knowledge Retrieval within Software Projects and their Graphical Representation for Collaborative Programming

**Ivan Polášek[1], Ivan Ruttkay-Nedecký[2], Peter Ruttkay-Nedecký[2], Tomáš Tóth[2], Andrej Černík[2], Peter Dušek[2]**

[1] Faculty of Informatics and Information Technology, Slovak University of Technology in Bratislava, Slovakia, e-mail: polasek@fiit.stuba.sk

[2] Gratex International, a.s., Bratislava, Slovakia, www.gratex.com, e-mails: peter.ruttkay-nedecky@gratex.com, ivan.ruttkay-nedecky@gratex.com, tomas.toth@gratex.com, andyc@gratex.com, pdusek@gratex.com

*Abstract: This paper proposes information and knowledge mining in the source code of medium and large enterprise projects. Our methods try to recognize structures and types of source code, identify authors and users to enhance collaborative programming, and support knowledge management in software companies. Developers within and outside the teams can receive and utilize visualized information from the code and apply it to their projects. This new level of aggregated 3D visualization improves refactoring, source code reusing, implementing new features and exchanging knowledge.*

*Keywords: information and knowledge mining; knowledge management; collaborative programming; visualization; recognition; authors; users; source code type; code tagging*

## 1 Introduction

This paper describes our approach to information and knowledge mining, which aims to support collaborative programming and to help software developers in medium and large teams to understand complicated code structures and extensive content as well as to identify source code authors and concrete people working with existing modules. Accordingly, newcomers as well as other colleagues can reference real source code authors and users more efficiently.

There are some works analyzing collaborative source code development and the impact of individual authors' contributions to the selected characteristics of source code [7]. Authors from the data mining domain often utilize code line oriented SW tools that determine only the latest author of every line [8]. There are also approaches to monitor source code users' work [3] [4] including reading, modifying or showing interest in the source code in some other way.

The determination of the source code topics can be used for purposes of identifying the domain expertise of developers [5]. It is also possible to support program comprehension by identifying common topics in source code [6].

For tagging (keywords for features, authors, patterns/anti-patterns, rating, etc.) and comments, we can use two approaches:

- attach tags using comments inserted directly into source code [1],
- separate tags into external DB or sources [2].

We have studied both approaches and now we prefer the second one.

The contribution of our method is to find new level of aggregated visualization for the identified and interconnected information from the source code for collaborative development.

We have decided to create an intelligent environment that takes advantage of data mining methods. It presents information and knowledge in a 3D graphical form and highlights the interconnections between the information. The key goals of this environment are to support the source code reuse, to support refactoring and the implementing of new features. Knowledge exchange and management is also increased through integration with *Gratex Knowledge Office* and source code tagging.

# 2 Visualization

For knowledge mining purposes, our method extracts *Abstract Syntax Tree* (AST) from the source code [10] [14] [15] and the resulting graph structure is then visualized with acquired information.

We have developed our own graph visualization engine using the graphical engine Ogre3D [11] and the Fruchterman-Reingold [12] force-directed layout algorithm (see Figure 1).

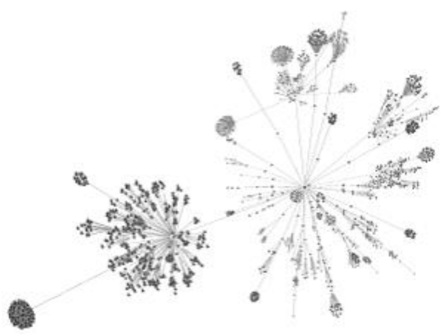


Figure 1
Visualization of the source code using Fruchterman-Reingold algorithm

We are also developing our own graph representations, which are more semantically oriented than a general force-directed algorithm. Snapshots of these graphs are provided in the following Figure 2.

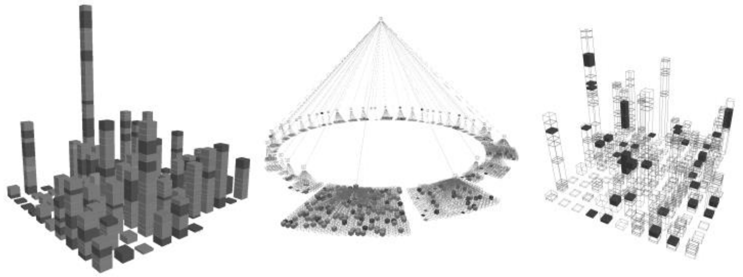


Figure 2
Semantically-oriented source code visualizations

# 3 Tier Recognition

The purpose of tier recognition is to identify domains in three-tier-based systems. These systems are composed of presentation *tier*, *application processing tier* and *data management tier* [16]. Each domain is identified by a tier of the associated source code. It is possible to correlate this information through associated entities in AST with other mined information. Therefore, it is possible, for example, to determine the developer’s orientation on a given tier. Tier recognition also simplifies navigation in the source code through clustering.

The recognition is performed on multiple levels - classes, namespaces, projects and others. Each code entity (class, method, field, project, etc.) is described by its child code entities in the sub level (e.g., a class is described by its methods), but also the code entity itself describes its child code entities (e.g., methods in a data tier class will most likely belong to the data tier).

In our approach, we determine tiers in several ways that can be combined to achieve more precise results.

## 3.1 Keywords

It is common that the name of the type in source code (class, interface, structure, enumeration) describes its purpose. For instance the name, “DbCommand”, from the first glance tells us that the type represents a kind of a database command encapsulation. In our method, we use this common practice to identify specific keywords in names of code identifiers with the intention of recognizing the code tier of a given type.

#### 3.1.1 Identifiers

Our method uses multiple identifier types as the source for searching. As can be expected, the primary identifier is *type name*. Additional identifiers are *base type name*, *namespace* and, in the case of an integrated development environment that supports grouping of source code into projects, also *project name*.

#### 3.1.2 Keywords Dictionary

This method requires a set of known keywords and their tier assignments as an input. We call this set *keyword dictionary*. It is possible for one keyword to be used in multiple tiers. Therefore, it is essential to perceive each tier assignment of a keyword as a rate that defines how much the keyword is specific for a given tier. This method requires that the sum of tier assignment rates is equal to one for each keyword in the dictionary. For our test, we used only a small and manually constructed dictionary, but we are planning to use an automatic keyword extraction to create a nontrivial dictionary. An example of a dictionary is presented in the following table.

Table 1

Example keywords dictionary

| Keyword | Data | App | Presentation |
|---|---|---|---|
| Data | 0,90 | 0,10 | 0,00 |
| Db | 1,00 | 0,00 | 0,00 |
| Table | 0,60 | 0,00 | 0,40 |
| Workflow | 0,10 | 0,80 | 0,10 |
| Control | 0,00 | 0,00 | 1,00 |
| **…** | … | … | … |

#### 3.1.3 Word Extraction

The task of the word extraction is to divide an identifier’s name into separate words that can be compared with keywords in the dictionary. The extraction result can be seen in the following figure.


Figure 3

Word extraction

For the purposes of word extraction, we created regular expressions which define the split points in the identifier’s name.

- **Class name**- [_<>,]|\d+|(?<=[^A-Z])(?=[A-Z])|(?<=[A-Z])(?=[A-Z][a-z])

- **Interface name**- ^I(?=[A-Z])| [_<>,]|\d+|(?<=[^A-Z])(?=[A-Z])|(?<=[A-Z])(?=[A-Z][a-z])
- **Namespace name**- [_.]|\d+|(?<=[^A-Z])(?=[A-Z])|(?<=[A-Z])(?=[A-Z][a-z])
- **Project Name** - [_.]|\d+|(?<=[^A-Z])(?=[A-Z])|(?<=[A-Z])(?=[A-Z][a-z])

### 3.1.4 Association Rate

After extracting the words from the identifiers, the partial association rate of each identifier type is computed. Each partial rate is in the range of <0.0, 1.0>. The partial rate is computed using the following pseudocode.

```
GetTierPartialRate(IN: words, IN: lookupTable, OUT: rate, OUT: baseRate)
  baseRate = 0
  for word in words :
     if lookupTable.Contains(word) :
        keyword = lookupTable[word]
        baseRate++
        tierRate += keyword.Rate
  if baseRate > 0 :
        rate = tierRate/baseRate
  else :
        rate = 0
```

Words that have not been successfully matched with any keyword are not included in the computation. Therefore, they do not lower the resulting rate.

In the next step, these partial rates are merged into final weighted rate. For this purpose, each identifier type has been given a weight in the range of <0.0, 1.0>. It is not required that the sum of these weights is equal to one. The computation of the final weighted rate is presented in the following pseudocode.

```
GetTierWeightedRate(IN: weights[], IN: unitRates[], IN:baseRates[], OUT:
weightedRate)
  weightSum = 0
 for i  in [0..3] :
     if baseRates[i] != 0 : weightSum+=weights[i]
 if weightSum == 0 :
    weightedRate = 0
 else:
   for i  in [0..3] :
     if baseRates[i] != 0 :
        weightedRate+=unitRates[i]*weights/weightSum
```

### 3.1.5 Case Study

In this part, we will demonstrate the application of this method on a very small set of types (Table 2).

Table 2
Example types

| Type | BaseType | Project | Namespace |
|---|---|---|---|
| IGetBlobHelper | | Frm.DataInterfaces | DataInterfaces.Blob |
| Graph3dControl | UserControl | Graph3d.WinForm | Graph3d.WinForm |
| IWorkflowHelper | | Frm.DataInterfaces | DataInterfaces.Workflow |
| IWorkflowEntity | | Frm.DataInterfaces | DataInterfaces.Workflow |
| WSFrmDataContext | DataContext | Frm.DataClasses | DataClasses |
| IDBHistoricalTable | | Frm.DataInterfaces | DataInterfaces.HistTable |

We will use the dictionary from Table 1 as the keyword dictionary. The following charts display computed partial rates for each keyword type.

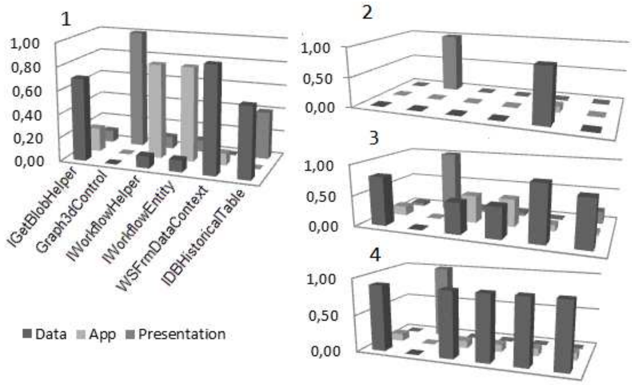



Figure 4
Unit rate assignments (1-by name; 2-by base types; 3-by namespace; 4-by project)

These rates are then composed to a final result using the following identifier type weights (Table 3).

Table 3
Example identifier type weights

| Identifier | Weight |
|---|---|
| Name | 1 |
| Base Type Name | 0.7 |
| Namespace | 0.6 |
| Project Name | 0.6 |

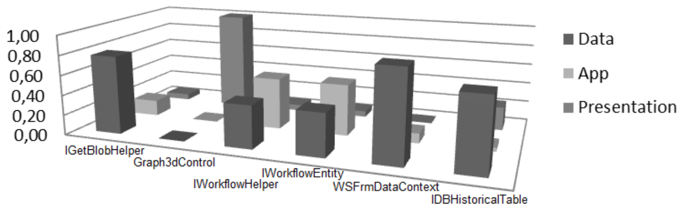



Figure 5
Weighted assignment rates

Figure 6 shows an application of this method on a real-life project represented as a *Manhattan graph*. Each bar represents a single type. Each sub bar represents a single method, the height of which is determined by its code line count. Bar darkness represents the rate of a DB tier assignment.

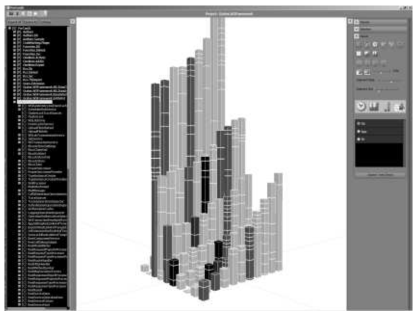


Figure 6
Code tiers as a Manhattan graph in our visualization environment

## 3.2 Standard Types

Standard types represent generally known types that are usually included in programming languages or frameworks. Source code is created using standard types and types that are recursively created from standard types. This method searches for used standard types in source codes and determines implemented tiers using a knowledge of the relationships between standard types and tiers.

Figure 7 shows how tiers are determined from given type declarations. For each type declaration several steps are performed. First, used standard types are extracted and their namespaces are identified. Using a predefined lookup table, the tier ratios of the extracted namespaces are determined.

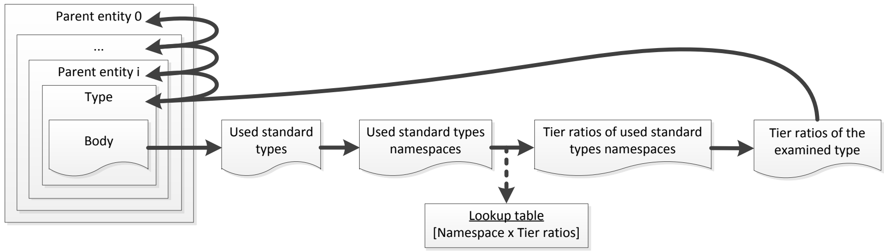



Figure 7
Determining tiers by examining used standard types

Table 4 presents a fragment of a lookup table that maps standard .Net namespaces to ratios for each of the three tiers. The table is constructed manually in our work, but we are planning to use automated crawling techniques and clustering algorithms in the future.

If a namespace is not found in the lookup table, its parent namespace is searched for, and so on up to the root namespace. If not even the root namespace is found, the given namespace is ignored.

Standard types themselves can also be present in the lookup table. Furthermore, identified types and their namespaces can also be placed back to the lookup table and extend the knowledge base of the process.

Table 4
A fragment of a lookup table, which maps .Net namespaces to tier ratios

| Namespace | Presentation tier | Application tier | Data tier |
|---|---|---|---|
| System.ComponentModel | 0.4 | 0.3 | 0.3 |
| System.Data | 0.05 | 0.05 | 0.9 |
| System.DirectoryServices | 0.15 | 0.7 | 0.15 |
| System.Drawing | 1.0 | 0.0 | 0.0 |
| System.Globalization | 0.6 | 0.1 | 0.3 |
| System.IdentityModel | 0.2 | 0.8 | 0.0 |

Final ratios for the examined type are calculated from extracted ratios.

In our approach we calculate an arithmetic average of all extracted ratios for each tier separately. Standard namespaces and their types could be weighted. Ratios of each entity are calculated from ratios of its child entities (e.g. namespace and its types).

Figure 8 shows example results of this method. Tiers were determined for a fragment of a system where mostly the presentation tier is present. Empty rows represent types that were not determined, as they can belong to any tier.

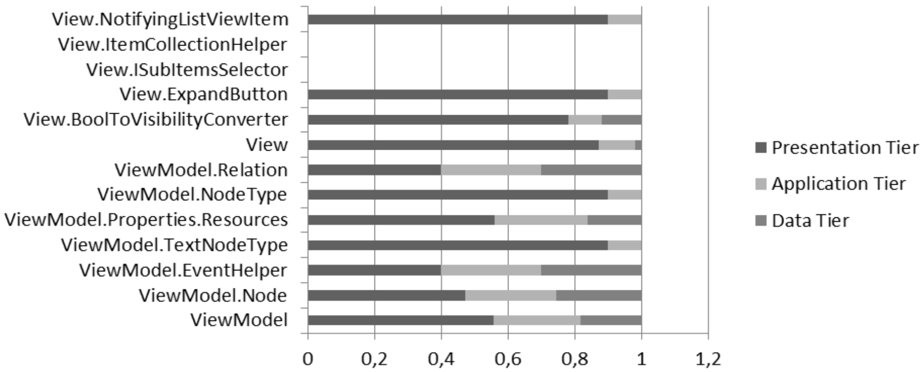



Figure 8
Example - final ratios of examined types

## 3.3 Project Meta-Information

Integrated development environments often store meta-information describing source codes in separate data sources. For example Microsoft Visual Studio (VS) stores information about source code files in project and solution files. We examine these files and extract two kinds of information - *project types* and *project output types*.

VS project has one or more project types. Each type describes what framework the project uses, whether it is an installer or an extension, and so on. Some frameworks support the development of specific tiers; for example Windows Presentation Foundation supports the development of a presentation tier.

VS projects can produce three kinds of outputs. *Windows application* is an executable file with a graphical user interface which, in general, implements the presentation tier. *Console application* is also an executable file, but without a graphical user interface; therefore we consider it as being without a presentation tier. *Class library* is a dynamic link library that can implement all three tiers.

# 4 Source Code Users

Our method acquires information about users and their activities on code entities. This information is used to model user behavior and represents a source of knowledge for our visualization. Two methods are presented in this chapter.

## 4.1 Code Entities Checkout

Revision control systems (RCS) usually allow users to view which source code files are currently being edited by which users. This helps to avoid conflicts when two users edit the same file. But conflicts mostly occur only when two users edit the same portion of the same file. It is not necessary to lock the whole file.

We look at source code not only as a set of files. We go deeper into these files to examine their code entities. We determine which code entities (not just whole files) are currently being edited by which users. This information could then be used to lock concrete code entities rather than whole files.

We use the following algorithm to determine which code entities are currently being edited:

```
GetChangedCodeEntities(OUT:changedCodeEntitySet)
  changedLocalFiles = GetChangedLocalFiles()
  for localFile in changedLocalFiles
    originalFile = DownloadOriginalVersionFromRCS(localFile)
    changedLineIndices = Compare(originalFile, localFile)
    originalFileAst = ExtractAst(originalFile)
    for lineIndex in changedLineIndices
      changedCodeEntity = originalFileAst.GetCodeEntityAt(lineIndex)
      changedCodeEntitySet.Add(changedCodeEntity)
```

## 4.2 User Activity on Code Entities

This method monitors the activities of users on individual code entities. Our goal is to determine on which code entities and how users are currently working and also to measure this activity.

In our approach, *activity* is a general term for anything a user performs with a single code entity: editing, pressing the mouse over it, reaching it in source code, etc. Each activity has a value in interval <0, 1> which represents a measurement of how active the user is on the code entity. *0* means no activity and *1* means maximum activity.

A user can perform more activities on a single code entity at the same time.

Figure 9 shows how the final activity of a user on a single code entity is determined. The final activity is composed of all activities the user performs on that code entity. This includes different activities and the same activities performed in different places (typing into a single code entity in two different editors).

First, final values for all different activities are calculated. For each different activity, this is the maximum value from all places where the activity is performed

– e.g. *typing* in two editors. This represents information: what different activities the user performs on the code entity and a measurement of "how much".

The final activity for the code entity is calculated from all different activities. We sum up all different activities for the code entity. At this point we have the total value, "how much" the user is active on the code entity.

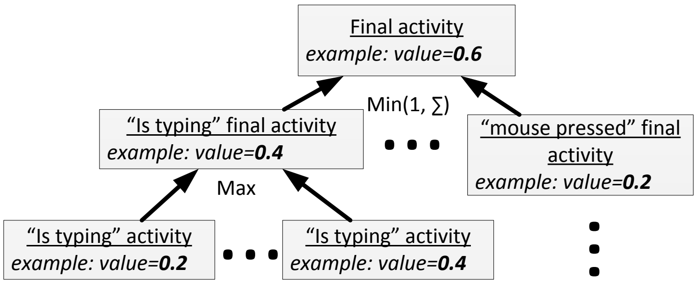



Figure 9

Computing of the final activity of a single user on a single code entity

### 4.2.1 Activity Initial Value and Cooling of Activities

Every activity has a predefined maximum and minimum value depending on its relevance. When an activity occurs, its value is maximal. When it is not performed for a period of time, its value starts to decrease down to the minimal value. We call this process the *cooling down of activities,* and it expresses the decreasing interest of the user in the code entity.

### 4.2.2 Case Study

Figure 10 shows a prototype where three activities are monitored for a single user. The cooling down of activities is also shown. These activities are the following:

- In viewport – the code entity is reached in a code editor (scrolling etc.)
- Mouse down – the user pressed the mouse over the code entity
- Is typing – the user types into the code entity

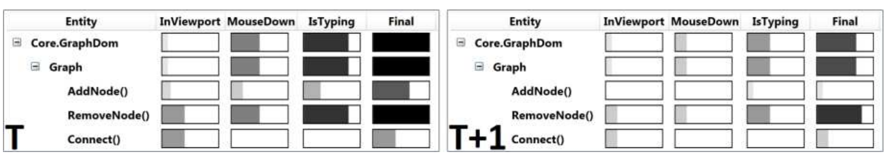



Figure 10

Activities for a single user. Activities are being cooled down.

# 5 Source Code Authors

To identify the programmer's skills and quality of his work, we must recognize the authors in the whole source code version history, evaluate their particular contributions and associate them with analyzed code characteristics. We can then identify the responsibilities of each programmer and their impact on team development.

## 5.1 Author Characteristics

An author of the source code can be anyone who creatively changed the content of the source code file. We can divide the authors into three basic groups:

- real authors of the content, who modify the logical nature of source code (adding, modifying or deleting code entities),
- editors who modify the form of source code record but not its logical meaning (refactoring, sorting elements, formatting code, …),
- reviewers who can comment on the code or an update of the code due to newer version of used libraries.

By a different criterion, we can separate authors to first, being the author of a source code part, and coauthors.

Alternatively, we can reduce every developer to a coauthor, because everyone can considerably modify previous versions of source code.

Also, there are authors whose source code part has persisted to the last, or a particular version of the source code and authors whose source code part was deleted or considerably modified over time.

Another criterion determines how the author can be mapped or bound to a given source code entity and how these bindings will be represented:

- Authors of structural and syntactical entities of source code (project, package or module, namespace, class, interface, field, property, method, statement),
- Authors of lines of a source code.

## 5.2 Presentation of Authorship

The presentation of authorship can be solved by various approaches. In the context of source code versions (changesets), it can be presented in the following ways:

- Authorship only in a particular version: the authors of the last changes of the source code elements.
- Authorship based on the life cycle of the source code development to a particular changeset.

- Authorship based on the whole life cycle of the source code, not only until the presented changeset, if it is not the latest one.

Information about authorship should cover: author, changeset, date of change, and the type of change that has been done (add, edit, delete).

## 5.3 Determination of Authorship

The author is defined by changes that the author made in some particular version and in parts of the source code. If we need to evaluate authorship in the whole history of the code entity, we must match code entities between several versions of the source code. This is not trivial due to change of identifiers of code entities and changes of entities positions in AST [14]. The identification of source code entities is given by their similarities or matching [10]. This approach determines the authorship of the source code entities in object oriented paradigm, where syntax units can be represented as AST. It is based on the extraction of source code entity changes and can be divided into several phases (Figure 11):

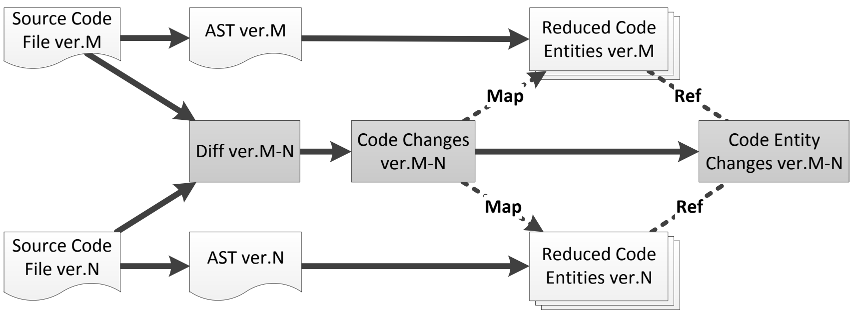



Figure 11
Phases of authorship determination based on extraction of source code entity changes

1) Extraction of source code files from a software solution. Files can be added into, moved within or deleted from the software solution. It is important to identify all source code files contained in the solution throughout its history.
2) Extraction of source code from all files valid for given versions. This is done repetitively for each two adjacent versions and supported by revision control systems. Some scenarios we consider as a problem are as follows: renaming the file, moving the file in a solution structure, creating a new file with the same name instead of the file previously removed.
3) To acquire a history of a source code means getting the history of the source code file(s). This operation is supported by revision control systems and managing the source code files. Some scenarios we consider as a problem are as follows: renaming the file, moving the file in a solution structure, creating a new file with the same name instead of the file previously removed.

4) Representation of each two adjacent versions as ASTs. Transformation to AST representation is restricted to level of meaningful syntactical units (classes, functions, properties). Lower level content is represented as lines.

5) Comparison of selected source code versions using Diff function based on solving the longest common subsequence problem [17]. The output of Diff function can be represented as a set of code changes, which can be of type Add, Delete or Modify. Each change holds information about the author and the range of affected lines: add in a newer version of source code, delete in older and modify in both versions.

6) Mapping of code differences to code entities and representation of source code entity changes. In this phase, we create source code entity changes, which are relations between source code elements and source code changes. Code change falls to code entity if the intersection of their ranges is not an empty set. One change can fall to:

   a) one code entity in new, old or both compared versions of source code

   b) to multiple code entities in one level of AST (for example two functions)

   c) to multiple code entities in multiple levels of AST (method, class, namespace)

   One code change can produce many code entity changes, so it is important to create relations (between change information unit, old and new version of a code entity) only between syntax units of the same type and on the same level of AST.

The following algorithm is based on previously described phases. The result is a set of changes related to the code elements. Based on this output, we can evaluate the authorship metrics.

```
ExtractCodeEntityChanges(IN: solutionPath, OUT: CodeEntityChanges[])
for filePath in ParseSolution(solutionPath)
  oldSrcCodeFile = null
  for newSrcCodeFile in GetHistory(filePath)
    ast = ParseAst(newSourceCodeFile)
    newCodeEntities = ReduceCodeEntities(ast.Root, empty)
    if oldSourceCodeFile is not null
      codeChanges[] = GetCodeChanges (oldSrcCodeFile, newSourceCodeFile)
      codeEntityChanges+=MapCodeEntityChange(oldCodeEntities,
newCodeEntities, codeChanges)
    oldSrcCodeFile = newSrcCodeFile
    oldCodeEntities = newCodeEntities
```

## 5.4 Authorship Metrics

We consider the following aspects in the process of authorship metrics evaluation. First, we calculate authorship for individual types of authors: real authors (as well as coauthors), editors, reviewers. The total authorship of an author can be evaluated as their weighted sum, where real author changes will have the heaviest weight and reviewer changes the lightest.

Next, authorship is based on the source code change types, where total authorship is weighted by the sum of the particular types: *Added* (weight 4), *Deleted* (1) and *Modified,* which in fact consists of an *old version* deleted (2) and a *new version* added (1). We can polemize which type of change tends better to the authorship of the code, and so weights for each type (in the brackets above) can be a part of the knowledge and can be calibrated. Code changes can be measured in lines of code or syntactic code entities, in absolute numbers or relative index.

Also, it is meaningful to consider and evaluate metrics depending on time in expression of change sets, where the oldest commit is ranked as least important and the latest commit as the most important.

We use these characteristics for visualization of source code entities authorship in *Tent graph*. Part of this graph is shown in Figure 12. Our colleagues from Vision & Graphics Group[1] analysed similar software visualization methods in 3D space [21] [25] without these additional features (authors and code users).

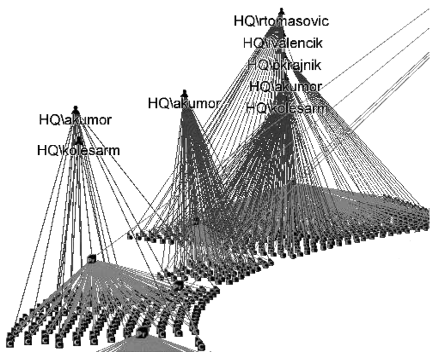



Figure 12
Authorship of source code entities presented in *Tent graph*

[1] vgg.fiit.stuba.sk

# 6 Information Tags for Knowledge Exchange

Information tags represent a way of binding knowledge with the particular source code. Tags can be created by automated machines or by users. We visualize both types of tags in an integrated development environment as well as in our own graph visualization engine. In this way the knowledge obtained through tags can be presented to programmers, managers and other members of the development team.

## 6.1 Tags Visualization

Tags are stored in a database and are referenced to source code entities. When a user opens a source code file in a development tool, the corresponding tags are loaded and shown as icons next to corresponding lines. Detailed information of each tag can be shown.

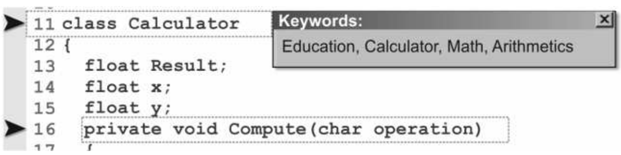



Figure 13
Displaying tags in a source code editor

The same plugin allows for creating new tags: a right mouse-click above selected lines (or above method header, class header, project file name etc.) defines an entity to be tagged, and context menu allows for creating its content. Tag content with timestamp and author's name will be stored in a database, where it is referenced to an existing source code entity.

## 6.2 Type of Tags

We defined the following types of tags:

- *note* – free text comment, annotation, remark, recommendation
- *links* - URL or file references to source, know-how, documentation, etc.
- *keywords* **-** categories, destination of code, etc.
- *features* - technical features like percentage of progress, etc.
- *rating, warnings* – good or bad rating, warning about mistake etc.
- *authors* **-** list of code authors, their kind of participation on code
- *history* - history of entity code updates

All types, except for the last two items, are user's marks, which may be created by programmers. A programmer can identify his own class with keywords, assign some feature to method and/or write some recommendation in foreign code.

Metadata like authors and history are read-only tags of each entity. It helps programmers to find who creates each class, who updates it mostly, or how the methods of some classes evolve.

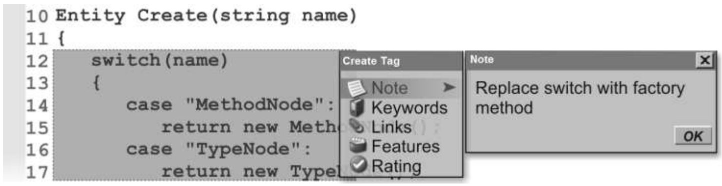



Figure 14

Creation of the note mark

## 6.3 Principle of Positioning

The key problem of tagging a source code is the dynamic changes of the code; therefore simply referencing tag positions to line numbers is unusable. We use referencing in Abstract Syntax Tree that remembers the identity of entities (like classes or methods) in historical file versions. This means that the system remembers that method x occupies lines (10-25) in version 1, lines (12-32) in version 2, etc.

When a tag is attached to whole method x, it will be shown next to the class header line in the actual version of the file.

If a tag is attached to a line range or a single line inside method x, the positioning is more complicated. The system remembers the file version in which the tag was created and the line position inside (in relation to) original entity. Thanks to the entity history, we know the position of method x in an actual file version (e.g. it occupies lines 12-32) and the relative position offset (e.g. lines 2-3). Therefore, tagged lines are 12-13 in the new file version. The system must verify if the lines (12-13) contain equivalent or fairly similar content as the corresponding line range in the original version. If so, the tag will be shown beside line 12, and if not, the tag will not be shown.

The tag's validity may be optionally limited to a code version range. For example, a link to class documentation has logically unlimited validity, but a warning note usually loses its sense when the code is corrected or changed.

## 6.4 Utilization of Code Tagging Data

The goal of code tagging is to save the programmers' time when they find patterns [13] [18], descriptions, or the same features of existing codes. Information accumulated in the marking databases allows versatile advanced usage. For example:

- *Searching* of projects, classes and methods *by keywords, features*, etc.:
  - Find existing methods for sending E-mail in existing source codes
  - Find all classes in projects with low rating
  - See the evolution of the class through historical versions
- *Comparing* projects, classes, methods *by features*
- *Classifying* projects, classes, methods *by keywords, feature,* etc.
- Visualization of *summaries*, like
  - *Good-rated / bad-rate* rated codes in projects
  - *Safe / unsafe or problematic* methods in projects
  - *Fast / slow* methods in projects
  - *Just developed / finished* codes in projects
  - *Documented / undocumented* methods in projects

**Future Work**

We plan to complete the existing environment with other methods, namely content recognition using multiagent systems [24] [19] [22], swarm intelligence [23], neural networks [20] especially Self-Organizing Maps, and pattern or anti-pattern matching.

In the tier recognition method we are currently using a simple direct method for word extraction from identifiers, but we are planning to apply also other and more sophisticated methods [9].

The presented algorithm for authorship determination does not solve the problem with proper identification of code entities through history in scenarios, where identity or position in AST of code elements has changed. To solve this deficiency, we must focus on methods for the determination of source code similarities.

**Acknowledgment**

This contribution is a partial result of the Research & Development Operational Programme for the project Research of methods for acquisition, analysis and personalized conveying of information and knowledge, ITMS 26240220039, co-funded by the ERDF.

## References

[1] Storey M., Li-Te Ch., Bull I., Rigby P.: Shared Waypoints and Social Tagging to Support Collaboration in Software Development, Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work (CSCW '06) ACM, New York, NY, USA, 195-198, 2006

[2] Guzzi A., Hattori L., Lanza M., Pinzger M.; van Deursen A.: Collective Code Bookmarks for Program Comprehension, Program Comprehension (ICPC) 2011 IEEE 19th International Conference on, Vol., No., pp. 101-110, 22-24 June 2011

[3] Peacock A., Xian Ke, Wilkerson M: Typing Patterns: a Key to User Identification, Security & Privacy, IEEE, Vol. 2, No. 5, pp. 40-47, Sept.-Oct. 2004

[4] Kot, B., Wuensche, B., Grundy, J., Hosking, J.: Information Visualisation Utilising 3D Computer Game Engines Case Study: a Source Code Comprehension Tool, In Proceedings of the 6th ACM SIGCHI New Zealand Chapter's International Conference on Computer-Human Interaction: Making CHI Natural (CHINZ '05) ACM, New York, NY, USA, 2005

[5] Sindhgatta, R.: Identifying Domain Expertise of Developers from Source Code, In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08) ACM, New York, NY, USA, 2008

[6] Kuhn, A., Ducasse, S., Gírba, T.: Semantic Clustering: Identifying Topics in Source Code, Information and Software Technology, Volume 49, Issue 3, March 2007

[7] Rahman, F., Devanbu, P.: Ownership, Experience and Defects: a Fine-grained Study of Authorship, In Proceedings of the 33rd International Conference on Software Engineering, ACM, 2011, pp. 491-500

[8] Bird C., Rigby C. P., Barr T. E., Hamilton J. D., German M. D., Devanbu , P.: The Promises and Perils of Mining Git, In Proceedings of the 2009 6th IEEE International Working Conference on Mining Software Repositories (MSR '09) IEEE Computer Society, 2009, pp. 1-10

[9] Enslen, E., Hill, E., Pollock, L., Vijay-Shanker, K.: Mining Source Code to Automatically Split Identifiers for Software Analysis, Mining Software Repositories, 2009, MSR '09, 6th IEEE International Working Conference, pp. 71-80, May 2009

[10] Fluri, B., Wursch, M. Pinzger, M., Gall H.: Change Distilling: Tree Differencing for Fine-grained Source Code Change Extraction, *IEEE Transactions on Software Engineering*, Vol. 33, No. 11 (2007) 725-743

[11] Object-Oriented Graphics Rendering Engine OGRE, www.ogre3d.org (Accessed on 19 August 2011)

[12] Fruchterman, T. M. J., & Reingold, E. M.: Graph Drawing by Force-Directed Placement. Software: Practice and Experience, 21(11) 1991

[13] Beck K.: Implementation Patterns, Addison-Wesley, 2007

[14] Neamtiu I., Foster J., Hicks M.: Understanding Source Code Evolution Using Abstract Syntax Tree Matching, MSR '05 Proceedings of the 2005 International Workshop on Mining Software Repositories, NYC (2005) ACM SIGSOFT Software Engineering Notes, Volume 30, Issue 4, July 2005, NYC

[15] Maletic, J. I., Collard, M. L., Marcus, A.: Source Code Files as Structured Documents, In: 10th IEEE International Workshop on Program Comprehension, Paris (2002)

[16] Sommerville, I: Software Engineering, 9th edition. Addison Wesley, 2011

[17] Eppstein, D.: Longest Common Subsequences. ICS 161: Design and Analysis of Algorithms Lecture notes for February 29, 1996, http://www.ics.uci.edu/~eppstein/161/960229.html (Accessed on 9 March 2012)

[18] Gamma E. et al.: Design Patterns: Elements of Reusable Object-oriented Software, Addison-Wesley, 1995

[19] Tomášek M.: Language for a Distributed System of Mobile Agents, Acta Polytechnica Hungarica, Vol. 8, No. 2, 2011

[20] Lovassy R., Kóczy L. T., Gál L.: Function Approximation Performance of FuzzyNeural Networks, Acta Polytechnica Hungarica, Vol. 6, No. 5, 2009

[21] Šperka M., Kapec P.: Interactive Visualization of Abstract Data, Science & Military, 1/2010, pp. 89-90

[22] Cerná A., Cerný J.: Position Identification of Moving Agents in Networks, Acta Polytechnica Hungarica, Vol. 7, No. 2, 2010

[23] Yilmaz A. E.: Swarm Behavior of the Electromagnetics Community as regards Using Swarm Intelligence in teir Research Studies, Acta Polytechnica Hungarica, Vol. 7, No. 2, 2010

[24] Kelemen J.: Agents from Functional-Computational Perspective, Acta Polytechnica Hungarica, Vol. 3, No. 4, 2006

[25] Kapec P.: Knowledge-based Software Representation, Querying and Visualization, Information Sciences and Technologies, ACM Slovakia, Vol. 3, No. 2, 2011

# Simulation Tools Evaluation using Theoretical Manufacturing Model

**Pavol Semanco, David Marton**

Faculty of Manufacturing Technologies with seat in Presov
Bayerova 1, 08001 Presov, Slovakia
pavol.semanco@tuke.sk; david.marton@tuke.sk

*Abstract: The simulation of the business and manufacturing processes plays an essential role in modern scientific research and decision-making in management. However, this simulation, especially its output performance measures, can be very tricky because these simulations depend on the calculations of particular simulation software. The objective of this paper is to confront the average values of performance indicators of the manufacturing simulation model in three well-known manufacturing-focused simulation tools. Thus, we applied advanced inferential statistical technique after normality test and the homogeneity of variances to analyze the output data of the model in three different simulation tools. We performed statistical analysis on the data of the average waiting time in queue and the number of completed parts obtained from 1500 replications together. The simulation model of this study employs a single-machine workstation driven by M/G/1 queue with FIFO queue discipline. The findings from the partial and overall results are offered in the conclusion and discussion part of the paper.*

*Keywords: simulation; technique; inferential statistics; queue; model; process*

# 1 Introduction

In these turbulent times, in order to remain competitive enterprises have to continually improve the underlying business and manufacturing processes and exploit the latest technologies for better performance. Simulation represents a very effective tool in production planning and decision-making, especially in small and medium enterprises (SMEs). R. E. Shannon in [1] defines simulation as “the process of designing a model of a real system and conducting experiments with this model for the purpose of understanding the behavior of the system and/or evaluating various strategies for the operation of the system.” Accordingly, simulation can be employed either to predict the effect of changes to the performance of an existing system or to compare two or more scenarios of a new system under a variety of circumstances. Current manufacturing and business discrete-event simulation gives us better accuracy using highly evolved simulation software with user interface enabling 3D modeling and visualization.

Despite all that, simulation is still faced with a limited understanding of its capabilities as regards the practice perspective. There is also the lack of know-how in best applying simulation within the SME user communities, which causes hindrance in the adoption process of simulation. The diversity of simulation software and the unclear boundaries of simulation types also cause difficulty in the utilization of simulation. Murphy and Perra in [2] helped clarify the unclear boundaries of simulation types by determining the key simulation functions for each simulation type. Due to application overlaps they developed the simulation type functionality matrix. Hlupic et al. [3] offered criteria that can be of practical use in evaluating and selecting the correct simulation software. To select the appropriate simulation software, it is further necessary to know whether the output performance indicators give us the correct values.

Our objective is to consider the output performance measures of three simulation tools. We used an advanced inferential statistical technique that is widely used to analyze the consistency of output data.

The rest of the paper is organized as follows. In Section, “Manufacturing simulation,” we shortly describe each of the tested discrete-event simulation tools, namely ARENA, SIMUL8 and WITNESS. Section 3 presents the inferential statistical techniques. In Section 4, “Computational experiment,” we introduce the manufacturing simulation model with the use of M/G/1 queue and experiment condition. Subsequently, there are also presented the results of the simulation runs and the statistical analysis of output data. Finally, partial and overall findings from experiment are summarized and discussed, and possible future research ideas are presented.

# 2 Manufacturing Simulation

Manufacturing simulation is considered one of the main application areas of simulation technology. It is able to reliably determine the system behavior for a set of pre-defined conditions. Simulation software involves the design and modeling of a large variety of aspects of the production facility, including layout, equipment selection, control strategies, material handling, buffer sizing, dispatching strategies, etc. Depending on the objectives, a simulation model can become very complex and data intensive. However, on the other side, the simulation software is just an analytical tool; and to optimize or rationalize a wide range of manufacturing systems requires expert intervention. Output performance measures (i.e. work in process, flowtime, total production, throughput, downtime, and machine utilization) are used by experts to evaluate the behavior of the system and to identify potential areas for possible improvement.

It should also be noted that fundamental to simulations is the method by which random numbers are generated. This method is often called the random number generator [4-10].

We begin by describing each simulation tool in a way that illustrates the important background features for our study. For the simulation software, we decided on the basis of availability. We also applied Hlupic's evaluation criteria to outline the main features of the simulation tools that we used in this study. In Table 1, we present the shortened version of evaluation framework for the simulation software products.

## 2.1 Arena

ARENA is built upon SIMAN, and in effect it is a high level graphical front end for SIMAN in which models are built by placing icons onto a drawing board and then linking these icons or blocks together to define the model logic [11]. ARENA provides 10 random number streams to select from, or one can accept the default stream. All of ARENA's distributions are based on a multiplicative congruential generator for uniformly distributed numbers between 0 and 1.

## 2.2 Simul8

SIMUL8 uses a modified multiplicative-congruential pseudo random number generator. There are 30,000 sets of random number streams available in SIMUL8. Since 2006, SIMUL8 supports replacement of its own generator with any coded in a dynamic link library (DLL) [12]. Programming in this simulation software product involves the five fundamental elements (work entry point, storage bin, work center, work complete and resource).

## 2.3 Witness

WITNESS is a discrete-event simulation package from Lanner Group that is mainly used to carry out the manufacturing simulation. It has an object-oriented modeling environment. The concept of WITNESS is based on the Queueing Theory. Programming itself involves the five most common elements (parts, machines, conveyors, buffers and labors). The random number generator of WITNESS uses 6 seeds (or streams) to generate random numbers. It also has the possibility to alter one or more of these seeds. WITNESS generates pseudo random numbers by using a combined multiple recursive generator. This method generates random numbers between 0.0 and 1.0. WITNESS then uses this number to sample from statistical distributions for activity timings, breakdown timings, setup intervals, PERCENT rules, etc. In order to calculate a different number each time, the pseudo random number generator uses the previous result to form the basis for calculating the next [13].

# 3 Computational Experiment

The experimental design and conditions are further detailed in terms of simulation model description, parameter settings, platform configuration, statistical analysis and results obtained. It is anticipated that despite the different random number generator, output performance measures from each simulation tool will be uniform. On the other hand, the simulation tool could compute average output performance indicators differently.

## 3.1 Simulation Model

In this section, we discuss the design of the simulation model, which is the same for all simulation software. A series of simulation experiments were carried out to study the diversity between simulation tools. The M/G/1 manufacturing queuing model is a terminating system that begins from an empty-and-idle state and ends after the simulation time passed. In other words, in the system there is no work in process when the next replication starts. We took each replication of the model to be exactly 20 minutes. We used 3 scenarios, where each scenario includes the same M/G/1 model in a different simulation tool. In Figure 1, the generic model and the three scenarios (Scenario 1 = Arena, Scenario 2 = Simul8, and Scenario 3=Witness) are depicted. The three scenarios were investigated via extensive simulation runs. For better evaluation and comparison of all above scenarios, it is assumed throughout the simulations that initially the buffer is empty and machine process is not initiated before the start of the simulation. Another way to say it: arrival of the first part in the system is at time $t = 0$.

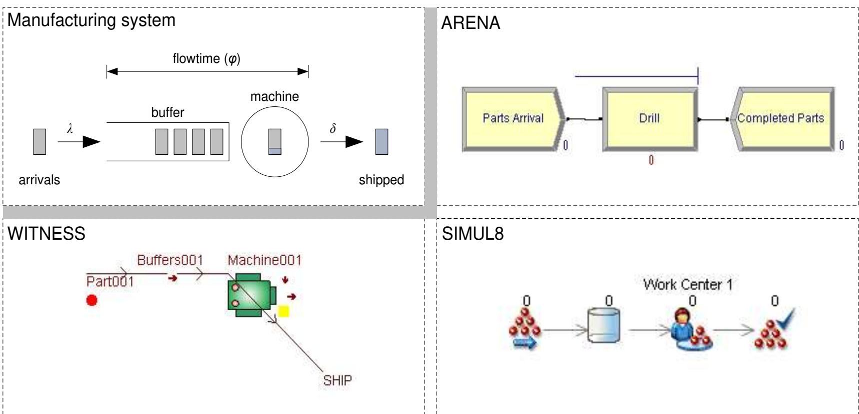



Figure 1
Simulation model with M/G/1 queue

For all three scenarios, we consider a shop with one machine and an unlimited buffer capacity for storing jobs, which arrive according to exponential distribution.

The buffer works according to a First-In-First-Out (FIFO) system. Job processing times are drawn from a triangular distribution. The input and output parameters of the simulation model are discussed next.

## 3.2 Parameters and Platforms

All the test simulation software products were run on a PC equipped with a 1.7 GHz Intel Celeron CPU and 1 GB of RAM. The simulation software products used to carry out the experiment were chosen based on availability: Witness 2008 (Full version release 1.00a, manufacturing performance edition), Arena 10.0 CPR 7 (student version) and Simul8 11.0 (student version).

The input parameters are the inter-arrival time, the processing time, the terminating condition and the number of replications. The inter-arrival time has an exponential distribution with a mean $\mu=1/\lambda=5$ minutes and the processing time has a triangular probability distribution with these parameters: *min*=1 minute, *mode* = 3 minutes, and *max*=6 minutes. The key parameters are described in Table 1.

Table 1

Overview of simulation model configuration and experimental factors

| **Factors** | |
|---|---|
| Number of machines | 1 |
| Inter-arrival time distribution | Exponential |
| Processing time distribution | Triangular |
| Queue discipline | FIFO |
| Simulation length (T) | 20 minutes |
| Number of replication (N) | 500 runs |
| Buffer capacity | infinite |
| *Experimental factors* | |
| Mean inter-arrival rate (λ) | 0.2 job per minute |
| Processing time | 1; 3; 6 minutes per job |

For ease in comparing the different scenarios and for a better performance evaluation of the system, each scenario was assessed using traditional output performance metrics, i.e. number of entered jobs, average waiting time in queue, average queue size, maximum queue size, average time in system, and total production.

## 3.3 Results

In order to gain better insights, the results from 1500 simulation runs together are presented in Table 2, where there are three numbers in brackets. The first number stands for the mean number, the second number refers to standard deviation, and the third is half the width of 95% confidence interval.

Table 2

Summary of the simulation for three scenarios (Arena, Simul8, and Witness)

| Performance measure | Arena* | Simul8* | Witness* |
|---|---|---|---|
| Number of entered jobs | (4.97;2.00;0.17) | (5.08;2.10;0.18) | (5.62;2.00;0.21) |
| Average waiting time in queue | (1.28;1.42;0.12) | (1.29;1.31;0.11) | (2.06;2.22;0.19) |
| Average queue size | (0.43;0.57;0.05) | (0.46;0.62;0.05) | (0.65;0.94;0.08) |
| Maximum queue size | (1.47;1.15;0.10) | (1.73;1.08;0.09) | (1.93;1.26;0.11) |
| Average time in system | (4.43;1.51;0.13) | (5.56;1.45;0.13) | (4.38;2.00;0.18) |
| Total production | (3.79;1.27;0.11) | (3.61;1.25;0.11) | (3.98;1.64;0.14) |

*(Mean; Standard deviation; Half width)

Figure 2 introduces the replication series of the three simulation tools for better illustration of the sample size and sample values. In the figure we display both performance indicators total production and average waiting time in buffer.

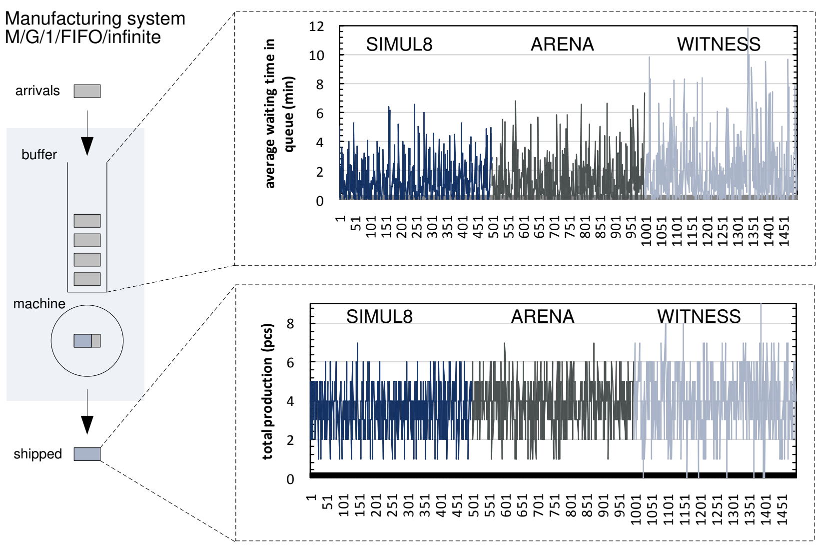



Figure 2

Overall views on simulation run in Arena, Simul8 and Witness

Due to the different random number generators used by the simulation tools, the output performance measures are slightly different. In order to determine if there is a significant difference among the performance measures obtained by these simulations, namely the average waiting time in queue and total production, a single-factor analysis of variance (ANOVA) was chosen. Like other parametric tests, the analysis of variance assumes that the data fit the normal distribution. To ensure this assumption, we did a normality test and also a test for homogeneity of variance. Table 3 presents results of the normality test from which we can conclude that three independent samples are non-normal distributed.

Table 3
Results of normality test using K-S and S-W tests

| | | **Kolmogorov-Smirnov** | | | **Shapiro-Wilk** | | |
|---|---|---|---|---|---|---|---|
| | **Simulation tool** | Statistic | df | p | Statistic | df | p |
| Average waiting time in queue (Minutes) | Arena | 0.184 | 500 | $<0.001$ | 0.824 | 500 | $<0.001$ |
| | Simul8 | 0.162 | 500 | $<0.001$ | 0.866 | 500 | $<0.001$ |
| | Witness | 0.177 | 500 | $<0.001$ | 0.827 | 500 | $<0.001$ |

In Figure 3 is depicted a histogram of non-normal distribution of 1500 simulation replications together with *mean*=1.54 minutes, *s*=1.735, and normal distributed curve. Figure 4 indicates the non-normal by graphically comparing observed values to expected normal values.

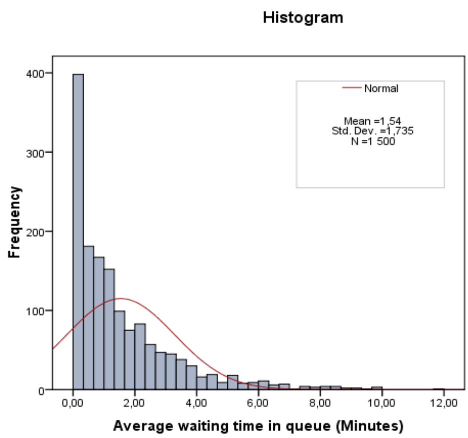



Figure 3
Data distribution (N=1500)


Figure 4
Normality test (N=1500)

We also conducted Levene's test for homogeneity of variances (see Table 4). Results from tests showed us that the obtained data from the simulations is non-normal, and the variances are different. Subsequently, we did a quick research and we found that the simulation studies, applying a variety of non-normal distributions, show that the false positive rate is not affected very much by this violation of the assumption, as in reference [14-16]. This is because a large number of random samples from a population is approximately normally distributed even when the population is not normal. However, we decided to perform a non-parametric Kruskal-Wallis test, which substitutes for the one-way ANOVA. This non-parametric test does not assume that the data fit the normal distribution. It does assume that the data obtained from three simulation tools have the same distribution as each other.

Table 4
Levene's test for homogeneity of variances (α=0.05)

| Average waiting time in queue (Minutes) | Levene Statistic | df1 | df2 | p |
|---|---|---|---|---|
| Based on Mean | 56.536 | 2 | 1497 | <0.001 |
| Based on Median | 33.045 | 2 | 1497 | <0.001 |
| Based on Median and with adjusted df | 33.045 | 2 | 1128.582 | <0.001 |
| Based on trimmed mean | 46.539 | 2 | 1497 | <0.001 |

Here we want to test a hypothesis that the results from all three simulation tools are equal. We state our hypothesis as:

$H_0$:There is no difference between the three simulation tools ($\mu_{Arena}$=$\mu_{Simul8}$= $\mu_{Witness}$).

$H_1$:At least one of the simulation tools gives different results due to generator number dissimilarity (not all $\mu$'s are equal).

The results are summarized and presented in Tables 5 and 6. At significance level $\alpha$=0.05, the final results show that there is significant difference between these outputs with respect to the average waiting time in queue. This seems to indicate that our preliminary expectations were not correct. According to our results, we do reject the null hypothesis. Another way to state this is: The difference between the three scenarios is statistically significant in this case.

Table 5

Summary of ranked data in each condition with Kruskal-Wallis test ($\alpha$=0.05)

| | **Simulation tool** | **N** | **Mean Rank** |
|---|---|---|---|
| Average waiting time in queue (Minutes) | Arena | 500 | 693.14 |
| | Simul8 | 500 | 715.81 |
| | Witness | 500 | 842.55 |
| | Total | 1500 | |
| **Test statistic** | Chi-Square | | 34.754 |
| | df | | 2 |
| | p-value | | <0.001 |

The Kruskal-Wallis test indicates a significant effect of a group so we rejected hypothesis $H_0$ and we applied a post-hoc test to determine which of the simulation tools is different. The Mann-Whitney test with Bonferroni correction compared all groups to each other. We compared Arena versus Simul8, Arena versus Witness, and Simul8versus Witness to identify the significance of the Kruskal-Wallis nonparametric test (see Table 6). We can conclude there is difference between Arena and Witness, and Simul8 and Witness.

Table 6

Test statistics for the Mann-Whitney test with Bonferroni correction ($\alpha$=0.0167)

| **Contrast** | **Average waiting time in queue (Minutes)** | | | |
|---|---|---|---|---|
| | Mann-Whitney U | Wilcoxon W | z | p-value |
| Arena vs. Simul8 | 120960.000 | 246210.000 | -0.888 | 0.375 |
| Arena vs. Witness | 100360.500 | 225610.500 | -5.410 | <0.001 |
| Simul8 vs. Witness | 103614.000 | 228864.000 | -4.694 | <0.001 |

Arena (*Mdn*=0.85) did not seem to differ in average waiting time in queue from Simul8 (*Mdn*=0.95), (*U*=120960, *p*=0.375, *r*=-0.03). The first comparison represents a very small effect because the effect size is close to zero. We can conclude that the effect on average waiting time in queue of Arena compared to Simul8 was negligible. On the other hand, Witness (*Mdn*=1.26) was significantly different from Arena (*U*=100360.5, *p*<0.001, *r*=-0.17) and Simul8 (*U*=103614,

$p$<0.001, $r$=-0.15). Because there was no significant difference between Arena and Simul8, we can assume that the random number generators of both simulation tools give the same random numbers. The small effect of Witness could lie in the equation (formula) for calculating the average performance indicators, which is different from Arena and Simul8. The average indicator in Witness counts all parts in the queue (or system) even when they are still in the queue (or in machine) at the end of the simulation. Arena and Simul8 determine the average waiting time indicators based on the parts that leave the queue (or system) at the end of the simulation run. Thus we can state that Witness determines the average performance indicator from an arrivals point of view, and Arena and Simul8 determine their average performance indicators from a departures point of view.

In this section, we investigated the congruence of the three different simulation software products based on their performance indicator, namely on the average waiting time in queue. We also identify the source of diversity between Witness and the other two simulation tools.

**Conclusions**

A simulation study was undertaken to investigate the effect of random number generators on performance indicators. The simulation study was applied to the M/G/1 manufacturing queuing model. The average waiting time in queue was used as the performance measure to explore the possible diversity between among the selected simulation tools.

After several simulation runs, we used the Kruskal-Wallis test, which revealed a significant effect of a group on value ($H(3)$=34.75, $p$<0.0001). A post-hoc test using the Mann-Whitney test was used to follow up the findings. The Bonferroni correction was also applied, and therefore all effects are reported at a 0.0167 level of significance. It appeared that the average waiting time in queue for Arena compared to Simul8 was no different ($U$=120960, $r$=-0.03). However, when Witness's performance indicator was compared to Arena and Simul8, average waiting times were significantly different with small effect size.

The distinction between Arena and Witness, and Simul8 and Witness, lies in the addition of a formula that counts the average waiting time in the queue. Non-significance between Arena and Simul8 is due to the fact that they use a similar random number generator. Our findings on the three simulation tools also indicate that although they use a slightly different random generator, their results should be directed towards the same output. The size of the gains is significantly influenced by the formula for average waiting time.

Thus far, the analysis indicates that Witness has significantly different average waiting time than Arena and Simul8. This is due to the different equation formulae used in those simulation tools. The Witness simulation software includes in average waiting time also for parts (or entity) that remained in queue after terminating a simulation run (which seems to be more practical in manufacturing planning and scheduling). Both Arena and Simul8 simulation software only count dispatched parts. This finding is also important from a practical perspective, because not knowing the behavior of performance indicators can lead to the wrong conclusion. This may occur by extending manufacturing lead time and increasing the cost of storing WIP, which was not considered in the beginning. This paper may be helpful for managers in selecting the appropriate simulation tool.

We hope that use of a simulation as presented will offer useful preliminary results for studying the throughput of a manufacturing system under various experimental conditions. We also proposed to add new criterion into evaluation framework of simulation software. This criterion is concerned with the type of the formula for average performance indicators (measures). Future research could address an approach to more difficult models, including the mean time between failures (MTBF) and the mean time to repair (MTTR).

**References**

[1] Shannon, R. E. (1998) Introduction to the Art and Science of Simulation. IEEE, USA

[2] Murphy, C. A., Perera T. (2002) The Definition of Simulation and Its Role within an Aerospace Company. Simul Practice and Theory 9, 273-291

[3] Hlupic, V., Irani, Z., Paul, R. J. (1999) Evaluation Framework for Simulation Software. International Journal of Advanced Manufacturing Technology 15, 366-382

[4] Sadiku, M. N. O., Ilyas, M. (1991) Simulation of Local Area Networks, CRC Press

[5] L'Ecuyer, P., Simard, R., Chen, E. J., Kelton, W. D. (2002) An Object-Oriented Random-Number Package with Many Long Streams and Substreams. Operational Research 50, 1073-1075

[6] Pusatli, O. T., Sanjay, M. (2011) Software Measurement Activities in Small and Medium Enterprises: an Empirical Assessment. Acta Polytechnica Hungarica 8(5), 21-42

[7] Gentle, E. J. (2003) Random Number Generation and Monte Carlo Methods. Springer, New York

[8] McHaney, R. (1991) Computer Simulation: A Practical Perspective. Academic Press

[9] Fishman, G. S. (1990) Multiplicative Congruential Random Number Generators with Modulus 2beta: An Exhaustive Analysis for Beta = 32 and a Partial Analysis for Beta = 48. Mathematics of Computation 54, 331-344

[10] Sanjay, M. (2011) Evaluation Criteria for Object-oriented Metrics. Acta Polytechnica Hungarica 8(5), 109-136

[11] Kelton, W. D., Sadowski, P. R., Sturock, T. D. (2007) Simulation with Arena. McGraw-Hill, New York

[12] Hauge, J. W., Paige, K. N. (2002) Learning SIMUL8: The Complete Guide. Plain Vu Publishers, Bethlingham

[13] Lanner Group (2000) WITNESS 2000 User Manual. Lanner Group, Oaks, UK

[14] Glass, G. V., Peckham, P. D., Sanders, J. R. (1972) Consequences of Failure to Meet Assumptions Underlying Fixed Effects Analyses of Variance and Covariance. Review of Educational Research 42, 237-288

[15] Harwell, M. R., Rubinstein, E. N., Hayes, W. S., Olds, C. C. (1992) Summarizing Monte Carlo Results in Methodological Research: the One- and Two-Factor Fixed Effects ANOVA cases. Joutrnal of Educational Statistics 17, 315-339

[16] Lix, L. M., Keselman, J. C., Keselman, H. J. (1996) Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance F test. Review of Educational Research 66, 579-619

# Investigating the Influence of Knowledge Management Practices on Organizational Performance: An Empirical Study

**Mohamad H. Gholami[1], Mehrdad Nazari Asli[2], Salman Nazari-Shirkouhi[3*], Ali Noruzy[4]**

[1] Department of Industrial Engineering & Management Systems, Amirkabir University of Technology, Tehran, Iran, P.O. Box: 15977-37615
E-mail: fgholami@aut.ac.ir

[2] Department of Management, Imam Khomeini International University, Qazvin, Iran, P.O. Box: 34149-16818, E-mail: akimiaee@mohaddes.ac.ir

[3,*] Young Researchers Club, Roudbar Branch, Islamic Azad University, Roudbar, Iran, P.O. Box 14579-44616, E-mail: snnazari@ut.ac.ir; corresponding author

[4] Department of Educational Administration, Faculty of Psychology and Education, University of Tehran, Tehran, Iran, P.O. Box 11365-4563
E-mail: ali.noruzy@ut.ac.ir

*Abstract: The aim of this study is to investigate the influence of knowledge management practices on organizational performance in small and medium enterprises (SMEs) using structural equation modeling (SEM). A number of 282 senior managers from these enterprises were chosen using simple random sampling and the data were analyzed with the structural equation model. The results showed that knowledge acquisition, knowledge storage, knowledge creation, knowledge sharing, and knowledge implementation have significant factor loading on knowledge management; and also productivity, financial performance, staff performance, innovation, work relationships, and customer satisfaction have significant factor loading on organizational performance. Finally, the results of this study suggest that knowledge management practices directly influence the organizational performance of SMEs.*

*Keywords: knowledge management; organizational performance; SMEs*

## 1 Introduction

Today's necessity of information and development of knowledge in the organizations and the acceptance of the human resources managers' role as knowledge managers has become the vision of the organizations which are interested in keeping their competitive advantage. KM helps the SMEs have a proper understanding of and insight into their internal experiences and external resources (customers, suppliers, and competitors). KM activities, including knowledge acquisition, knowledge storage, knowledge creation, knowledge sharing and knowledge implementation can help the SMEs achieve necessary capabilities, such as problem solving, dynamic learning, strategic planning, decision-making, and improving their organizational performance as a whole[1]. The main goal of KM is the rapid, effective and innovative utilization of the resources and knowledge assets, infrastructures, processes and technologies in order to promote organizational performance [2].

Many researchers have attempted to explain why certain firms behave better than others by linking different organizational elements with performance measures. These studies include linking performance with strategy, structure, environment, learning capabilities, market orientation, resources, and employees' abilities [3]. As KM involves valuable processes which can influence the productivity, financial performance, staff performance, innovation, work relationships and customer satisfaction and finally organizational performance, studying the influence of KM practices on organizational performance in SMEs is important [4]. However, studying KM practices in the SMEs has not been sufficiently considered in literature, and limited studies have been conducted o investigate the effect of KM practices on their organizational performance. SMEs can achieve a higher degree of productivity, innovation, efficiency, customer satisfaction and competitive advantage with the use of KM practices, with the result finally of in improvement in organizational performance [5].

# 2 Knowledge Management Practices and Organizational Performance

KM practices means the process of acquiring, storing, understanding, sharing, implementing knowledge, and these actions are taken in the organizational learning process with regard to the culture and strategies of the organizations [6]. On the other hand, Bhatti and Qureshi [7] stated that KM means efforts to explore the tacit and explicit knowledge of individuals, groups, and organizations and to convert this treasure into organizational assets so that individuals and managers can use it in various levels of decision making. KM is a systematic and integrated management strategy that develops, transfers, transmits, stores, and implements knowledge so that it can improve efficiency and effectiveness of the organization's manpower [8].

The relevant theory that helps significantly towards realizing the important role of knowledge management is the knowledge-based theory. This theory supposes that knowledge management practices such as knowledge acquisition, knowledge storage, knowledge creation, knowledge sharing and knowledge implementation play a critical role in achieving high level productivity, financial and human resource performance and finally improving sustainable competitive advantage [9, 10].

In order for SMEs to be more successful and survive in a competitive market, they need to consider adaptive and intelligent strategies, including KM processes and best practices [11, 12].Some scholars have developed conceptual models based on knowledge-based theory which contain critical KM practices. On the other hand, KM practices could be defined in various forms and utilized in different configuration. For example, the life cycle model by Nissen et al. [13] divided a knowledge flow into six phases. These six phases are creation, organization, formalization, distribution, application or implementation, and evolution. Moreover, Wiig et al. [14] described KM as including eight practices: reviewing, analyzing the KM processes, analyzing the application risks, executing the proposed plans, developing knowledge, consolidating knowledge, sharing knowledge, and combining knowledge.

As discussed in the research background, different models are considered to describe KM practices in various ways. In this research, five main practices are adapted from the models of Nonaka et al. [15], Dahiya et al. [8] and Bhatti and Qureshi [7]. These practices encompass knowledge creation, acquisition, sharing, storage, and implementation, which have been frequently applied in evaluation of KM practices.

*Knowledge creation*: Knowledge creation involves the utilization of internal and external resources of an organization to generate new knowledge for achieving the organizational goals. Brainstorming methods and conducting research to make the best use of the knowledge assets of customers, suppliers and staffs are strategies applied in many prosperous SMEs for creating knowledge [16].

*Knowledge Acquisition*: This practice encompass the process of acquiring and learning appropriate knowledge from various internal and external resources, such as experiences, experts, relevant documents, plans and so forth. Interviewing, laddering, process mapping, concept mapping, observing, educating and training are the most familiar techniques for knowledge acquisition.

*Knowledge sharing*: Knowledge sharing is a process through which personal and organizational knowledge is exchanged. In the other words, knowledge sharing refers to the process by which knowledge is conveyed from one person to another, from persons to groups, or from one organization to other organization [17].

*Knowledge storage*: Knowledge storage involves both the soft or hard style recording and retention of both individual and organizational knowledge in a way so as to be easily retrieved. Knowledge storage utilizes technical systems such as modern informational hardware and software and human processes to identify the knowledge in an organization, then to code and index the knowledge for later retrieval [18]. In the other words, organizing and retrieving organizational knowledge means knowledge storage by providing the ability to retrieve and use the information by the individuals.

*Knowledge implementation*: This means the application of knowledge and the use of the existing knowledge for decision-making, improving performance and achieving goals. Organizational knowledge should be implemented in the services, processes and products of the organization.

Organizational performance is one of the most important structures discussed in management research and could be considered as the most important criterion for testing the success of SMEs. Performance is one of the most critical areas of SME management, which many management scholars and practitioners have focused on improving via strategic variables such as KM practices [3]. Past studies have conceptualized firms' performance with measures of return on assets, sales growth, new product success [19], market share and overall performance [20], sales growth, market share and profitability [21], overall performance, new product success, change in relative market share [22], profitability, sales growth, and overall customer satisfaction [23]. In this field, Venkatraman and Ramanujam [24] found that financial measures (return on equity, return on investment) and operational measures (market share, sales growth, and, profit growth) were frequently employed to measure organizational performance. On the other hand, there is no full consensus among academic researchers on the variables and indices of organizational performance. In the other words, organizational performance indices are different in SMEs. Researchers have considered different indices for the assessment of the performance of SMEs. For example, Johnsen and McMahon [25] considered return on assets, return on shareholders' salary, and return on investment and dividend as the performance indices of SMEs. Koh et al. [26] used three criteria to measure organizational performance, including organizational effectiveness (the relative quality of the products, success in the provision of new products, and the ability of the organization to keep the customers), share and growth of market (sales levels, sales growth, and relative market share), and profitability (capital return rate and profit margin). In addition, Huang [27] considered the indices of effectiveness, efficiency, productivity, life quality, innovation, and profitability for measuring the organizational performance of SMEs. Since the considered indices for measuring performance of the SMEs are different, some of the most important indices applied in previous researches have been selected for this study. The indices which are considered here for measuring the performance of these enterprises include productivity, financial performance, staff performance, innovation, work relationships and customer satisfaction.

Customer satisfaction is an essential component for the survival of the firm, and firms that are responsive to changes in customer needs, requirements and wants are expected to achieve a sustainable competitive advantage [28]. Additionally, innovation can be considered as a critical factor in achieving high performance.

Innovation is about using technology and knowledge to offer customers a new product or service via improved features or lower prices [29].

These six indices are of the highest importance in measuring the performance of SMEs, and few studies have been done on effect of KM activities on organizational performance [30]. However, a few researchers have been able to identify KM practices and relate them to the firm's performance. Some research indicates that firms which use suitable KM practices might enhance their capabilities, which may in turn result in better firm performance [1, 31, 32, 33].

Roland [34] in his study indicated that performance depends on a firm's ability to integrate knowledge into the value creation process and into core competency based strategies. Furthermore, his findings revealed that to achieve and maintain a high level of performance, an organization must develop efficient mechanisms for creating, transferring, and integrating knowledge. Also, Noruzy et al. [35] conducted research to study the influence of KM on organizational performance among 106 companies. Their results showed that KM positively influences the organizational performance of manufacturing firms. Moreover, Garcia-Morales et al. [30] suggested that strategic variables of knowledge (knowledge slack, absorptive capacity, tacitness) play a positive mediating role between transformational leadership and organizational performance.

Furthermore, Zack et al. [1] revealed that KM practices have a positive and indirect influence on financial performance. Also, Kiessling et al. [6] suggested that KM positively affects organizational outcomes, such as the firm's innovation, product improvement and employee improvement. Other research has indicated that KM positively and directly influences the SMEs' organizational performance [5, 36, 37] (Fugate et al., 2009; Chen and Liang, 2011; Hussain et al., 2011). According to the reviewed literature, we propose the following hypothesis:

***Hypothesis:*** *Knowledge management practices positively influence organizational performance.*

# 3 Methods

## 3.1 Sample and Procedure

The statistical population of this study consists of SMEs in Iran within the food industries, auto industries, textile industries, pipe and faucet industries, electronic industries, and clothing industries. Three hundred and eighty questionnaires were distributed among the sample group. All of the questionnaires were distributed during one month. A preliminary survey instrument was pre-tested by 30 senior managers and the reliability of the instrument estimated by using Cronbach's Alpha. Results with Cronbach's Alpha showed that the instruments had acceptable reliability (more than 0.7). Finally, a total of 282 completed questionnaires were obtained.

## 3.2 Instruments

*Knowledge management practices:* The knowledge management practices instrument was adapted from Cho [3], Chen and Huang [38], Chen and Liang [36], and Fugate et al. [5]. This questionnaire consists of five components included: knowledge acquisition, knowledge storage, knowledge creation, knowledge sharing and knowledge implementation. A five point Likert scale was used to measure these components (strongly disagree=1, to strongly agree=5).The validity and reliability was confirmed using a confirmatory factor analysis ($\chi 2$/df=1.26, RMSEA= 0.036, NFI=0.91, NNFI=0.90, CFI=90) and Cronbach's Alpha ($\alpha$=0.81, 0.78, 0.73, 0.84 and 0.85 for each components). Results showed that our scale has high validity and reliability.

*Organizational performance:* For measuring the organizational performance we developed a scale by adapting some items from previous studies, such as Cho et al. [39]; Chen and Liang [36] and Fugate et al. [5]. This scale consists of five components included: productivity, financial performance, staff performance, innovation, work relationships and customer satisfaction. A five point Likert scale was used to measure these components (strongly disagree=1, to strongly agree=5).To examine its validity and reliability we performed confirmatory factor analysis ($\chi 2$/df=1.3, RMSEA= 0.069, NFI=0.93, NNFI=0.90, CFI=90) and Cronbach's Alpha ($\alpha$=0.71, 0.75, 0.82, 0.80, 0.69 and 0.85 for each components respectively).The results showed that this scale has high validity and reliability for measuring organizational performance.

# 4 Findings

Structural equation modeling (SEM) was conducted to estimate the fitness of the model, and to perform the SEM analysis the LISREL 8.30 program was used. The most practical indices were used to estimate the model fitness, including:$\chi 2$/df, Root Mean Square Error of Approximation (RMSEA), Goodness of Fit Index (GFI), Comparative Fit Index (CFI) and Adjusted Goodness of Fit Index (AGFI).Scores lower than 5 for the $\chi 2$/df index reveals an acceptable rate; in other words, smaller scores in this index indicate a better fitness of the model [40, 41]. In this study, the $\chi 2$/df was 1.051, which attests to the appropriate fitness of the model. An RMSEA equal to or lower than 0.05 is suitable for tested models, but scores above 0.05 and up to 0.08 propose an agreeable error of approximation in the model. Models with their RMSEA at 0.10 and higher are considered to have low fitness. GFI and AGFI show to what degree the model has better fitness when compared to the model's non-existence. For the model to be acceptable, GFI, AGFI and CFI should be equal to or higher than 0.90 [40].

Figure 1 shows the factor loading of KM components (knowledge acquisition, knowledge storage, knowledge creation, and knowledge sharing and knowledge implementation) and organizational performance components (productivity, financial performance, staff performance, innovation, work relationships and customer satisfaction). As this figure shows, KM practices in SMEs significantly and positively influenced organizational performance (β= 0.41).

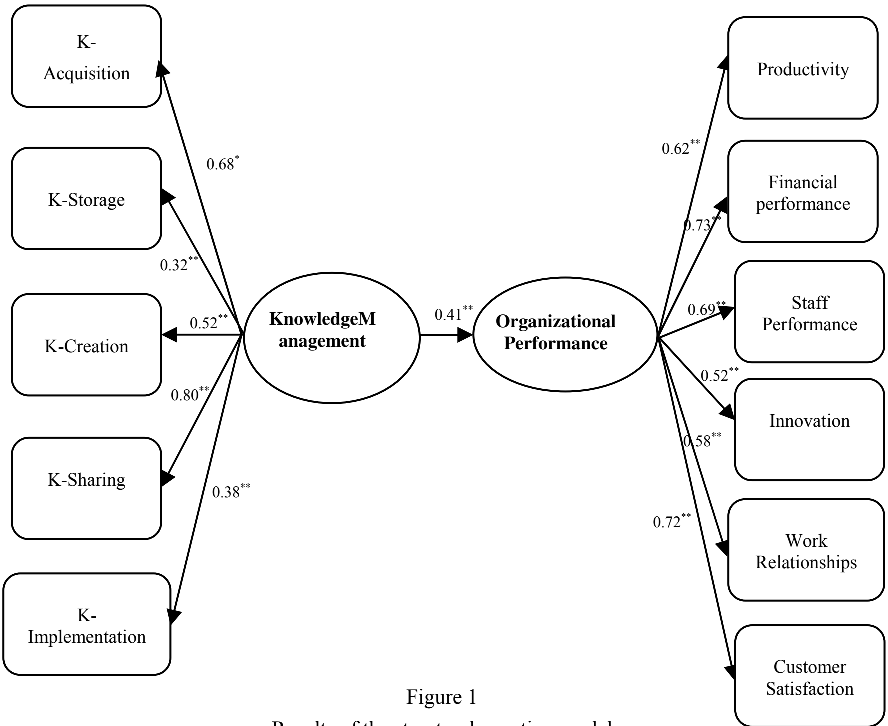



Figure 1
Results of the structural equation model

The results of Table 1 indicate scores of factor loadings and explain variances for KM practices and organizational performance. Factor loadings for knowledge acquisition, knowledge storage, knowledge creation, knowledge sharing and knowledge implementation are 0.68, 0.32, 0.52, 0.80, and 0.38, respectively. These factor loadings are statistically significant ($p<0.01$). Moreover factor loadings for productivity, financial performance, staff performance, innovation, work relationships and customer satisfaction are 0.62, 0.73, 0.69, 0.52, 0.58 and 0.72, respectively, while the explained variances for these variables are 0.46, 0.53, 0.47, 0.27, 0.33 and 0.51. These factor loadings are statistically significant at $P<0.01$.

Table1

Factor loadings and estimated common variance of the variables

| Variable | Knowledge Management | | Organizational Performance | |
|---|---|---|---|---|
| | Factor Loading | Explained Variance | Factor Loading | Explained Variance |
| Knowledge Acquisition | 0.68 | 0.46 | | |
| Knowledge Storage | 0.32 | 0.10 | | |
| Knowledge Creation | 0.52 | 0.27 | | |
| Knowledge Sharing | 0.80 | 0.64 | | |
| Knowledge Implementation | 0.38 | 0.14 | | |
| Productivity | | | 0.68 | 0.46 |
| Financial performance | | | 0.73 | 0.53 |
| Staff Performance | | | 0.69 | 0.47 |
| Innovation | | | 0.52 | 0.27 |
| Work Relationships | | | 0.58 | 0.33 |
| Customer Satisfaction | | | 0.72 | 0.51 |

Moreover, KM practices in SMEs positively influenced organizational performance ($\beta$= 0.41), the $\beta$ of which is statistically significant ($p<0.01$). Table 2 shows the goodness of fit indices for the estimated model. As can be seen, the $\chi2$/df score is 2.1, which lies within the acceptable range. Additionally, the RMSEA score is 0.064, which is lower than 0.08, and is thus within the acceptable range. The scores for AGFI, GFI and CFI, which are 0.90, 0.91 and 0.91, respectively, reveal the acceptability of these scores for the model. In total, one can say the examined model has an appropriate fitness.

Table 2

Goodness of fit indices for the second-rank model

| Model | $\chi^2$/df | RMSEA | GFI | AGFI | CFI |
|---|---|---|---|---|---|
| Scores | 2.1 | 0.064 | 0.91 | 0.90 | 0.91 |

## Discussion and Conclusion

As discussed in previous sections, KM encompasses knowledge acquisition, knowledge storage, knowledge creation, knowledge sharing, and knowledge implementation, and organizational performance includes critical components such as productivity, financial performance, staff performance, innovation, work relationships, and customer satisfaction. By considering these components, the research model has been conceptualized and operationalized among SMEs. Results showed that knowledge sharing has higher factor loading compared with other KM practices, and financial performance has higher factor compared with other organizational performance components. Other results showed that the SMEs' KM practices positively and significantly influenced their organizational performance. Generally, based on our findings, we can say that the improvement of KM practices can play a significant role in improving productivity, financial performance, staff performance, innovation, work relationships, and customer satisfaction, and thus in improving the SMEs' organizational performance [25, 26]. Moreover, the conclusions of this research suggest that KM practices are the critical elements for promoting the performance of SMEs.

When knowledge is recognized, acquired, and stored, SMEs can implement this knowledge to explore problems and create solutions, producing a structure for facilitating efficiency and effectiveness. In the modern dynamic and complex environment, SMEs need to acquire, create, share, save and implement new knowledge in order to make strategic decisions that can lead to improvements in productivity, financial and staff performance, innovation, work relationships, and customer satisfaction. Thus, SME managers should be committed to providing a supportive climate and culture, one that motivates employees and supervisors to implement the mentioned KM practices, in order to foster the SMEs results.

## Managerial Implications

This research makes a contribution by providing SMEs with better insights into KM practices, including knowledge acquisition, storage, creation, sharing, and implementation, in order to improve organizational performance. Further, by linking these issues to performance, this study demonstrates the importance of KM for better firm performance. Moreover, SME managers should perceive the benefits of KM practices that can increase productivity, financial performance, staff performance, innovation, work relationships, and customer satisfaction. KM leadership in SMEs should invest in internal and external resources in employing of an appropriate knowledge. Therefore, improved performance can be one of the long-term and strategic benefits of fulfilling KM best practices. SME managers should properly change the workplace culture and environmental circumstances so that employees adopt, support, commit to, and employ KM practices in fulfilling their activities.

SMEs could easily collect information from their customers, suppliers and other stakeholders, organize the collected knowledge through modern informational technologies or even traditional means, share the organized knowledge throughout all organizational levels, and finally implement the shared knowledge to overcome challenges and improve performance. Therefore, integrating KM practices as a strategic element is one of the most important tasks of SME managers.

## References

[1] Zack, M., McKeen, J., and Singh, S. (2009) Knowledge Management and Organizational Performance: an Exploratory Analysis. Journal of Knowledge Management, 13(6), 392-409

[2] Darroch, J. (2005) Knowledge Management, Innovation, and Firm Performance. Journal of Knowledge Management, 9(3), 101-115

[3] Cho, J. K. J. (2001) Firm Performance in the e-commerce Market: the Role of Logistics Capabilities and Logistics Outsourcing. University of Arkansas

[4] Chang, S., and Lee, M. (2008) The Linkage between Knowledge Accumulation Capability and Organizational Innovation. Journal of Knowledge Management, 12(1), 3-20

[5] Fugate, B. S., Stank, T. P., & Mentzer, J. T. (2009) Linking Improved Knowledge Management to Operational and Organizational Performance. Journal of Operations Management, 27(3), 247-264

[6] Kiessling, T. S., Richey, R. G., Meng, J., and Dabic, M. (2009) Exploring Knowledge Management to Organizational Performance Outcomes in a Transitional Economy. Journal of World Business, 44(4), 421-433

[7] Bhatti, K. K., and Qureshi, T. M. (2007) Impact of Employee Participation on Job Satisfaction, Employee Commitment and Employee Productivity. International Review of Business Research Papers, 3(2), 54-68

[8] Dahiya, D., Gupta, M., and Jain, P. (2012) Enterprise Knowledge Management System: A Multi Agent Perspective. Information Systems, Technology and Management, 285(4), 271-281

[9] Soderberg, A. M., and Holden, N. (2002) Rethinking Cross Cultural Management in a Globalizing Business World. International Journal of Cross Cultural Management, 2(1), 103-121

[10] Spender, J. C. (1996) Making Knowledge the Basis of a Dynamic Theory of the Firm. Strategic Management Journal, 17(Special Issue), 45-62

[11] Davenport, T., and Prusak, L. (2000) Working Knowledge: How Organizations Manage What They Know. Boston: Harvard Business School Press

[12] Tiwana, A. (2003) The Knowledge Management Toolkit: Orchestrating IT, Strategy, and Knowledge Platforms (2nd ed.) Upper Saddle River, NJ: Prentice-Hall

[13] Nissen, M., Kamel, M. N., and Sengupta, K. C. (2000) Integrated Analysis and Design of Knowledge Systems and Processes. Information Resources Management Journal, 13(1), 24-43

[14] Wiig, K. M., DeHoog, R., and Vander Spek, R. (1997) Supporting Knowledge Management: A Selection of Methods and Techniques. Expert Systems with Applications, 13(1), 15-27

[15] Nonaka, I., Umemoto, K., and Senoo, D. (1996) From Information Processing to Knowledge Creation: a Paradigm Shift in Business Management. Technology in Society, 18(2), 203-218

[16] Moodysson, J. (2008) Principles and Practices of Knowledge Creation: On the Organization of "Buzz" and "Pipelines" in Life Science Communities. Economic Geography, 84(4), 449-469

[17] Frappaolo, C. (2006) Knowledge Management. West Sussex, England: Capstone Publishing

[18] Karadsheh, L., Mansour, E., Alhawari, S., Azar, G., and El-Bathy, N. (2009) A Theoretical Framework for Knowledge Management Process:

towards Improving Knowledge Performance. Journal of Communications of the IBIMA, 7(1), 67-79

[19] Narver, J. C., and Slater, S. F. (1990) The Effect of a Market Orientation on Business Profitability. The Journal of Marketing, 54(3), 20-35

[20] Jaworski, B. J., and Kohli, A. K. (1993) Market Orientation: Antecedents and Consequences. The Journal of marketing, 75(3), 53-70

[21] Deshpande, R., Farley, J. U., & Webster Jr, F. E. (1993) Corporate Culture, Customer Orientation, and Innovativeness in Japanese Firms: a Quadrad Analysis. The Journal of Marketing, 57(1), 23-37

[22] Baker, W. E., and Sinkula, J. M. (1999) The Synergistic Effect of Market Orientation and Learning Orientation on Organizational Performance. Journal of the Academy of Marketing Science, 27(4), 411-427

[23] Ellinger, A. E., Daugherty, P., & Keller, S. (2000) The Relationship between Marketing/Logistics Interdepartmental Integration and Performance in U.S. Manufacturing Firms: An Empirical Study. Journal of Business Logistics, 21(1), 112-128

[24] Venkatraman. N., and VasudevanRamanujam (1986) Measurement of Business Performance in Strategy Research: A Comparison of Approaches. Academy of Management Review, 11(4), 60-75

[25] Johnsen, G. J., and McMahon, R. G. P. (2005) Owner-Manager Gender, Financial Performance and Business Growth Amongst SMEs from Australia's Business Longitudinal Survey. International Small Business Journal, 23(2), 115-142

[26] Koh, S. C. L., Demirbag, M., Bayraktar, E., Tatoglu, E., and Zaim, S. (2007) The Impact of Supply Chain Management Practices on Performance of SMEs. Industrial Management & Data Systems, 107(1), 103-124

[27] Huang, T. C. (2001) The Relation of Training Practices and Organizational Performance in Small and Medium Size Enterprises. Education+ Training, 43(8/9), 437-444

[28] Griffa, A. F. (2008) A paradigm Shift for Inspection: Complementing Traditional CMM with DSSP Innovation. Sensor Review, 28(4), 334-341

[29] Halliday, S. V. (2008) The Power of Myth in Impeding Service Innovation. Journal of Management Inquiry, 17(1), 44-55

[30] Garcia-Morales, V. J., Llorens-Montes, F. J., and Verdu-Jover, A. J. (2008) The Effects of Transformational Leadership on Organizational Performance through Knowledge and Innovation. British Journal of Management, 19(4), 299-319

[31] Lee, H., and Choi, B. (2003) Knowledge Management Enablers, Processes, and Organizational Performance: An Integrative View and Empirical Examination. Journal of Management Information Systems, 20(1), 179-228

[32] Gloet, M., and Terziovski, M. (2004) Exploring the Relationship between Knowledge Management Practices and Innovation Performance. Journal of Manufacturing Technology Management, 15(5), 402-409

[33] Marques, D. P., and Simon, F. J. G. (2006) The Effect of Knowledge Management Practices on Firm Performance. Journal of Knowledge Management, 10(3), 143-156

[34] Roland, N. (2006) Knowledge Management in the Business-driven Action Learning Process. Journal of Management Development, 25(9), 896-907

[35] Noruzy, A., Dalfard, V., Azhdari, B., Nazari-Shirkouhi, S., and Rezazadeh, A. (2012) Relations between Transformational Leadership, Organizational Learning, Knowledge Management, Organizational Innovation, and Organizational Performance: an Empirical Investigation of Manufacturing Firms. The International Journal of Advanced Manufacturing Technology, in press, 1-13

[36] Chen, D. N., and Liang, T. P. (2011) Knowledge Evolution Strategies and Organizational Performance: A Strategic Fit Analysis. Electronic Commerce Research and Applications, 10(1), 75-84

[37] Hussain, I., Xiaoyu, Y. U., Si, L. W. S., and Ahmed, S. (2011) Organizational Knowledge Management Capabilities and Knowledge Management Success (KMS) in Small and Medium Enterprises (SMEs) African Journal of Business Management, 5(22), 8971-8979

[38] Chen, C. J., and Huang, J. W. (2009) Strategic Human Resource Practices and Innovation Performance—The Mediating Role of Knowledge Management Capacity. Journal of Business Research, 62(1), 104-114

[39] Cho, J. J., Ozment, J., and Sink, H. (2008) Logistics Capability, Logistics Outsourcing and Firm Performance in an e-commerce Market. International Journal of Physical Distribution & Logistics Management, 38(5), 336-359

[40] Browne, M. W., and Cudeck, R. (1993) Alternative Ways of Assessing Model Fit. Sage Focus Editions, 154, 136-136

[41] Kline, R. B. (2005) Principles and Practice of Structural Equation Modeling (2nd ed) New York: Guilford Press

# Active Knowledge for the Situation-driven Control of Product Definition

**László Horváth and Imre J. Rudas**

Óbuda University, John von Neumann Faculty of Informatics, Institute of Applied Mathematics; Bécsi út 96/b, H-1034 Budapest, Hungary
e-mail: horvath.laszlo@nik.uni-obuda.hu, rudas@uni-obuda.hu

*Abstract: The current product model consists of features and unstructured contextual connections in order to relate features. The feature modifies the previous state of the product model producing contextual connections with previously defined features. Active knowledge is applied for the adaptive modification of product model features in the case of a changed situation or event. Starting from this state-of-the-art, the authors of this paper introduced a new method to achieve higher-level and more advanced active feature driven product model definition. As part of the related research program, new situation driven model definition processes and model entities are explained in this paper. Higher-level knowledge representation in the product model is motivated by a recent trend in industrial product modeling systems towards more advanced and efficient situation-based self-adaptive model generation. The proposed model represents one of the possible future ways of product model development for product lifecycle management (PLM) systems on the global or product level of decisions. Its implementation will be new application-oriented model entity generation and representation utilizing existing modeling resources in industrial PLM systems by use of application programming interfaces (API).*

*Keywords: product model; situation-based product definition; active knowledge in product model; product behavior driven feature definition*

## 1 Introduction

The work of engineers has been supported by computer applications for a long time. Digital drawing and documentation representations in the early years of computer-supported engineering were changed to geometric and then to product modeling during the eighties and nineties. Nowadays, the product model has the capability to represent arbitrary engineering objects together with their relationships. The trend towards virtual engineering space [1] can be recognized easily in current developments in order to represent all characteristics of a planned product and its connections. This paper introduces some results about a new concept, approach and method as a contribution towards achieving a complete and consistent model for virtual engineering space. According to this concept, the communication between engineer and model entity generation must be raised from the present product feature definition level to the function and knowledge request level. The approach covers communication of this request as human intent, its harmonization with other intents, and its repeated application for a changed situation and event in order to assist decisions on product features. For this purpose, the method is being developed for request definition and processing in the product model in the form of the appropriate structure of new features.

The proposed modeling is intended to realize the organic extension of current product modeling in product lifecycle management (PLM) systems towards better human communication and decision making. The results of the research are planned to be implemented experimentally into an open PLM system where application programming interface (API) is available for the programming of modeling in the application environment and to provide access to existing features, modeling procedures, and user surfaces in that PLM system. This also facilitates the strong application orientation of industrial product definition.

This paper is an extended and upgraded version of a contribution to the proceedings of the 6th International Workshop on Knowledge Management [2]. It starts with an introduction to the achievements in product modeling that are in close connection with the proposed function level human request based feature definition. The proposed product modeling extends the currently applied product modeling. A detailed introduction of the currently prevailing and well proven engineering technology is impossible within the scope of this paper. The reader is supposed to know recent advanced feature-based and knowledge-ware-driven contextual modeling in the PLM environment.

After an introduction to the currently prevailing advanced product modeling and its model representation capabilities, the knowledge processes in the current and proposed product definition are compared. Then the replacement of the current formal knowledge and object parameter value communication with the proposed influence request communication between engineer and product feature generation process is explained. Following this, current product model entities, including influence requests, actual objective structure, product feature context structure, and actual behavior structure, are detailed. Finally, affect analysis along contextual chains propagating a change of product feature is explained.

## 2 Preliminaries and Purpose

Engineering practice is moving into product lifecycle management (PLM) systems. A PLM system represents a high level of product information integration. The trend in this integration have led to complex and large product models with a very high number of entities and a crowd of unorganized relationships. Work in this paper was motivated by the recognition that the definition and processing of this current product model need more organized relationships with more efficient knowledge representation. In an earlier work [3], it was concluded that background content controlled engineering object definition would be required for product development in model space. The proposed new method of function level human request based feature definition is devoted to realizing this concept. It is grounded on former works in the modeling of engineering intent [4], the definition of product behavior in the product model [5], knowledge representation in the product model [6], and information content based product modeling [7]. Change propagation to engineering objects and adaptive action was described in [8]. These publications cover different aspects of knowledge based engineering modeling and lay the groundwork for the development of the method introduced in this paper.

The term ‘product model’ is applied for the definition of consistent engineering purposed computer representation of the product. It was grounded theoretically during development of the STEP (Standard for the Exchange of Product Model Data, ISO 10303) product model standard by the International Standards Organization (ISO). This is the only standard for a product model and is under implementation in different industries by using application protocols. Knowledge related engineering research works generally consider STEP. As a representative example, a framework was proposed for integrated data and knowledge model, using reference protocols in [9].

A step-by step organic development of engineering modeling and simulation has led to the present product lifecycle management (PLM) technology. In order to identify the place of the proposed modeling in product modeling, the development of engineering modeling and simulation is tracked in Fig. 1. Geometry was the first area of intensive research because mathematic description of shape was urged by the early control of machine tools and robots. Nowadays, the only representation of curves and surfaces in PLM systems is the sophisticated non-uniform rational B-spline (NURBS). The connection of knowledge and geometry for higher quality shapes is also a concern in the proposed modeling. On this topic, paper [10] introduces an application of knowledge-guided NURBS at repair imported incompatible geometric models. Boundary representation (B-rep) was involved in geometric modeling in order to complete geometry by topology for the representation of geometric structure. Euler operators are applied to establish the topological structure of the solid body using vertex, edge, and face entities, while geometry is defined during construction of the solid shape in the model space. In the seventies, a breakthrough in simulation technology was the application of finite element modeling and analysis (FEM/FEA). This facilitated the approximation of location dependent parameters at nodes in a mesh generated on a geometric model. By now, mesh is associative with geometry, parametric, and adaptive. FEM/FEA fits well into the knowledge based product model. By active application of FEM/FEA, design variables acting on product performance can be controlled. B-rep and FEM/FEA are essential elements of the current product definition in PLM systems.

The application and variant orientation of a product model is allowed by feature and parametric modeling principle, respectively. These principles and the related modeling achievements are in close connection with development of the proposed modeling concept, approach, and method. Form feature is an application oriented shape aspect and uses B-rep representation. The definition of relative placing and the moving of solids in model space represent placing constraints and degrees of freedom, respectively. Feature based product modeling is still a frequent area of research. In related research [11], parametric and semantics based feature is defined in order to support function-oriented product modeling together with better communication in an engineering modeling environment. Currently, the feature principle has been extended to all entities in the product model. During the nineties, parametric modeling allowed for the definition of dimensions as variables. The parameter model was introduced to represent the relationships among dimensions. In current product models, the arbitrary parameter can be defined and related.

This century started with two new essential techniques in product modeling. They are contextual modeling and knowledge ware. In contextual modeling, feature parameters are to be defined in the context of other feature or product level parameters. Knowledge ware is applied for the collection of local or corporate knowledge in order to facilitate situation and event based control of feature generation. Among others, knowledge ware includes representation of relationships and functional connections of parts and their features as well as related algorithms.

The development of knowledge ware driven product entity generation placed emphasis on knowledge management in PLM systems. The personalization and codification in knowledge management in engineering design are introduced for the development of a multidisciplinary framework in [12]. Paper [14] introduces an approach to the definition and mapping of expert knowledge. In [15], functional behavior and structure model is introduced as design conceptual space. Function behavior and structure model, and design process schema are suggested as way to understand the typical life cycle of a decision-making process in design. All knowledge sources are involved in this conceptual model. In [16], knowledge acquisition processes are defined to capture structured knowledge systematically, and knowledge representation technology is characterized to store knowledge, preserving important relationships. The objective is to establish a better solution than conventional engineering data bases. Paper [17] shows multiple expert inputs in knowledge learning. Input is processed by the Bayesian Network into a structure. Empirical data is applied to refine and parameterize this structure. The product model is in strong contextual connection with business processes. In [13], a method is shown for P-graph based workflow modeling of business processes. The above knowledge methodology developments are efforts to solve problems produced by the emerging knowledge based PLM. Modeling in this paper is devoted to a global solution on product level that can utilize the above methods at local level problem solving.

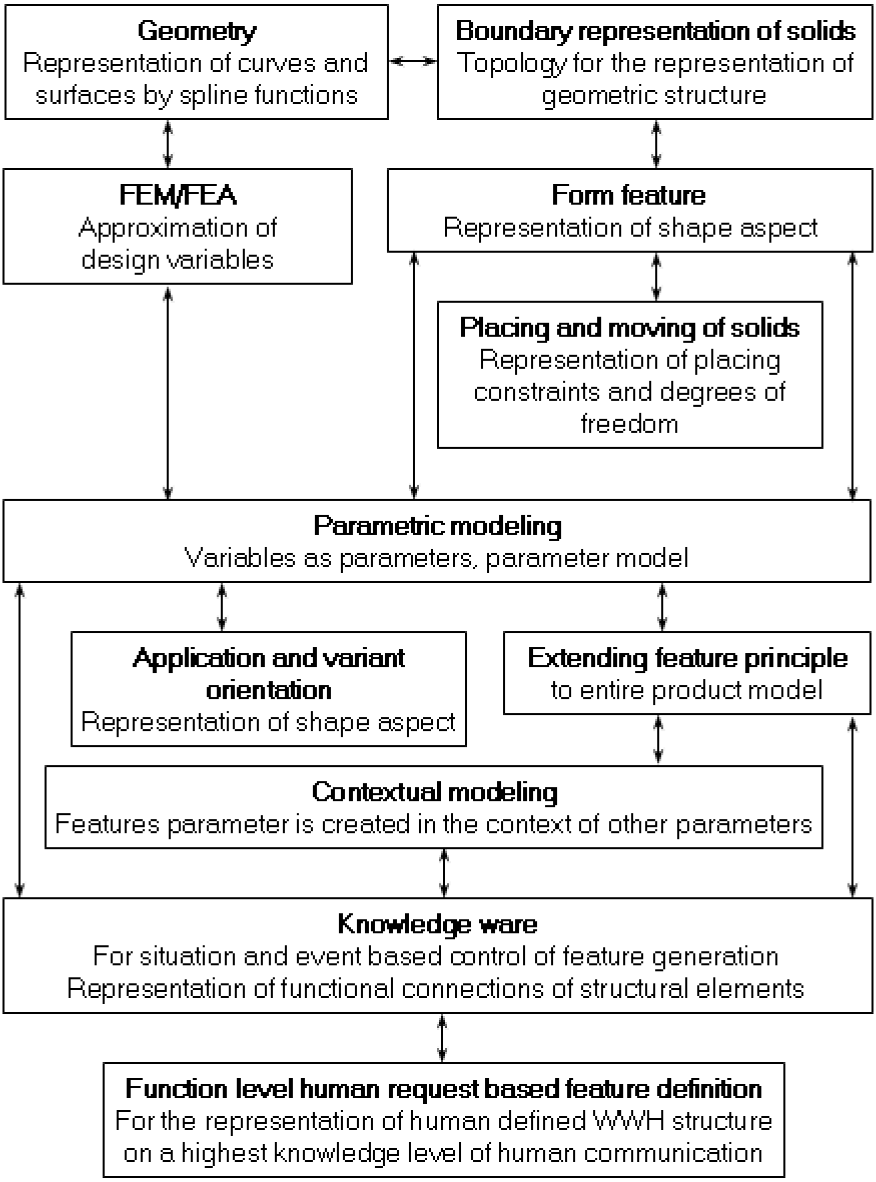



Figure 1

Step-by-step development of engineering modeling

The above characterization and summary of currently prevailing product modeling introduce the modeling proposed in this paper. The proposed modeling, called *Function level human request based feature definition,* is the result of efforts to bring product definition dialogue closer to human thinking. On a higher knowledge level of the human-modeling process communication, the human defines advanced model features in order to contribute via model representation of why, what, and how (WWH). The human requires product functions together with demanded quality and specified or proposed methods in a contextual structure. A decision on product feature definition is based on coordinated requests of authorized humans. It results in affect analyzed feature definition and applies adaptive actions for the control of product feature generation.

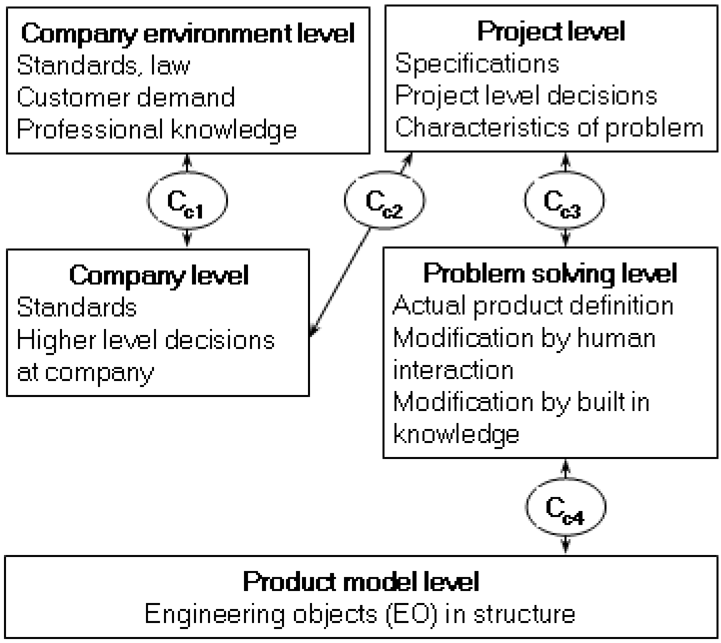



Figure 2
Levels in an engineering system

Decisions on engineering objects for a product are made in the context of relevant knowledge. Knowledge is defined on company environment, company, project, problem solving, and product model levels (Fig. 2). These levels constitute a contextual chain. On the company environment level, standards, laws and customer demand, as well as professional knowledge, are defined. Company level knowledge is defined in the context of company environment level knowledge ($C_{c1}$) and represents company standards and higher level decisions. Project level knowledge is defined in the context of company level knowledge ($C_{c2}$) and represents specifications, project level decisions, and the characteristics of problems. Problem solving level knowledge is defined in the context of project level knowledge ($C_{c3}$) and represents actual product definition, modification by human interaction, and modification by built in knowledge. Finally, the product model level includes engineering objects (EO) in an appropriate structure and is defined in the context of the problem solving level ($C_{c4}$). The modeling in this paper is proposed for the problem solving level. The trend to include higher level knowledge and specification in contextual chains within product model was considered.

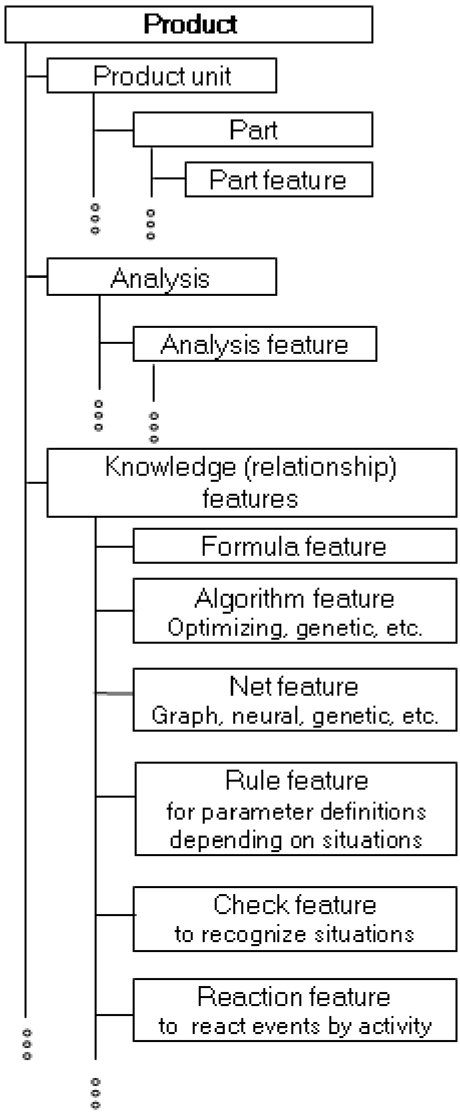



Figure 3

Product model level with engineering objects in structure

# 3 Current Practice

Including active knowledge representations in the product model is one of the most important trends in the current developments of product modeling. Purposeful knowledge features are available in PLM systems for the current product model.

Product objects as features are organized into units in the structure of the product model according to the model representation demanded by the actual product (Figure 3). Analysis and knowledge features are separated from these units in the structure for organizational and acceptance causes.

Simple, engineer-understandable but efficient knowledge representations are preferred by engineers. The formula is free to define as a means of context definition for product and other features. Optimizing, genetic, and other algorithms make user development of product modeling possible. Connectivity knowledge is included in the product model by various graphs, neural networks, genetic support, and other connective entities. The active definition and the checking of engineering objects are driven by rules, checks, and reactions (Fig. 3). When a rule, a reaction, or value of its parameter is defined or changed, the related engineering objects are changed automatically by the execution of the rule or reaction. Similarly, check notices change request and offers change for the relevant engineering objects. Check and rule are involved for handling situations, while reaction for handling events in product model. Check is applied to recognize the situation, whereas rule is applied for the parameter definition depending on the situation. Reaction activates predefined activity for new or changed event.

The current product model (Fig. 3) is considered in this paper as a classical way of product definition. It represents a well-proven principle, method and system background. Modeling in this paper is built as an extension to this strong basis. The knowledge process in the current or classical product definition was analyzed in order to identify knowledge-related problems during model construction and communication amongst engineers. In Fig. 4, two engineers communicate through a product model. One of the engineers is in interaction with the product feature definition processes. The knowledge background is available for human activities. This background is available as a representation or should be represented in the product model. However, a high proportion of knowledge is applied by the human but not communicated with the modeling system, because no capability is available for representation. The engineer communicates knowledge and object parameter values. Product feature generation processes generate features for the product model and place them in a structure. Similarly to human communication, a high proportion of knowledge is applied by processes, but no capability is available for representation in the product model. The other engineer is in interaction with product feature application processes. Due to the above shortage in knowledge representation capabilities, this engineer is not properly informed.

Information is not included in the product model about the process to a decision, allowed modifications, the strength of a decision, and some of the relationship definitions. The main purpose of the modeling proposed in this paper is to fill this gap caused by a lack of knowledge in the dotted boxes in Fig. 4.

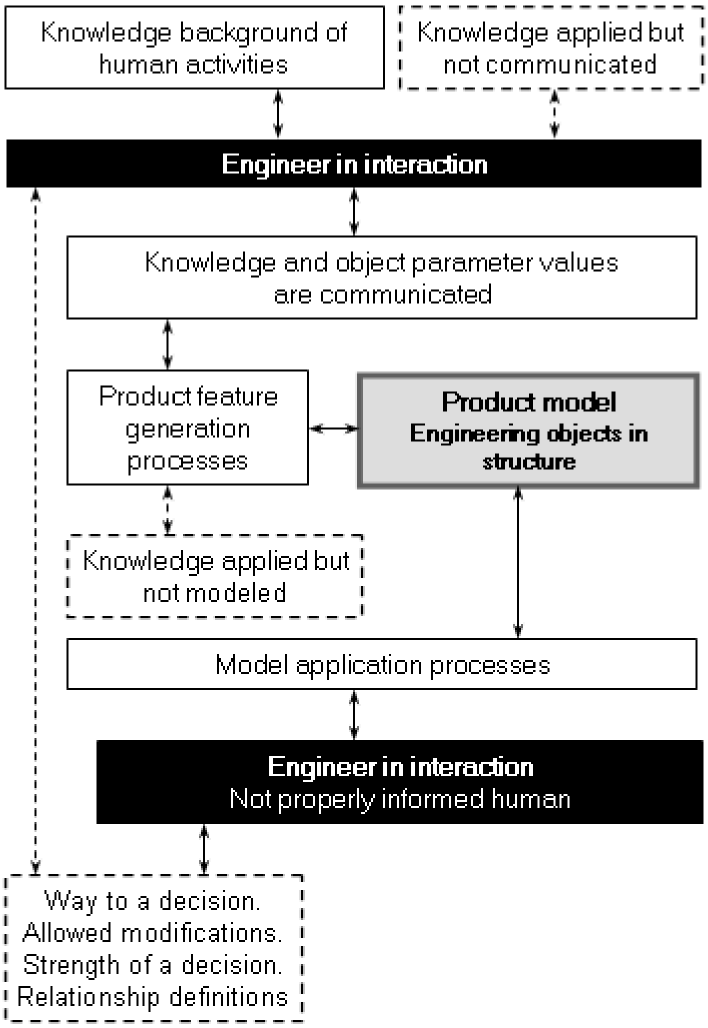



Figure 4

Knowledge process in current product definition

# 4 The Proposed Product Definition

Starting from the knowledge process in the current product definition, the extension to knowledge representation and its application for product model feature generation are introduced in Fig. 5. Beyond the knowledge background of human activities, outside expert sources and activities are to be connected to the modeling in order to allow for any knowledge to be included in the product model. Depending on the feature choice available in a PLM system, extensive knowledge representation in the product model is a real possibility in future engineering. Theoretically, lost knowledge is not allowed. Otherwise, the self-adaptive product model cannot work properly. However, knowledge that intentionally is not communicated or cannot be represented in the product model may be allowed to be applied by an engineer only in interaction. The resulted feature parameter values are applied as constraints on those can be modified according to coordination requests from different engineers. Constraining and contextual connection handling functionality in the current product definition is utilized at the proposed knowledge feature definition.

In the proposed method, knowledge and object parameter value communication between engineer and product feature generation processes are replaced by influence, or in other words, by product definition request communication (Fig. 5). Influence requests are represented for product model. Different requests for the same features are coordinated, and an actual influence definition is generated. Model feature generation always works by using the actual influence definition and modifies the affected features when this definition is modified. Compliance of the actual influence definition with the existing product model is evaluated by affect analysis. Product feature generation processes are controlled by using product behavior definitions. Consequently, any request against product feature generation must be defined as product behavior. For this reason, an extended definition of behavior was proposed in [5].

Extended model application processes communicate information carried by extension entities in the product model with an engineer who applies product model information. This engineer is a properly informed human depending on the model representation provided by the implementation of the extension. The product model to be extended includes engineering objects in product structure and product function structure as the main relevant elements of the current product model (Fig. 5). The extension involves new entities for the new modeling tasks including influence requests, actual objective structure, product feature context structure, and actual behavior structure. These entities will be detailed and explained in the rest of this paper (Figs. 6 and 7).

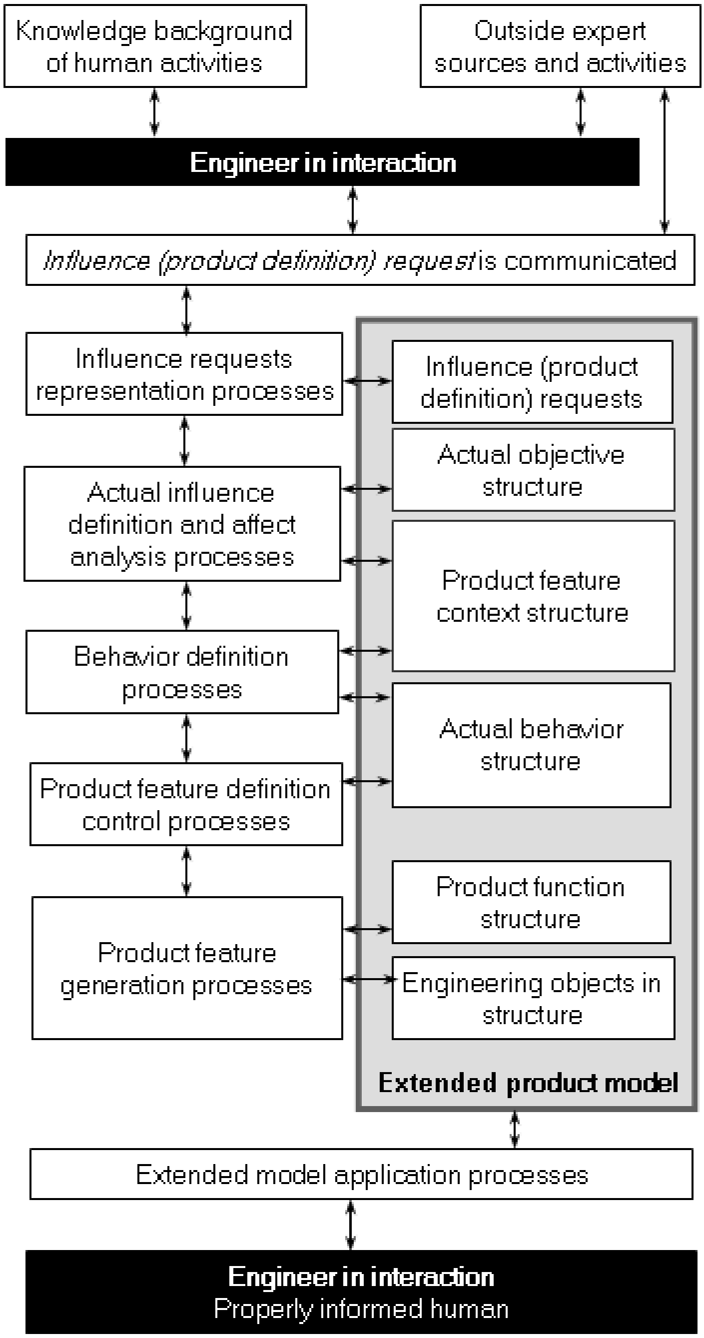



Figure 5
Knowledge process in the proposed product definition

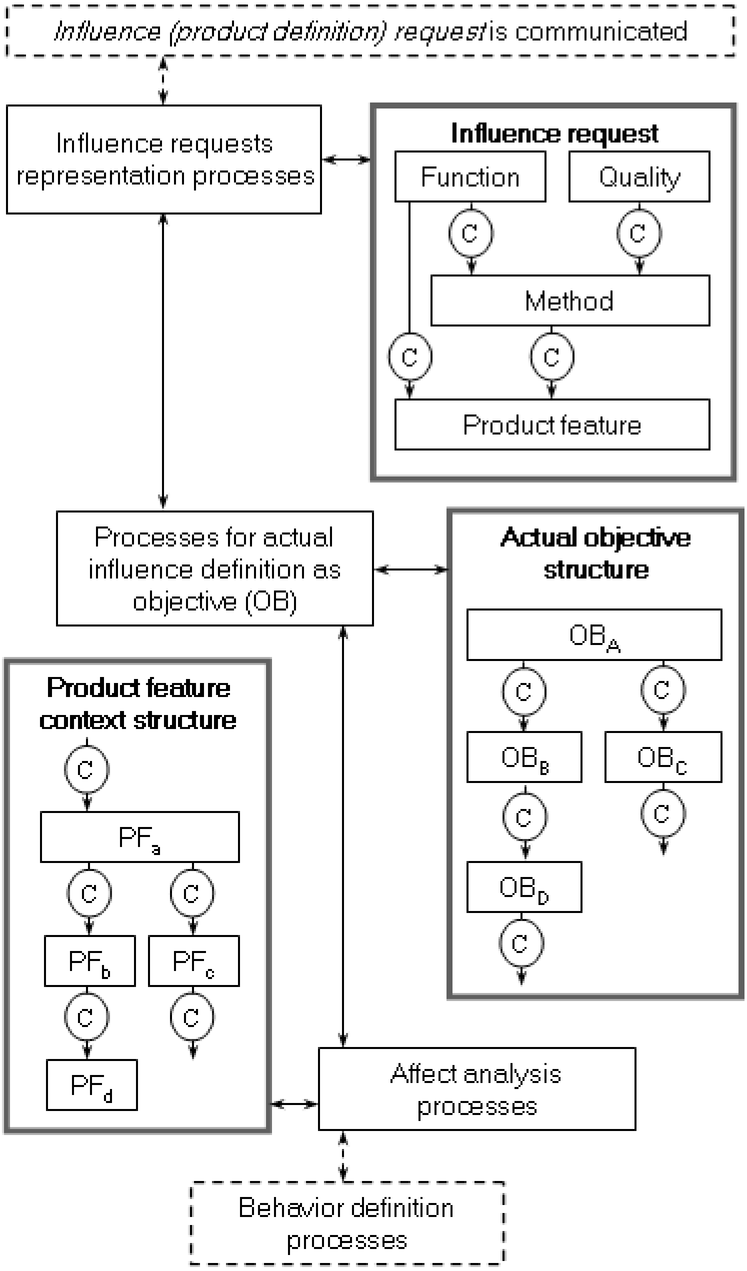



Figure 6
Influence request and objective

# 5 New Product Model Entities

The new entities are represented as features for the proposed modeling and are in appropriate contextual connection with current product model features. The influence request is communicated with the appropriate processes in the extension and then processed into an influence request feature.

The influence request feature is contextual with the function to be realized, the quality to be kept, the method to be applied, and the product features to be generated for the product under development (Fig. 6). Letter 'C' in a circle symbolizes a contextual connection entity, while the arrow shows the direction of context.

The method for the definition of the product features is defined in the context of function. Quality is a specification for the awaited function related characteristics of the product. A product feature is defined in the context of function and quality. Entity contextual connection is free to be defined using a wide range of tools available in PLM systems, such as formula, function, algorithm, rule, etc. The well proven construction tools are available in the PLM system.

Product feature generation is controlled by the actual behavior structure (Fig. 7). Behavior provides objective driven product feature generation instead of the current conventional direct product feature generation. The behavior feature will be detailed and explained in the rest of this paper (Fig. 8). The behavior definition is assisted by the situation and circumstance set definition.

In the background of a behavior, the situation collects circumstances as product functions, objects, parameter definitions, etc. and provides all information to be considered at the product object definition. The control of object generation is governed by the use of information in adaptive action. Adaptive action is also handled as a feature and it is made ready to execute by the appropriate decision on product objects.

The main contextual connections in the product model extension constitute contextual main chains, as shown in Fig. 8. The function in the product function structure and the objective in the actual objective structure are defined in the context of the influence request. The product function structure connects product features by functions and it is available in current advanced product models. Behaviors are organized for the product in the actual behavior structure. This structure includes behaviors (BH) connected by contextual definitions. Behavior in the actual behavior structure is defined in the context of the relevant objective in the actual objective structure, and it also depends on the contextual connections inside of this structure. Basically, behavior is defined as a contextual chain of behavior (as defined), situation, circumstances, and adaptive action. Product features and their contextual connections are defined in the context of adaptive action. At the same time, engineering objects in their structure are generated in the context of product features and their contextual connections in product feature context structure.

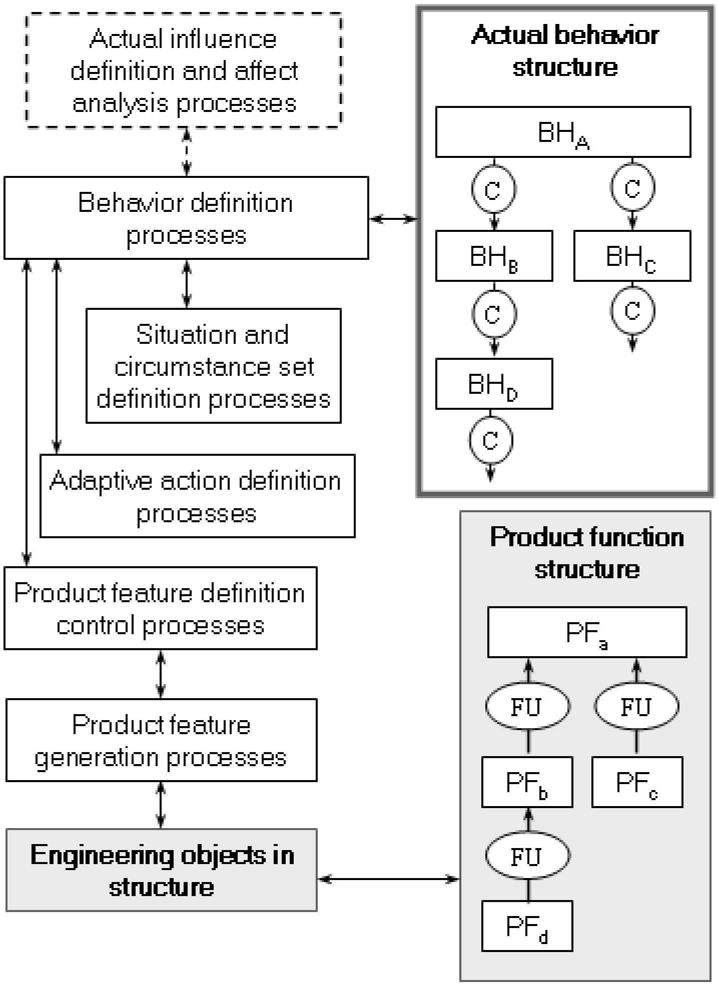



Figure 7
Behavior and control of product feature generation

Influence requests from different engineers in the PLM system are synthesized into an actual objective structure. Parallel and overlapping influence definitions are decided and harmonized into actual influence definitions, as was conceptualized in [7]. These definitions are handled as objectives. The elaboration of the methods for these modeling functions is a future plan. At this time, the application of dialogues provided by current PLM systems is considered. The actual objective structure consists of objectives and their contextual connections. As was stated above, a lack of organized contexts in the current product model requires a special feature for the product feature context structure in the product model extension. This structure consists of product features (PF) and their contextual connections and is used via affect analysis processes. Affect analyzed and corrected actual objective structure is applied by behavior definition processes.

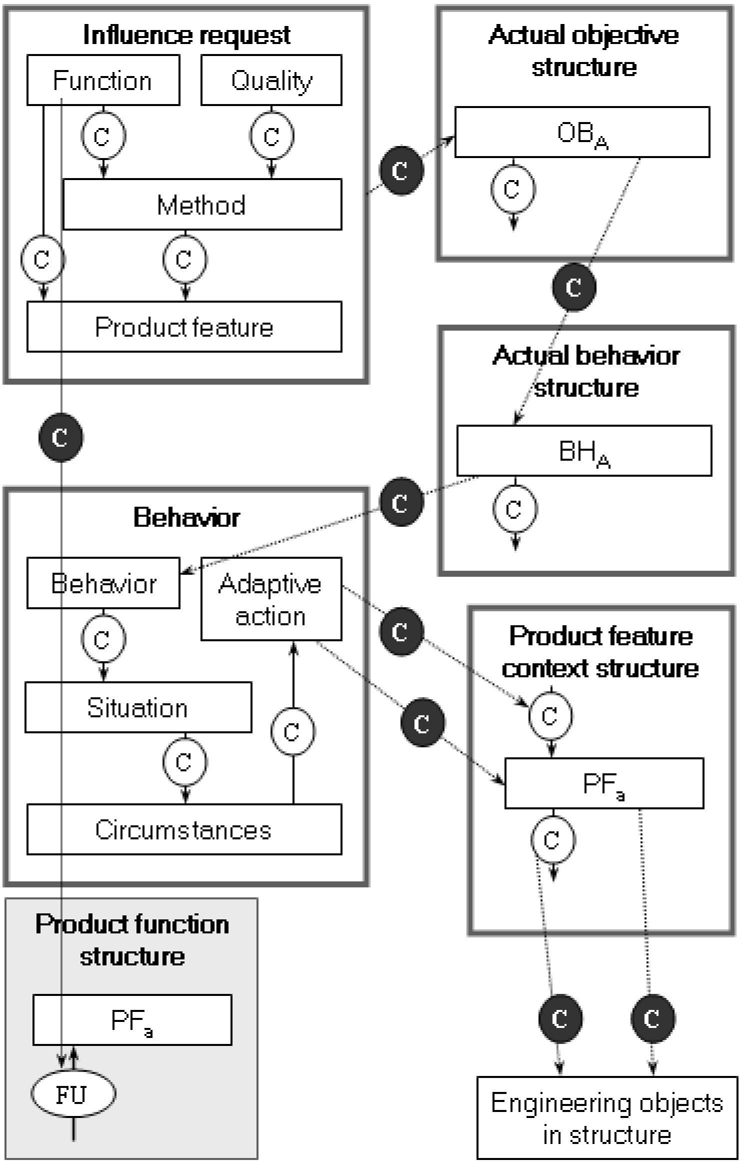



Figure 8

Main contextual connections in the extension of product model

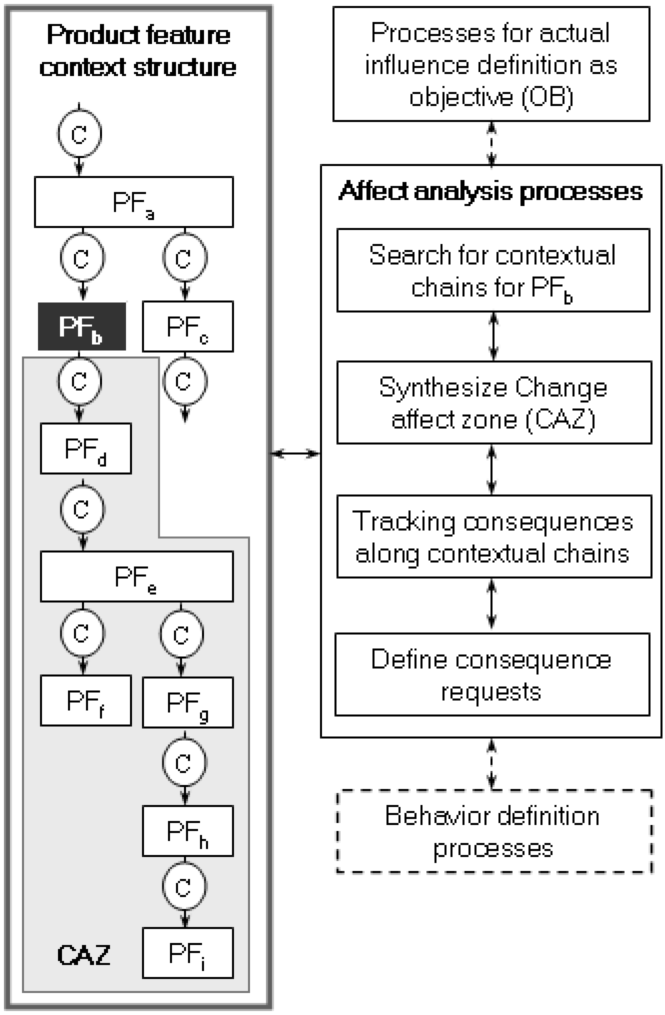



Figure 9
Affect analyses in CAZ

Analysis for the affects of actual objectives on existing or already decided product objects is done, as explained in Fig. 9. Affect analysis is done along the contextual chains propagating change of a product feature through other product features. In Fig. 9, this product feature is $PF_b$. These chains are searched in the change affect zone (CAZ). The CAZ includes all potentially affected product features and synthesizes contextual product feature chains. Tracking consequences along the contextual chains results in consequence requests to be included in the product model extension. Objectives and behaviors must be updated accordingly.

## Conclusions

This paper introduces new results in modeling methodology and related model entities in order to enhance knowledge communication between human and product feature generation processes and to establish an organized knowledge background of the product feature definition in the product model. The currently applied and well-proven product modeling technology is extended. The well-established direction of development of this modeling technology makes continuous development of the proposed extension to the currently applied modeling possible

Due to a shortage in the knowledge representation capabilities of the current product modeling, information about the process to a decision, allowed modifications, the strength of a decision, and some relationship definitions cannot be included in the product model. The proposed extension applies influence request for communication between human and model generation, active objective structure for representation of coordinated and organized objectives of product development, actual behavior structure to represent new features for the control of product feature generation. The work has relied upon earlier results by the authors in this area of engineering modeling.

## Acknowledgement

The authors gratefully acknowledge the financial support by the Óbuda University research fund.

## References

[1] Horváth, L., Rudas, I. J.: Product Definition in Virtual Space Using Background information at Description of Engineering Objects, in *WSEAS Transactions on Systems and Control*, ISSN 1991-8763,Volume 3, Issue 9, 2008, pp. 748-758

[2] Horváth, L., Rudas, I. J.: Knowledge Representations for the Practice of PLM Systems, in proc. of the *6th International Workshop on Knowledge Management*, Trencin, Slovakia, 2011, pp. 65-76

[3] Horváth, L., Rudas, I. J.: Modeling of the Background of Human Activities in Engineering Modeling, Proc. of the *27th Annual Conference of the IEEE Industrial Electronics Society*, Denver, Colorado, USA, 2001, pp. 273-278

[4] Horváth, L., Rudas, I. J.: Intelligent Human-Computer Communication of Engineers at Extended Companies, in *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Volume 10, Number 4, 2006, pp. 510-516

[5] Horváth, L., Rudas, I. J.: Adaptive Modeling for Robot Systems Based on Behaviors and Feature Driven Shape Descriptions, in *WSEAS Transactions on Information Science and Applications*, Volume 11, Number 2, 2005, pp. 1761-1771

[6] Horváth, L., Rudas, I. J.: Knowledge Technology for Product Modeling, Chapter 5 of the book *Knowledge in Context – Few Faces of the Knowledge Society*, Walters Kluwer, 2010

[7] Horváth, L., Rudas, I. J.: Engineering Objective Driven Product Lifecycle Management with Enhanced Human Control, In book *Towards Intelligent Engineering and Information Technology*, in Series: Studies in Computational Intelligence, Vol. 243, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 543-556

[8] Horváth, L.: New Design Objective and Human Intent Based Management of Changes for Product Modeling, Acta Polytechnica Hungarica, Volume 4, No. 1, 2007, pp. 17-30

[9] Jardim-Goncalves, R. Figay, N. Steiger-Garcao, A.: Enabling Interoperability of STEP Application Protocols at Meta-Data and Knowledge Level. International Journal of Technology Management, Volume 36, Number 4, 2006, pp. 402-421

[10] Rajab, K., Piegl, L.A., Smarodzinava, V.: CAD Model Repair Using Knowledge-guided NURBS, *Engineering with Computers*, Volume 28, Number 2, 2012, pp. 1-10

[11] Bronsvoort, W. F., Bidarra, R., Nyirenda, P. J.: Developments in Feature Modeling, *Computer-Aided Design and Applications*, Volume 5, No. 3, 2006, pp. 655-664

[12] McMahon, C., Lowe, A. Culley, S.: Knowledge Management in Engineering Design: Personalization and Codification, *Journal of Engineering Design*, Volume 15, Number 4, 2004, pp. 307-325

[13] J. Tick: Visualization and Simulation Tool for Analyzing P-Graph-based Workflow Systems, in *Applied Computational Intelligence in Engineering and Information Technology*, Springer-Verlag, Berlin Heidelberg 2012, pp. 71-82

[14] Renaud, J.: Improvement of the Design Process through Knowledge Capitalization: an Approach by Know-how Mapping, *Concurrent Engineering*, Volume 12, Number 1, 2004, pp. 25-37

[15] Colomboa, G., Moscaa, A., Sartori, F.: Towards the Design of Intelligent CAD Systems: An Ontological Approach, *Advanced Engineering Informatics*, Volume 21, Number 2, 2007, pp. 153-168

[16] Preece, A., Flett, A., Sleeman, D., Curry, D, Meany, N., Perry, P.: Better Knowledge Management through Knowledge Engineering, *Intelligent Systems*, Volume 16, Number 1, 2005, pp. 36-43

[17] Richardson, M., Domingos, P.: Learning with Knowledge from Multiple Experts, in proc. of the *Twentieth International Conference on Machine Learning*, Washington, DC, Morgan Kaufmann, 2003, pp. 624-631